

# Probing Unlearned Diffusion Models: A Transferable Adversarial Attack Perspective

Xiaoxuan Han

University of Chinese Academy of Sciences  
Beijing, China  
hanxiao2023@ia.ac.cn

Songlin Yang

University of Chinese Academy of Sciences  
Beijing, China  
yangsonglin2021@ia.ac.cn

Wei Wang\*

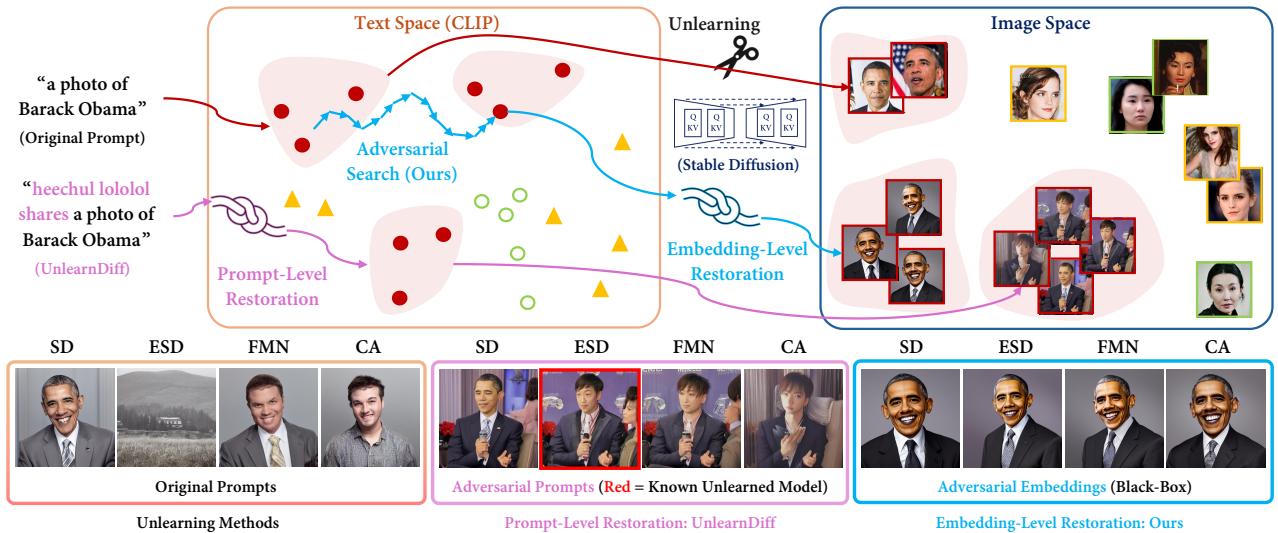
Institute of Automation, Chinese Academy of Sciences  
Beijing, China  
wwang@nlpr.ia.ac.cn

Yang Li

University of Chinese Academy of Sciences  
Beijing, China  
liyang2022@ia.ac.cn

Jing Dong

Institute of Automation, Chinese Academy of Sciences  
Beijing, China  
jdong@nlpr.ia.ac.cn



**Figure 1:** The overview of concept erasure and restoration for text-to-image Stable Diffusion [25] (SD) model. Due to content safety concerns, unlearning methods (e.g., ESD [7], FMN [39], and CA [15]) have been investigated to erase target concepts by shifting the text-to-image mapping. But these erasure methods leave a fatal flaw for restoring the erased concepts. Towards this, we propose an adversarial attack to probe erasure trustworthiness. Compared with previous methods, our black-box method can transfer across different unlearned models and is effective for the challenging task of celebrity identity restoration.

## ABSTRACT

Advanced text-to-image diffusion models raise safety concerns regarding identity privacy violation, copyright infringement, and Not Safe For Work (NSFW) content generation (e.g., nudity). Towards this, unlearning methods have been developed to erase these involved concepts from diffusion models. However, these unlearning methods only shift the text-to-image mapping and preserve the visual content within the generative space of diffusion models, leaving a fatal flaw for restoring these erased concepts. This erasure trustworthiness problem needs probe, but previous methods are sub-optimal from two perspectives: **(1) Lack of transferability:** Some methods operate within a white-box setting, requiring access to the unlearned model. And the learned adversarial input often

fails to transfer to other unlearned models for concept restoration; **(2) Limited attack:** The prompt-level methods struggle to restore narrow concepts from unlearned models, such as celebrity identity. Therefore, this paper aims to leverage the transferability of the adversarial attack to probe the unlearning robustness under a black-box setting. This challenging scenario assumes that the unlearning method is unknown and the unlearned model is inaccessible for optimization, requiring the attack to be capable of transferring across different unlearned models. Specifically, we first analyze the reasons for the poor transferability of previous methods. Then, we employ an adversarial search strategy to search for the adversarial embedding which can transfer across different unlearned models. This strategy adopts the original Stable Diffusion model as a surrogate model to iteratively erase and search for

\*Corresponding author.

embeddings, enabling it to find the embedding that can restore the target concept for different unlearning methods. Extensive experiments demonstrate the transferability of the searched adversarial embedding across several state-of-the-art unlearning methods and its effectiveness for different levels of concepts. Our code is available at <https://github.com/hxxdtd/PUND>. **CAUTION: This paper contains model-generated content that may be offensive.**

## CCS CONCEPTS

- Security and privacy → Privacy protections; • Computing methodologies → Image representations; *Image manipulation*.

## KEYWORDS

Diffusion Model, Machine Unlearning, and Adversarial Attack

## 1 INTRODUCTION

Developing Artificial Intelligence Generated Content (AIGC) is a double-edged sword. Although text-to-image (T2I) generative models [10, 19, 22, 25, 26, 37] can generate high-quality and diverse images according to the given prompts, they also raise significant safety concerns regarding identity privacy [2], copyright [29], and Not Safe For Work (NSFW) content [23, 27]. For example, these models can generate portraits of celebrities known as deepfakes [30], while some painting styles of artists can be imitated. Besides, these models can generate Not Safe For Work (NSFW) content, such as nudity and violence. To mitigate these issues, as shown in Figure 1, concept erasure methods [5, 7, 8, 12, 15, 39] for erasing the involved concepts have been developed, falling under the category of machine unlearning [14, 18, 28, 32, 34]. However, existing methods accomplish the “erasure” task by shifting the text-to-image mapping and fail to erase the visual content within the generative space of diffusion models, leaving a fatal flaw for restoring these erased concepts. This inspires the question: **How trustworthy are unlearning methods for text-to-image diffusion models?**

Previous methods are sub-optimal for this question from two perspectives. **(1) Lack of transferability:** Some methods operate within a white-box setting, requiring access to the unlearned model. And the learned adversarial input often fails to transfer to other unlearned models for concept restoration. As shown in Figure 1, methods like UnlearnDiff [41] make a strong assumption that the unlearned generative model is accessible, which lacks transferability and is less practical in real-world scenarios. **(2) Limited attack:** The prompt-level methods struggle to restore narrow concepts from unlearned models, such as celebrity identity (ID). As depicted in Figure 1, UnlearnDiff [41] tries to optimize an adversarial prompt to restore the target concept (i.e., “Barack Obama”). But its representative capability is constrained by the discrete nature of prompt tokens, making accurate identity restoration challenging.

In this paper, we aim to probe the erasure robustness of unlearned diffusion models under the black-box setting, where the attacker lacks knowledge of the unlearning methods and the unlearned models are inaccessible for the optimization. This is significantly challenging, especially for narrow concepts such as identity. We first analyze the reasons for the poor transferability of previous method. To tackle this challenge, we utilize an adversarial search

strategy to find the adversarial embedding transferable across different unlearned models. This strategy adopts an original Stable Diffusion model as a surrogate model to alternately erase and search for embeddings, guiding the embedding search from high-density regions to low-density ones. These embeddings located in low-density regions are difficult to erase, enabling them to restore the target concept for different unlearned methods.

**Our contributions are summarized as follows:**

- We propose a transferable adversarial attack to probe the unlearning robustness, which can transfer across diverse unlearned models and tackle the challenge of ID restoration.
- We improve the transferability by iteratively erasing and searching for the embeddings that can restore the target concept. The obtained embeddings are located in low-density regions and very likely to be overlooked by erasure methods, thus possessing greater restoration capabilities.
- Extensive experiments demonstrate the transferability of the searched adversarial embedding across various state-of-the-art unlearning methods, along with its effectiveness across diverse levels of concepts ranging from broad to narrow.

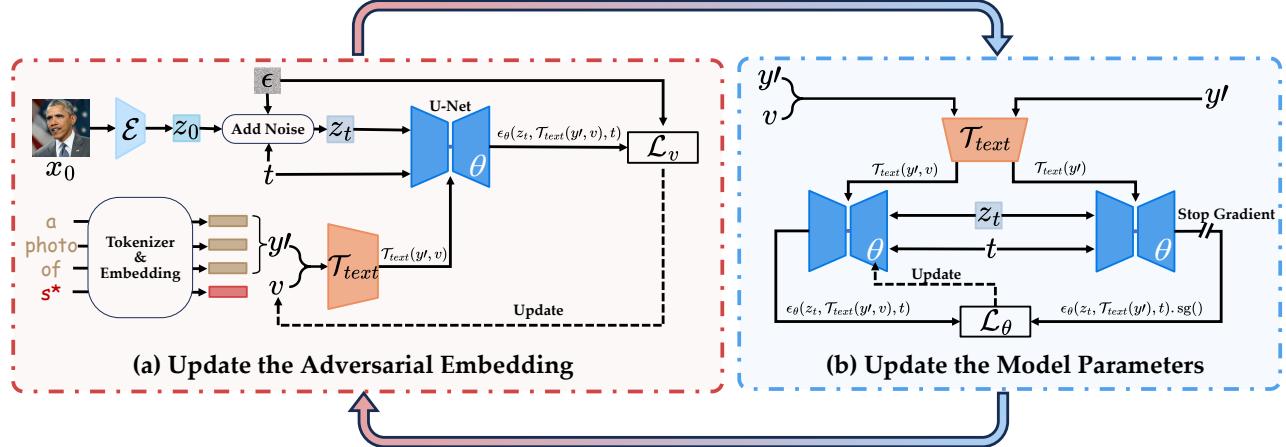
## 2 RELATED WORK

### 2.1 Text-to-Image Diffusion Models

Text-to-Image (T2I) Diffusion models are a variant of diffusion models, providing text-conditional guidance for image generation. The training of diffusion models includes two processes [13, 31]. In the diffusion process, noise is gradually added to the image  $x_0$  over multiple steps. In the reverse process [16, 17, 31], the model learns to predict the noise given the time step  $t$  and the noised version image  $x_t$ . T2I diffusion models [10, 22, 25] typically conduct the diffusion and reverse process in the latent space [4] of lower dimension for better computation efficiency. And various conditional mechanisms [25, 36, 40] (such as text prompt), are introduced for diverse controllable T2I generation. But the inappropriate text input can result in undesirable generated images [27, 35, 38], prompting the development of concept erasure methods to address this issue.

### 2.2 Diffusion Unlearning for Concept Erasure

A direct way to remove undesirable concepts is to filter out the inappropriate images and use the remaining data to retrain the model. But it is hard to find a perfect classifier to detect inappropriate images, and retraining the model from scratch requires large amounts of computation resources. Therefore, existing unlearning methods focus on erasing concepts from the trained model. Erased Stable Diffusion (ESD) [7] guides the predicted noise in the opposite direction of the target concept (i.e., the concept to erase) to decrease the probability of target concept generation. Concept Ablation (CA) guides the distribution of target concept towards a broader concept called the anchor concept. Forget-Me-Not (FMN) [39] doesn’t direct the target concept distribution to a specified distribution; instead, it minimizes the attention map corresponding to the target concept. Unlike the above fine-tuning-based methods, Unified Concept Editing (UCE) [8] is a model editing method that applies a closed-form editing to the liner projection weights of cross-attention parts. While existing unlearning methods demonstrate good performance



**Figure 2: The Adversarial Search (AS) strategy for concept restoration. We adopt the original Stable Diffusion model as a surrogate model to alternately erase and search for embeddings which can restore the target concepts.**

under benign input, their reliability under adversarial scenarios requires further investigation.

### 2.3 Adversarial Concept Restoration

UnlearnDiff attack [41] uses the classifier nature of diffusion models and employs Projected Gradient Descent (PGD) to optimize in the token space for concept restoration. But it requires access to the unlearned model. In contrast, Ring-A-Bell [33] assumes that the unlearned model is inaccessible and utilizes the text encoder to search for adversarial prompts. It employs prompts with and without the target concept to obtain its representation in the embedding space, then utilizes a genetic algorithm to search for adversarial prompts in the discrete token space. While Ring-A-Bell mainly focuses on NSFW content, it overlooks the probe of narrower concepts like celebrity identity. Pham et al. [20] probe a wide range of concepts, including celebrity identity, artist styles and NSFW content, with the aid of Textual Inversion [6]. However, they also require access to the unlearned model like [41]. To overcome the limitations of existing methods, we exploit the transferability of adversarial embedding to probe the restoration of diverse concepts, particularly narrow ones, without needing access to the unlearned model during the optimization process.

## 3 METHOD

In this section, we first introduce the preliminaries for the Stable Diffusion model, and explain how to conduct concept restoration at the embedding level. Then, we analyze the reasons for the poor transferability of the naive method and present the motivation behind the proposed strategy for improving transferability. Finally, we introduce the formulation and implementation of this strategy.

### 3.1 Preliminaries

**Stable Diffusion.** We focus on the widely employed latent diffusion models [25], also known as Stable Diffusion. In the training process, the image  $x$  undergoes encoding through the encoder  $\mathcal{E}$ , yielding its latent representation  $z = \mathcal{E}(x)$ . Subsequently, Gaussian noise is gradually added to  $z_0$ , generating a sequence of noised versions

of the latent representation  $\{z_1, z_2, \dots, z_T\}$ , where  $T$  denotes the total number of time steps. The training objective is to predict the Gaussian noise  $\epsilon$  as follows:

$$\mathcal{L} = \mathbb{E}_{z \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[ \|\epsilon - \epsilon_\theta(z_t, \mathcal{T}_{text}(y), t)\|_2^2 \right], \quad (1)$$

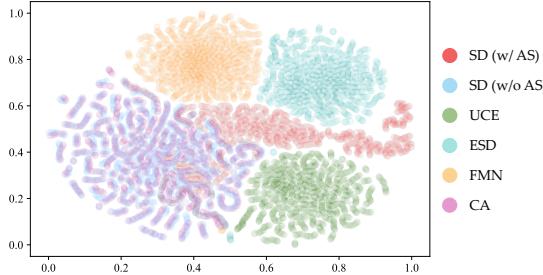
where  $y$  represents the token embeddings of the input prompt,  $\mathcal{T}_{text}$  denotes the text encoder used to obtain the conditional guidance of  $y$ ,  $t$  is the current time step, and  $\theta$  denotes model parameters.

**Concept Restoration.** To restore the target concept from the unlearned model parameterized by  $\hat{\theta}$ , Pham et al. [20] utilized Textual Inversion [6] to obtain the adversarial embedding for restoration. Concretely, they replaced the target concept in the input prompt with the placeholder token  $S^*$  (e.g., "a photo of <target-concept>" is modified as "a photo of  $S^*$ "), and optimized the corresponding embedding of  $S^*$  denoted by  $v$ . The modified prompt is then fed into the text encoder to obtain the new text condition  $\mathcal{T}_{text}(y', v)$ , where  $y'$  denotes the token embeddings of the original prompt but without the target concept (e.g., "a photo of"). Pham et al. [20] applied Textual Inversion to the unlearned model to find the optimal embedding for the token  $S^*$  by solving the optimization problem below:

$$v^* = \arg \min_v \mathbb{E}_{z \sim \mathcal{E}(x), y', \epsilon \sim \mathcal{N}(0,1), t} \left[ \|\epsilon - \epsilon_{\hat{\theta}}(z_t, \mathcal{T}_{text}(y', v), t)\|_2^2 \right] \quad (2)$$

where  $v^*$  denotes the learned adversarial embedding.

The method proposed by Pham et al. [20] proves effective for restoring diverse concepts, including narrow ones such as celebrity identity, which are challenging to restore using prompt-level attack methods such as Ring-A-Bell [33] and UnlearnDiff [41] due to the discrete nature of text representation. However, Pham et al. [20] made a strong assumption that the attacker had access to unlearned models. In this paper, we aim to perform concept restoration with access only to the original Stable Diffusion model (i.e., the model



**Figure 3: The visualization of the embeddings obtained from different unlearned models for the restoration of “Barack Obama”. Purple, yellow, cyan, and green points represent embeddings obtained from the models that have been unlearned by CA [15], FMN [39], ESD [7], and UCE [8], respectively. The blue and red points represent the embeddings acquired from the original Stable Diffusion (SD) model, while the red ones are obtained with our Adversarial Search (AS) strategy.**

before unlearning) parameterized by  $\theta$ , and leverage the transferability of the adversarial embedding to probe the robustness of different unlearning methods.

### 3.2 Transferable Adversarial Search Strategy

**Analysis for Poor Transferability.** Concept restoration aims to find the embedding capable of restoring the target concept within the model. A naive approach involves directly applying Textual Inversion [6] to the original model. However, as illustrated in Figure 3, the scatter plot displays that embeddings, represented by cyan, blue, yellow, purple, and green points, obtained from a single model—whether it is the original or an unlearned one—tend to exhibit distinctly separated cluster distributions. In other words, these embeddings obtained from the original model are highly likely to be ineffective for unlearned models. One possible explanation is that, for the original model, the learned embedding closely resembles the token embedding of the target concept, thus having a high probability of being erased. The transferable embedding we seek needs to diverge from the token embedding of the target concept. However, the embedding optimized for the original model may be stuck around the token embedding of the target concept due to the local minimum nature inherent in gradient-based optimization methods. Hence, additional guidance is necessary to facilitate the optimization process and explore more potential regions.

**How to Improve Transferability?** In the continuous embedding space, the distribution of the target concept may conform to a specific distribution, with high-density regions likely centered around the token embedding of the target concept. Given that most unlearning methods [7, 8, 15] utilize the token of the target concept for erasure, these high-density regions are mostly erased. Consequently, the transferable embedding for target concept restoration is likely to reside in low-density regions. Thus we need to guide the embedding search from high-density regions to low-density ones. To achieve this, we employ an Adversarial Search (AS) strategy. This strategy initially seeks the embedding from the original model, attempts to erase it, and then repeats the process of embedding search and erasure iteratively. After a number of iterations, we can

obtain a series of embeddings, among which lies the transferable one. As depicted in Figure 3 by the red points, after applying our adversarial search strategy, it is possible to avoid falling into the distribution of a single model. Therefore, the red portion exhibits a bar-like distribution, showing a tendency to intersect with all distributions where restorations are successful.

**Algorithm 1:** Searching for the transferable embedding for concept restoration

---

**Input:** Original model parameterized by  $\theta$ , training images  $X$ , total epochs  $E$ , image encoder  $\mathcal{E}$ , embedding update iterations per epoch  $I_v$ , total time steps  $T$ , noise coefficient sequence  $(\bar{\alpha}_t)_{t=1}^T$ , text encoder  $\mathcal{T}_{text}$ , neutral token embeddings  $y'$ , parameter update frequency  $f$ , parameter update iterations  $I_\theta$

**Output:** Candidate embedding set  $V$

```

1 Initialize:  $v \leftarrow v_0$ ;
2 for  $e$  in range( $E$ ) do
3    $x_0 \in X$ ; // randomly pick the training image
4    $z_0 = \mathcal{E}(x_0)$  ;
5   // update the embedding for  $I_v$  iterations
6   for  $i$  in range( $I_v$ ) do
7      $t \sim \text{Uniform}([1..T])$ ,  $\epsilon \sim \mathcal{N}(0, I)$  ;
8      $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$  ;
9      $\mathcal{L}_v = \|\epsilon - \epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t)\|_2^2$  ;
10    Update  $v$  with gradient descent to minimize  $\mathcal{L}_v$  ;
11  end
12   $V \leftarrow V \cup \{v\}$ ; // add  $v$  to the candidate set
13  // update model parameters every  $f$  epochs
14  if  $f \mid e$  then
15    for  $i$  in range( $I_\theta$ ) do
16       $t \sim \text{Uniform}([1..T])$ ,  $\epsilon \sim \mathcal{N}(0, I)$  ;
17       $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$  ;
18       $\mathcal{L}_\theta = \|\epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t) - \epsilon_\theta(z_t, \mathcal{T}_{text}(y'), t).\text{sg}()\|_2^2$  ;
19      Update  $\theta$  with gradient descent to minimize  $\mathcal{L}_\theta$  ;
20    end
21 return  $V$ 

```

---

**Strategy Formulation.** Adversarial search is formulated as a Min-Max optimization process, expressed as follows:

$$\min_v \max_\theta \mathbb{E}_{z \sim \mathcal{E}(x), y' \sim \mathcal{N}(0, 1), t} \left[ \|\epsilon - \epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t)\|_2^2 \right]. \quad (3)$$

The inner maximization aims to update the model parameters to maximize the noise prediction loss, preventing the learned embedding from restoring the target concept. However, directly maximizing the loss can significantly impair the model’s capacity in image generation, as also noted in [15]. Thus, we relax the inner maximization by introducing  $\tilde{\epsilon}$ , satisfying  $\|\epsilon - \tilde{\epsilon}\|_2 \geq d >$

0, where  $d$  is a constant. In the inner maximization, maximizing  $\|\epsilon - \epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t)\|_2^2$  is equivalent to maximizing  $\|\epsilon - \epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t)\|_2$ , which can be rewritten as follows:

$$\begin{aligned} \|\epsilon - \epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t)\|_2 &= \|(\epsilon - \tilde{\epsilon}) - (\epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t) - \tilde{\epsilon})\|_2 \\ &\geq \|\epsilon - \tilde{\epsilon}\|_2 - \|\epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t) - \tilde{\epsilon}\|_2 \\ &\geq d - \|\epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t) - \tilde{\epsilon}\|_2. \end{aligned} \quad (4)$$

Instead of maximizing the loss itself, we relax it by maximizing its lower bound, which is equivalent to a minimization problem as below:

$$\min_{\theta} \|\epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t) - \tilde{\epsilon}\|_2^2. \quad (5)$$

Then we need to determine the assignment of  $\tilde{\epsilon}$ . Naturally,  $\tilde{\epsilon}$  should maintain a certain distance  $d$  from  $\epsilon$  to ensure a large lower bound. In practice, we adopt the following assignment:

$$\tilde{\epsilon} = \epsilon_{\theta_0}(z_t, \mathcal{T}_{text}(y'), t), \quad (6)$$

where  $\theta_0$  is the original model parameters (i.e., parameters before updating),  $y'$  is token embeddings of a neutral prompt (i.e., "a photo of"). Then, the inner optimization becomes:

$$\min_{\theta} \|\epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t) - \epsilon_{\theta_0}(z_t, \mathcal{T}_{text}(y'), t)\|_2^2. \quad (7)$$

But the above minimization process has a relatively large memory burden, as the attacker needs to make a copy of the model parameters (i.e.,  $\theta_0$ ) before updating. To address this, we use  $\epsilon_\theta(z_t, \mathcal{T}_{text}(y'), t)$  to approximate  $\epsilon_{\theta_0}(z_t, \mathcal{T}_{text}(y'), t)$ , assuming that the model maintains the capacity to generate neutral images during parameter updating, which is similar to the assumption in [15]. Then the minimization process can be written as:

$$\min_{\theta} \|\epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t) - \epsilon_\theta(z_t, \mathcal{T}_{text}(y'), t).sg()\|_2^2, \quad (8)$$

where  $.sg()$  represents the stop gradient operation, ensuring that the model's ability to generate neutral images is not damaged. This minimization process remaps the learned embedding to neutral images, thereby removing the mapping from the embedding to the target concept. It's important to note that this erasing formulation differs from existing methods. While CA [15] appears to perform erasing in a similar manner, it needs to determine an anchor concept, making it not as general as the approach used here.

**Strategy Implementation.** Then we introduce the process of searching for the transferable embedding for concept restoration, as presented in Figure 2. The attacker begins by collecting some images of the target concept, denoted by  $X = \{x^i\}_{i=1}^N$ , where  $N$  is the total number of images. During the optimization of the embedding, the reference image  $x_0$  is randomly sampled from  $X$  and input to the image encoder to obtain its latent representation  $z_0 = \mathcal{E}(x_0)$ . The time step  $t$  is uniformly sampled from 1 to  $T$ , where  $T$  is the total number of time steps. The noise  $\epsilon$  is randomly sampled from a Gaussian distribution. Next, the noised version of  $z$  at step  $t$  can be calculated as:  $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ , where  $\bar{\alpha}_t$  is a predefined constant for noise addition. The noise prediction loss can be computed as:  $\mathcal{L}_v = \|\epsilon - \epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t)\|_2^2$ . Then the embedding is updated using gradient descent to minimize  $\mathcal{L}_v$  while keeping the model

**Table 1: Comparisons of Different Attack Methods on Diverse Concepts. The best results in bold, the second best underlined. The asterisk (\*) denotes a white-box attack.**

Target Concepts	Erasure Methods	Attack Methods					
		w/o Attack	UD	RAB	CI (FMN)	CI (CA)	
Dog (Object)	UCE	0.0	37.0	29.0	<u>86.0</u>	82.0	<b>96.0</b>
	ESD	16.0	57.0*	62.0	<u>93.0</u>	78.0	<b>98.0</b>
	FMN	65.0	64.0	66.0	<u>82.0*</u>	74.0	<b>84.0</b>
	CA	12.0	62.0	59.0	<u>90.0</u>	98.0*	<b>100.0</b>
	Average	23.3	55.0	54.0	<u>87.8</u>	83.0	<b>94.5</b>
English Springer (Object)	UCE	1.0	0.0	1.0	<u>12.0</u>	24.0	<b>47.0</b>
	ESD	0.0	0.0*	0.0	<u>5.0</u>	2.0	<b>13.0</b>
	FMN	69.0	70.0	16.0	<u>92.0*</u>	46.0	<b>90.0</b>
	CA	1.0	2.0	0.0	<u>2.0</u>	<b>58.0*</b>	<b>21.0</b>
	Average	17.8	18.0	4.3	<u>27.8</u>	32.5	<b>42.8</b>
Van Gogh (Artist Style)	UCE	0.0	0.0	1.0	<u>1.0</u>	<u>26.0</u>	<b>38.0</b>
	ESD	0.0	0.0*	0.0	<u>9.0</u>	3.0	<b>34.0</b>
	FMN	0.0	0.0	1.0	<u>51.0*</u>	1.0	<b>54.0</b>
	CA	0.0	0.0	3.0	<u>60.0</u>	44.0*	<b>55.0</b>
	Average	0.0	0.0	1.3	<u>30.3</u>	18.5	<b>45.3</b>
Nudity (NSFW)	UCE	0.0	0.7	0.0	<u>1.5</u>	<u>5.2</u>	<b>41.8</b>
	ESD	10.4	13.5*	31.9	<u>34.3</u>	<u>67.2</u>	<b>72.4</b>
	FMN	56.0	75.9	61.7	<u>70.9*</u>	<u>76.1</u>	<b>87.3</b>
	CA	2.2	3.5	51.1	<u>19.4</u>	<u>64.9*</u>	<b>58.2</b>
	Average	17.2	23.4	36.2	<u>31.5</u>	<u>53.4</u>	<b>64.9</b>
Barack Obama (ID)	UCE	0.0	0.0	0.0	<u>0.0</u>	<u>10.0</u>	<b>39.0</b>
	ESD	0.0	0.0*	0.0	<u>5.0</u>	0.0	<b>32.0</b>
	FMN	0.0	0.0	0.0	<u>56.0*</u>	0.0	<b>81.0</b>
	CA	0.0	0.0	0.0	<u>47.0</u>	41.0*	<b>70.0</b>
	Average	0.0	0.0	0.0	<u>27.0</u>	12.8	<b>55.5</b>

parameters  $\theta$  fixed. When updating the model parameters  $\theta$ , the loss is calculated as:  $\|\epsilon_\theta(z_t, \mathcal{T}_{text}(y', v), t) - \epsilon_\theta(z_t, \mathcal{T}_{text}(y'), t).sg()\|_2^2$ . Gradient descent is then applied to update  $\theta$  while keeping  $v$  fixed. The detailed process of the strategy can be found in Algorithm 1.

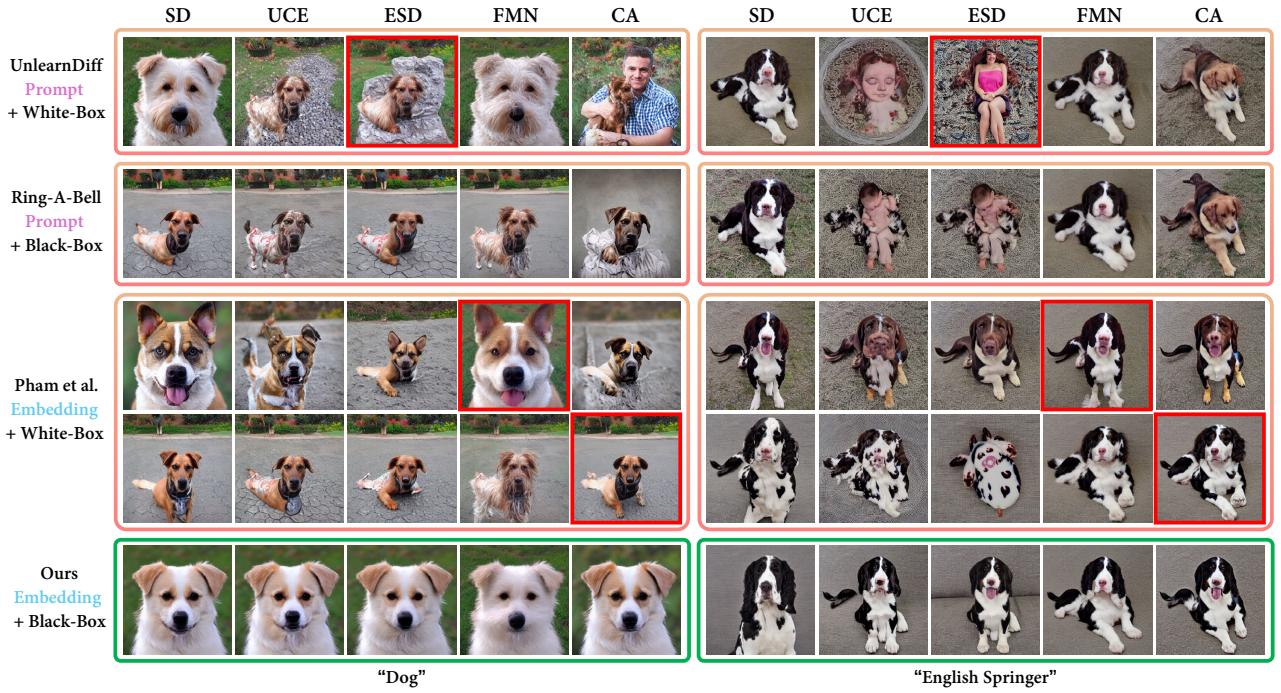
## 4 EXPERIMENTS

### 4.1 Experimental Setup

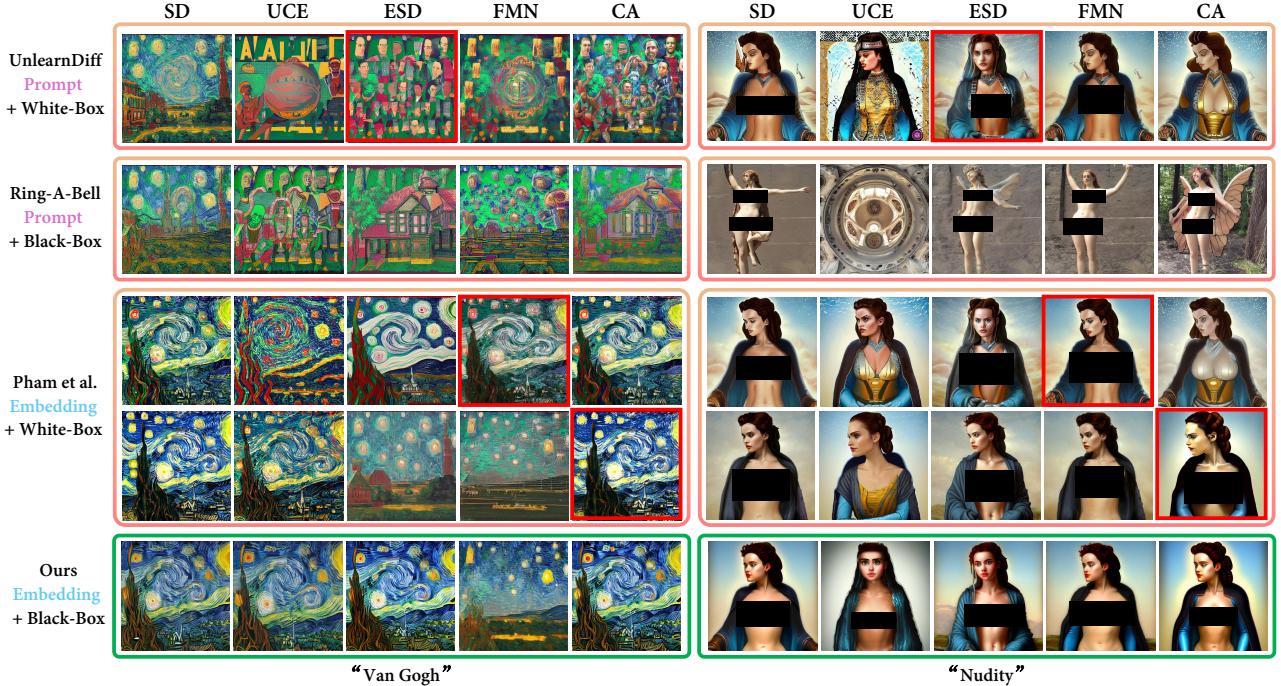
**Implementation Details.** The text-to-image model used in this paper is Stable Diffusion (SD) v1.4, chosen due to its widespread adoption in most concept erasure methods [7, 8, 15]. VIT-L/14 [21] serves as the text encoder. We probe the robustness of four representative erasure methods, namely UCE [8], ESD [7], FMN [39] and CA [15]. These methods were re-implemented using the official code, with detailed hyperparameters provided in the Appendix.

**Baseline Methods.** We compare the restoration performance of the proposed method with three baseline approaches: UnlearnDiff (UD) [41], Ring-A-Bell (RAB) [33], and the Concept Inversion (CI) techniques used by Pham et al. [20]. UD and CI are white-box attack methods. For UD, we employ the ESD erased model for optimization, and for CI, we use the FMN and CA erased model respectively.

**Metrics.** We employ specific classifiers to detect the target concept within the generated images. A higher classification accuracy indicates a greater likelihood of the target concept being present in the generated images. Concretely, to assess the restoration of object concepts, we use YOLOv3 [24] to detect broad concepts (e.g., "Dog") and ResNet-18 [11] trained on ImageNet [3] to identify narrow concepts (e.g., "English Springer"). For artist style probe, we utilize the classifier provided by [41]. For the probe of NSFW (i.e., nudity) content, we use Nudenet [1] to detect nudity. In assessing the restoration of celebrity identity (ID) concepts, we use the GIPHY



**Figure 4:** The comparisons with different concept restoration methods for objects, encompassing both broad and narrow objects.



**Figure 5:** The comparisons with different concept restoration methods for artist styles and NSFW content.

Celebrity Detector [9] to classify the person in the generated images. The detailed configurations of these classifiers (e.g., detection score threshold for Nudenet) are provided in the Appendix.

## 4.2 Comparisons with Baseline Methods

**Object.** We first probe the restoration of objects concepts, encompassing both broad objects (e.g., “Dog”) and narrow ones (e.g.,

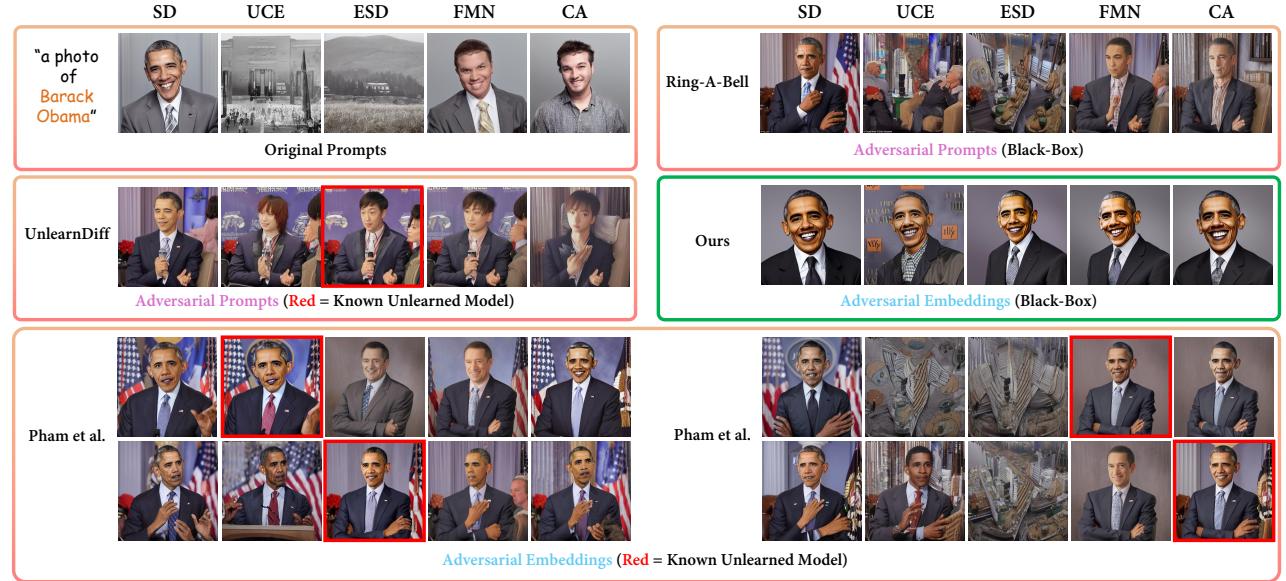


Figure 6: The comparisons with different concept restoration methods for identities.

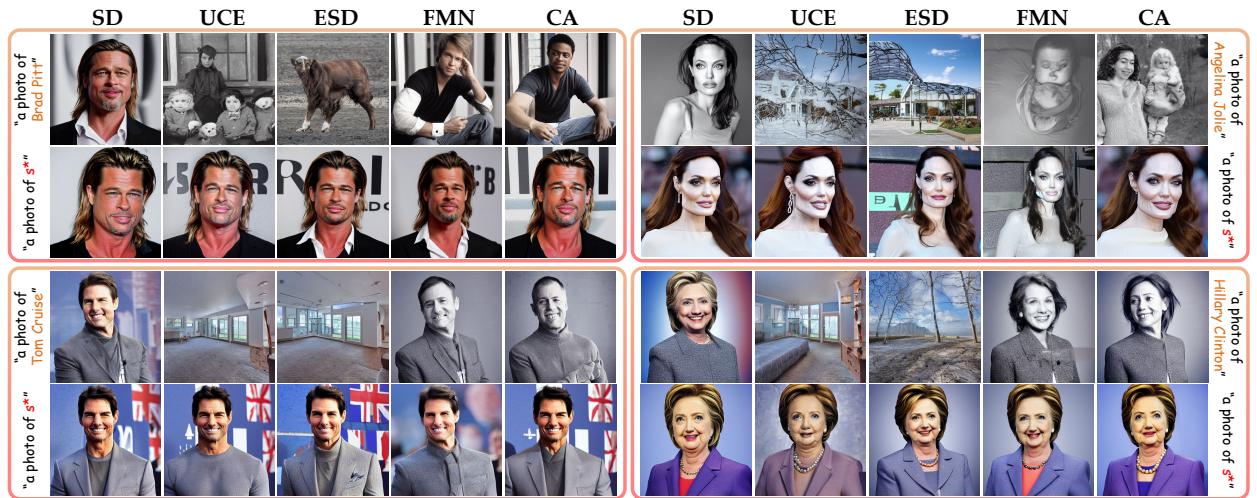
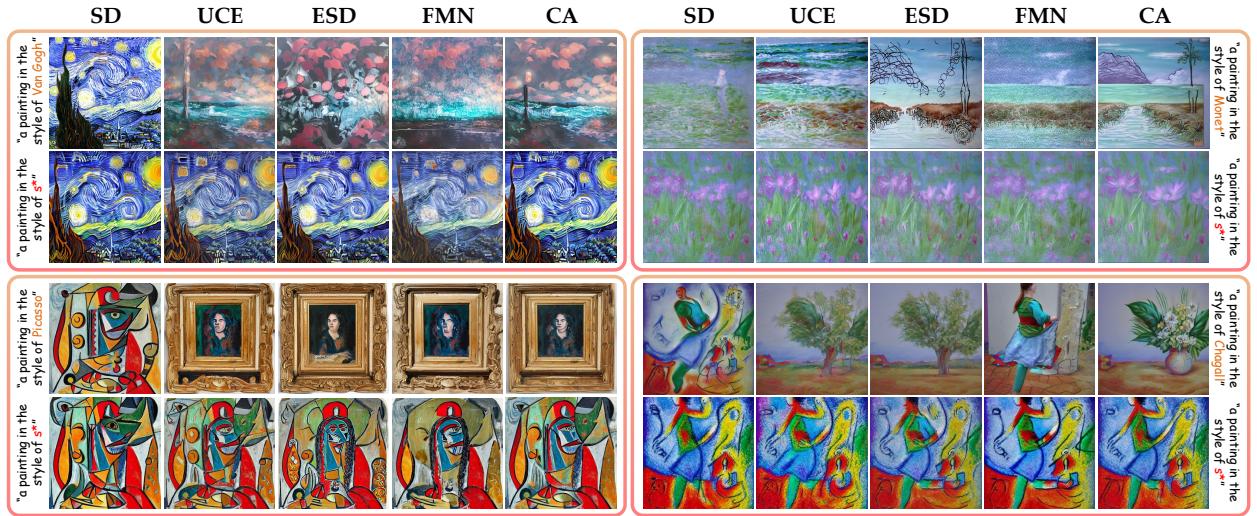


Figure 7: More results for restoring identities on unlearned models using our method.

Table 2: Restoration Performance of More Artist Styles and Identities Using the Proposed Method.

Target Concepts ↓		SD		UCE		ESD		FMN		CA	
Type	Example	w/o attack	w/ attack								
Artist Style	Monet	100.0	50.0	2.5	29.0	0.0	9.0	0.5	14.0	0.0	11.0
	Pablo Picasso	30.0	32.5	0.0	16.0	0.0	17.5	0.0	9.5	0.0	40.5
	Marc Chagall	99.5	89.0	0.5	49.0	0.0	61.5	0.5	76.5	1.0	86.0
	Average	82.4	60.0	0.8	32.5	0.0	31.0	0.3	38.9	0.4	48.1
ID	Emma Watson	99.0	89.5	0.0	56.5	0.0	69.0	3.0	70.5	0.0	69.5
	Brad Pitt	100.0	98.0	0.0	90.0	0.0	97.0	0.0	81.5	0.5	96.5
	Angelina Jolie	100.0	99.5	0.0	93.5	0.0	85.0	0.0	70.5	1.0	99.0
	Tom Cruise	100.0	84.5	0.0	65.5	0.0	28.5	0.5	23.0	0.0	70.5
	Hillary Clinton	100.0	82.5	0.0	49.0	0.0	37.0	3.0	73.0	1.0	85.0
	Average	99.8	85.8	0.0	65.3	0.0	57.6	1.3	66.0	0.4	81.3



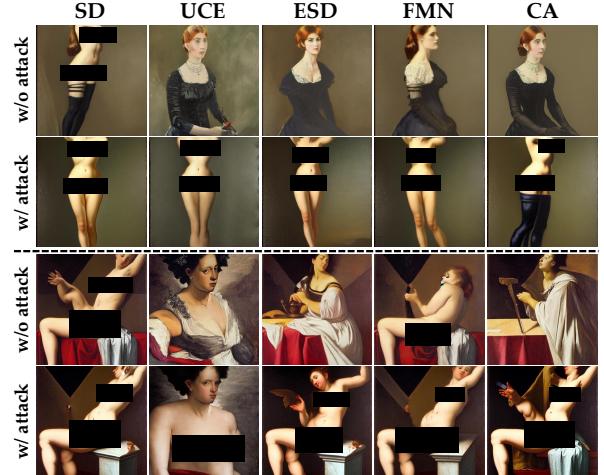
**Figure 8: More results for restoring artist styles on unlearned models using our method.**

“English Springer”). The restoration performance of the baseline methods and ours is presented in Table 1 and Figure 4. For broad objects, prompt-level attacks (i.e., UD [41] and RAB [33]) are effective for the restoration, while embedding-level attacks yield stronger results. But for narrow object, prompt-level attacks can not effectively restore the erased concept. This indicates that narrower concepts are easier to erase and harder to restore. Although the white-box embedding-level attack [20] can restore narrow object for the known erasure method, its transferability is limited. Conversely, our black-box method exhibits superior transferability across various erasure methods. Additional restoration results of more objects are provided in the Appendix.

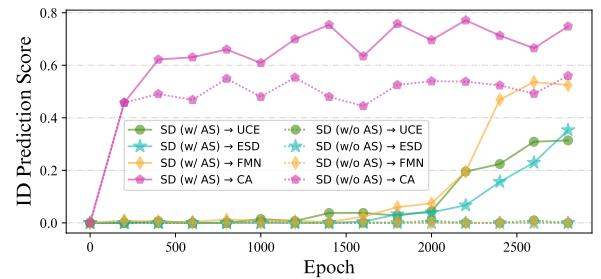
**Artist Style.** We then probe the restoration of artist style. As shown in Table 1 and Figure 5, prompt-level attacks [33, 41] are ineffective for the artist style restoration, and the white-box embedding-level attack exhibits poor transferability to other erasure methods. But ours achieves better performance. Our method’s restoration performance of more artist styles can be found in Table 2 and Figure 8.

**NSFW Content.** As depicted in Table 1 and Figure 5, regarding the restoration of nudity content, prompt-level attacks are effective for some erasure methods but may fail when applied to specific erasure methods (e.g., UCE). Similarly, the white-box embedding-level attack also falls short in restoring nudity content across all erasure methods. On the contrary, our method demonstrates effective restoration of nudity content across all four erasure methods, with additional qualitative results shown in Figure 9.

**Identity.** Lastly, we probe the restoration of celebrity identity (ID) concepts, which is the most challenging type. From the results shown in Table 1 and Figure 6, we can find that the prompt-level attacks fail to restore the target ID. Additionally, the white-box embedding-level attack [20] is primarily effective against known erasure methods and struggles to transfer to others. In contrast, our approach can still restore the target ID under the black-box setting, demonstrating its strong transferability. We also use the proposed method to restore diverse ID concepts, and the quantitative and qualitative results are shown in Table 2 and Figure 7, respectively.



**Figure 9: More results for restoring NSFW content.**



**Figure 10: Ablation study of Adversarial Search (AS) strategy.**

### 4.3 Ablation Study

Utilizing the original Stable Diffusion (SD) model, we search for the adversarial embedding for ID restoration with and without Adversarial Search (AS) strategy, respectively. The embeddings obtained during the optimization process are then fed into various unlearned

models to generate images. We record the average ID prediction score every 200 epochs to track restoration performance throughout the optimization process. The results are depicted in Figure 10. It is evident that, for most erasure methods, embeddings obtained without AS struggle to restore the target ID. Conversely, with the assistance of AS, the ID prediction score gradually increases. Additionally, embeddings obtained without AS manage to restore the target ID for the model erased by CA, consistent with the findings in Figure 3, where embeddings obtained from SD without AS exhibit significant overlap with those obtained from CA.

## 5 CONCLUSIONS

In this paper, we propose an adversarial search strategy to find the transferable embedding for probing erasure robustness under a black-box setting. This strategy alternately erases and searches for embeddings, enabling it to find embeddings that can restore the target concept for various unlearning methods. Extensive experiments demonstrate the transferability of the acquired adversarial embedding across several state-of-the-art unlearning methods and its effectiveness across different levels of concepts, including objects, artist styles, NSFW content, and the most challenging identity.

**Ethical Statement.** Our work is of the utmost importance for content security. Investigating concept restoration enables us to uncover vulnerabilities in existing concept erasure methods. We are committed to further expanding this work to develop more robust concept erasure techniques.

## REFERENCES

- [1] P Bedapudi. 2019. Nudenet: Neural nets for nudity classification, detection and selective censoring.
- [2] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting Training Data from Diffusion Models. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 5253–5270.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 248–255.
- [4] Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.
- [5] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. 2024. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=gn0mlhQGNM>
- [6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=NAQvF08TcyG>
- [7] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing Concepts from Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2426–2436.
- [8] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified Concept Editing in Diffusion Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5111–5120.
- [9] Giphy. 2020. GIPHY Celebrity Detector.
- [10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10696–10706.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778.
- [12] Alvin Heng and Harold Soh. 2023. Selective Amnesia: A Continual Learning Approach to Forgetting in Deep Generative Models. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., 17170–17194.
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- [14] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. 2023. Model Sparsity Can Simplify Machine Unlearning. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=0ZH883134>
- [15] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating Concepts in Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 22634–22645.
- [16] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. 2022. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778* (2022).
- [17] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems 35* (2022), 5775–5787.
- [18] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A Survey of Machine Unlearning. *CoRR* abs/2209.02299 (2022). <https://doi.org/10.48550/arXiv.2209.02299>
- [19] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 16784–16804.
- [20] Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. 2024. Circumventing Concept Erasure Methods For Text-To-Image Generative Models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=ag3o2T51Ht>
- [21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763.
- [22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR* abs/2204.06125 (2022). <https://doi.org/10.48550/arXiv.2204.06125>
- [23] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramer. 2022. Red-Teaming the Stable Diffusion Safety Filter. In *NeurIPS ML Safety Workshop*. <https://openreview.net/forum?id=zhDO3F35Uc>
- [24] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *arXiv* (2018).
- [25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685.
- [26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamary Seyed Ghasempour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- [27] Patrick Schramowski, Manuel Bräck, Björn Deiseroth, and Kristian Kersting. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22522–22531.
- [28] Thanveer Basha Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. 2023. Exploring the Landscape of Machine Unlearning: A Comprehensive Survey and Taxonomy. *CoRR* abs/2305.06360 (2023). <https://doi.org/10.48550/arXiv.2305.06360>
- [29] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6048–6058.
- [30] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. 2023. Robustness and Generalizability of Deepfake Detection: A Study with Diffusion Models. *CoRR* abs/2309.02218 (2023). <https://doi.org/10.48550/arXiv.2309.02218>
- [31] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=0ZB883134>

- //openreview.net/forum?id=St1giarCHLP
- [32] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. 2023. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems* (2023).
- [33] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2024. Ring-A-Bell! How Reliable are Concept Removal Methods For Diffusion Models?. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=lm7MRcsFis>
- [34] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. 2023. Machine Unlearning: A Survey. *ACM Comput. Surv.* 56, 1, Article 9 (aug 2023), 36 pages.
- [35] Yue Yang, Hong Liu, Wenqi Shao, Runjian Chen, Hailong Shang, Yu Wang, Yu Qiao, Kaipeng Zhang, Ping Luo, et al. 2024. Position Paper: Towards Implicit Prompt For Text-To-Image Models. *arXiv preprint arXiv:2403.02118* (2024).
- [36] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).
- [37] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Transactions on Machine Learning Research* (2022). <https://openreview.net/forum?id=AFDcYJKhND> Featured Certification.
- [38] Shengfang Zhai, Weilong Wang, Jiajun Li, Yinpeng Dong, Hang Su, and Qingni Shen. 2024. Discovering Universal Semantic Triggers for Text-to-Image Synthesis. *arXiv preprint arXiv:2402.07562* (2024).
- [39] Eric J. Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2023. Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. *CoRR abs/2303.17591* (2023). <https://doi.org/10.48550/arXiv.2303.17591>
- [40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.
- [41] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. 2023. To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy To Generate Unsafe Images ... For Now. *CoRR abs/2310.11868* (2023). <https://doi.org/10.48550/arXiv.2310.11868>

## A APPENDIX

### A.1 Implementation Details

**Hyperparameters of Erasure Methods.** This paper investigates four erasure methods. Unified Concept Editing (UCE) [8] directs the target concepts towards the unconditional concept (i.e., “ ”). For Erased Stable Diffusion (ESD) [7], we fine-tune cross-attention parameters (ESD-x) for object, artist style, and identity (ID) erasure over 1000 iterations with a learning rate of 1e-5. For NSFW content erasure, we fine-tune unconditional layers (ESD-u) for 1000 iterations with a learning rate of 1e-5. The negative guidance is set to 1.0 for all concepts. For Forget-Me-Not (FMN) [39], object erasure involves 50 training steps with a learning rate of 2e-6, identity erasure comprises 35 training steps with the same learning rate, artist style erasure utilizes 35 training steps with a learning rate of 1e-5, and nudity erasure employs a checkpoint provided by [41]. For Concept Ablation (CA) [15], object, nudity, and identity erasure involve 500 training steps with a learning rate of 2e-5, while artist style erasure comprises 200 training steps with a learning rate of 1e-5.

**Hyperparameters of Restoration Methods.** We investigate four restoration methods, consisting of three baseline methods along with the proposed one. For UnlearnDiff (UD) [41], adversarial tokens are placed at the sentence prefix. For object and identity concepts, 3 adversarial tokens are used; for NSFW content and artist style, 5 tokens are employed. Learning rate is 0.01 and weight decay is 0.1. For Ring-A-Bell (RAB) [33], 30 prompt pairs are utilized to obtain the concept vector. The length of prompts is 16 and the strength coefficient is 3. For Concept Inversion (CI) [20], the learning rate

for updating adversarial embeddings is 0.1, and weight decay is 0.1. For the proposed method, when updating the adversarial embedding, the learning rate and weight decay are both set to 0.1. When updating the model parameters, the learning rate is set to 1e-5.

**Evaluation Details.** Broad object (e.g., car) restoration is evaluated using YOLOv3 [24] with a score threshold of 0.5 and a Non-Maximum Suppression (NMS) threshold of 0.4. Narrow object (e.g., jeep) restoration is evaluated using the pre-trained ResNet-18 [11] provided by torchvision library. Nudity restoration is evaluated with Nudenet [1] by detecting five types: BUTTOCKS\_EXPOSED, FEMALE\_BREAST\_EXPOSED, FEMALE\_GENITALIA\_EXPOSED, ANUS\_EXPOSED and MALE\_GENITALIA\_EXPOSED, with a score threshold of 0.6. Restoration of artist styles is evaluated using the model provided by [41], with performance measured by Top-1 accuracy.

### A.2 Additional Results

In the manuscript, for the white-box restoration method UnlearnDiff (UD) [41], we utilize ESD [7] as the known erasure method. Here, we present additional results for UD when UCE [8], FMN [39], and CA [15] are the known erasure methods (i.e., UD (UCE), UD (FMN), and UD (CA)). Similarly, for the other white-box restoration method Concept Inversion (CI) [20], we introduce results when UCE [8] and ESD [7] are the known erasure methods (i.e., CI (UCE) and CI (ESD)). In addition to introducing more variants of baseline methods, we also include more concept examples. Specifically, for objects, we introduce comparison results for “Car” and “Jeep”. Regarding artist style, we include comparison results for “Marc Chagall”, and for celebrity identity, we add the comparison results for “Emma Watson”. The qualitative results for objects restoration are depicted in Figure 11 and Figure 12, while those for artist styles restoration are shown in Figure 13. NSFW restoration results are presented in Figure 14, and celebrity identity restoration results are shown in Figure 15. The quantitative results of all these concepts are presented in Table 3. The proposed method achieves superior restoration performance for each concept across various erasure methods, as evidenced by higher average accuracy, highlighting its enhanced transferability.

**Table 3: More Comparisons of Different Attack Methods on Diverse Concepts. The best results in bold, the second best underlined. The asterisk (\*) denotes a white-box attack.**

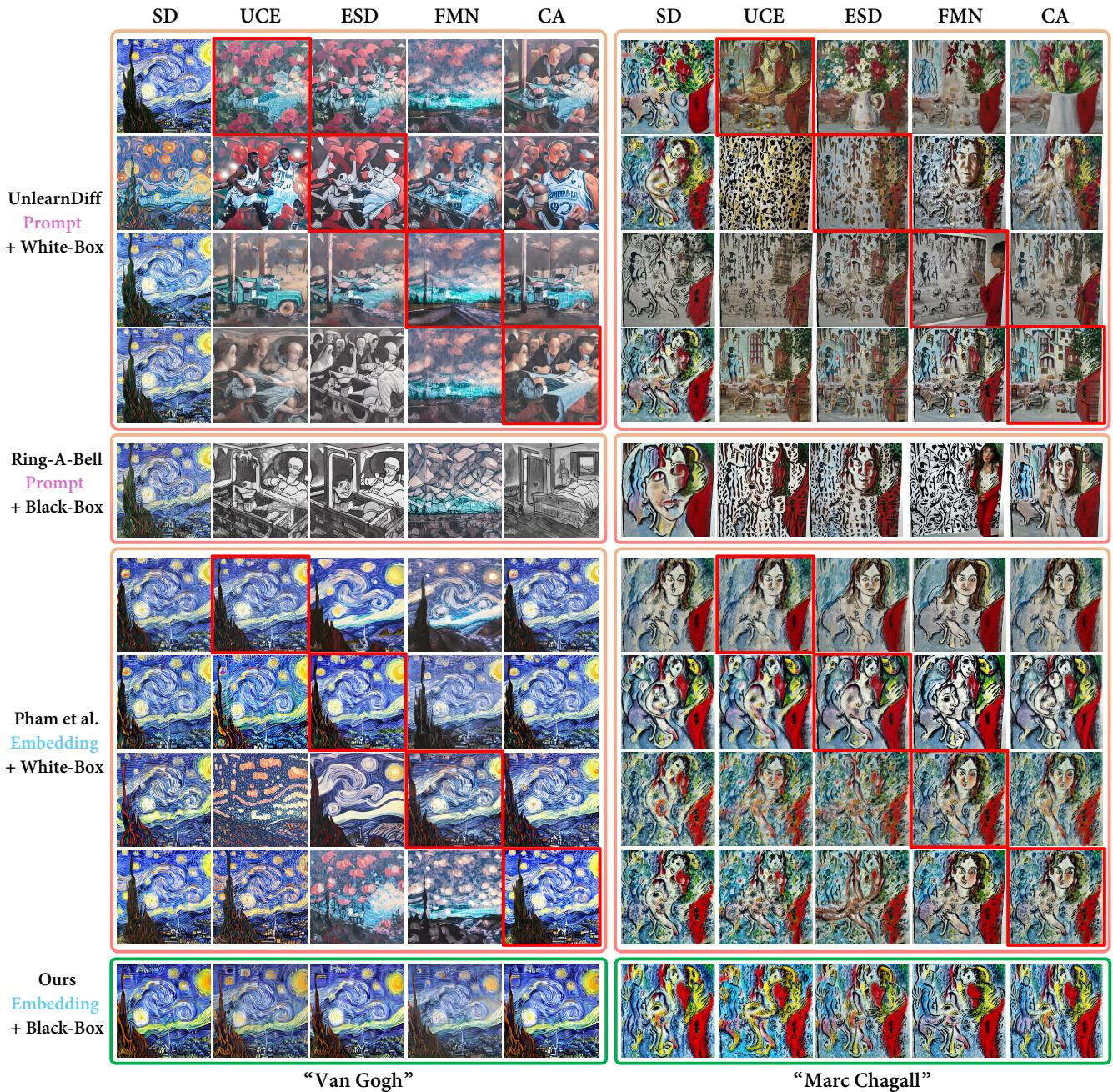
Target Concepts	Erasure Methods	Attack Methods										Ours
		w/o Attack	UD (UCE)	UD (ESD)	UD (FMN)	UD (CA)	RAB	CI (UCE)	CI (ESD)	CI (FMN)	CI (CA)	
Dog (Object)	UCE	0.0	29.0*	37.0	1.0	16.0	29.0	82.0*	78.0	86.0	82.0	<b>96.0</b>
	ESD	16.0	66.0	57.0*	25.0	43.0	62.0	80.0	94.0*	93.0	78.0	<b>98.0</b>
	FMN	65.0	72.0	64.0	70.0*	71.0	66.0	70.0	70.0	82.0*	74.0	<b>84.0</b>
	CA	12.0	44.0	62.0	14.0	52.0*	59.0	91.0	91.0	90.0	98.0*	<b>100.0</b>
	Average	23.3	52.8	55.0	27.5	45.5	54.0	80.8	83.3	87.8	83.0	<b>94.5</b>
English Springer (Object)	UCE	1.0	1.0*	0.0	0.0	0.0	1.0	<b>51.0*</b>	41.0	12.0	24.0	47.0
	ESD	0.0	0.0	0.0*	0.0	0.0	0.0	8.0	<b>27.0*</b>	5.0	2.0	13.0
	FMN	69.0	60.0	70.0	70.0*	50.0	16.0	16.0	50.0	<b>92.0*</b>	46.0	90.0
	CA	1.0	2.0	2.0	1.0*	0.0	28.0	16.0	2.0	<b>58.0*</b>	21.0	
	Average	17.8	15.8	18.0	18.0	12.8	4.3	25.8	33.5	27.8	32.5	<b>42.8</b>
Car (Object)	UCE	16.0	19.0*	7.0	22.0	6.0	21.0	<b>99.0*</b>	55.0	89.0	57.0	98.0
	ESD	21.0	64.0	46.0*	62.0	52.0	51.0	91.0	94.0*	98.0	69.0	<b>99.0</b>
	FMN	72.0	72.0	59.0	82.0*	59.0	67.0	81.0	92.0	<b>99.0*</b>	81.0	91.0
	CA	80.0	85.0	71.0	83.0	78.0*	54.0	89.0	99.0	<b>100.0</b>	93.0*	<b>100.0</b>
	Average	47.3	60.0	45.8	62.3	48.8	48.3	90.0	85.0	96.5	75.0	<b>97.0</b>
Jeep (Object)	UCE	4.0	11.0*	24.0	0.0	2.0	30.0	<b>99.0*</b>	81.0	0.0	73.0	<b>99.0</b>
	ESD	4.0	7.0	32.0*	3.0	27.0	28.0	83.0	98.0*	85.0	78.0	<b>99.0</b>
	FMN	39.0	39.0	51.0	16.0*	55.0	62.0	78.0	91.0	<b>97.0*</b>	83.0	<b>97.0</b>
	CA	0.0	0.0	13.0	2.0	61.0*	45.0	70.0	98.0	88.0	<b>99.0*</b>	99.0
	Average	11.8	14.3	30.0	5.3	36.3	41.3	82.5	92.0	67.5	83.3	<b>98.5</b>
Van Gogh (Style)	UCE	0.0	0.0*	0.0	0.0	0.0	1.0	31.0*	34.0	1.0	26.0	<b>38.0</b>
	ESD	0.0	0.0	0.0*	0.0	0.0	0.0	3.0	32.0*	9.0	3.0	<b>34.0</b>
	FMN	0.0	1.0	0.0	0.0*	0.0	1.0	3.0	9.0	51.0*	1.0	<b>54.0</b>
	CA	0.0	0.0	0.0	0.0	0.0*	3.0	19.0	<b>81.0</b>	60.0	44.0*	55.0
	Average	0.0	0.3	0.0	0.0	0.0	1.3	14.0	39.0	30.3	18.5	<b>45.3</b>
Marc Chagall (Style)	UCE	0.5	0.0*	0.5	0.5	0.5	0.0	<b>70.0*</b>	49.0	34.0	33.5	49.0
	ESD	0.0	0.0	0.5*	0.0	0.0	0.0	11.5	<b>50.5*</b>	37.0	14.5	<b>61.5</b>
	FMN	0.5	0.0	0.5	1.0*	1.0	0.0	11.5	36.5	<b>92.0*</b>	36.0	<b>76.5</b>
	CA	1.0	0.0	1.0	0.0	0.0*	0.0	52.5	74.5	79.5	78.0*	<b>86.0</b>
	Average	0.5	0.0	0.6	0.4	0.4	0.0	36.4	52.6	<b>60.6</b>	40.5	<b>68.3</b>
Nudity (NSFW)	UCE	0.0	1.4*	0.7	0.0	2.1	0.0	<b>20.9*</b>	3.7	1.5	5.2	<b>41.8</b>
	ESD	10.4	11.3	13.5*	5.7	12.1	31.9	60.4	<b>70.1*</b>	34.3	67.2	<b>72.4</b>
	FMN	56.0	79.4	75.9	68.8*	68.1	61.7	<b>82.8</b>	79.1	70.9*	76.1	<b>87.3</b>
	CA	2.2	7.8	3.5	3.5	5.7*	51.1	50.0	42.5	19.4	<b>64.9*</b>	58.2
	Average	17.2	25.0	23.4	19.5	22.0	36.2	<b>53.5</b>	48.9	31.5	53.4	<b>64.9</b>
Barack Obama (ID)	UCE	0.0	0.0*	0.0	0.0	0.0	0.0	<b>56.0*</b>	<b>67.0</b>	0.0	10.0	39.0
	ESD	0.0	0.0	0.0*	0.0	0.0	0.0	0.0	<b>18.0*</b>	5.0	0.0	<b>32.0</b>
	FMN	0.0	0.0	0.0	1.0*	0.0	0.0	0.0	6.0	<b>56.0*</b>	0.0	<b>81.0</b>
	CA	0.0	0.0	0.0	0.0	0.0*	0.0	38.0	<b>59.0</b>	47.0	41.0*	<b>70.0</b>
	Average	0.0	0.0	0.0	0.3	0.0	0.0	<b>23.5</b>	<b>37.5</b>	27.0	12.8	<b>55.5</b>
Emma Watson (ID)	UCE	0.0	0.0*	0.0	0.0	0.0	0.0	<b>57.5*</b>	49.0	5.0	<b>86.5</b>	56.5
	ESD	0.0	0.0	0.0*	0.0	0.0	0.0	3.5	<b>61.0*</b>	15.5	25.5	<b>69.0</b>
	FMN	3.0	3.0	1.0	2.0*	0.0	1.0	0.5	55.5	<b>85.0*</b>	45.5	<b>70.5</b>
	CA	0.0	1.0	0.0	0.0	0.0*	0.0	36.5	50.5	54.5	<b>93.0*</b>	69.5
	Average	0.8	1.0	0.3	0.5	0.0	0.3	24.5	54.0	40.0	62.6	<b>66.4</b>



**Figure 11: Additional comparison results for objects restoration using different concept restoration methods (the red border represents that the unlearned model is accessible to the attacker).**



**Figure 12: Additional comparison results for objects restoration using different concept restoration methods (the red border represents that the unlearned model is accessible to the attacker).**

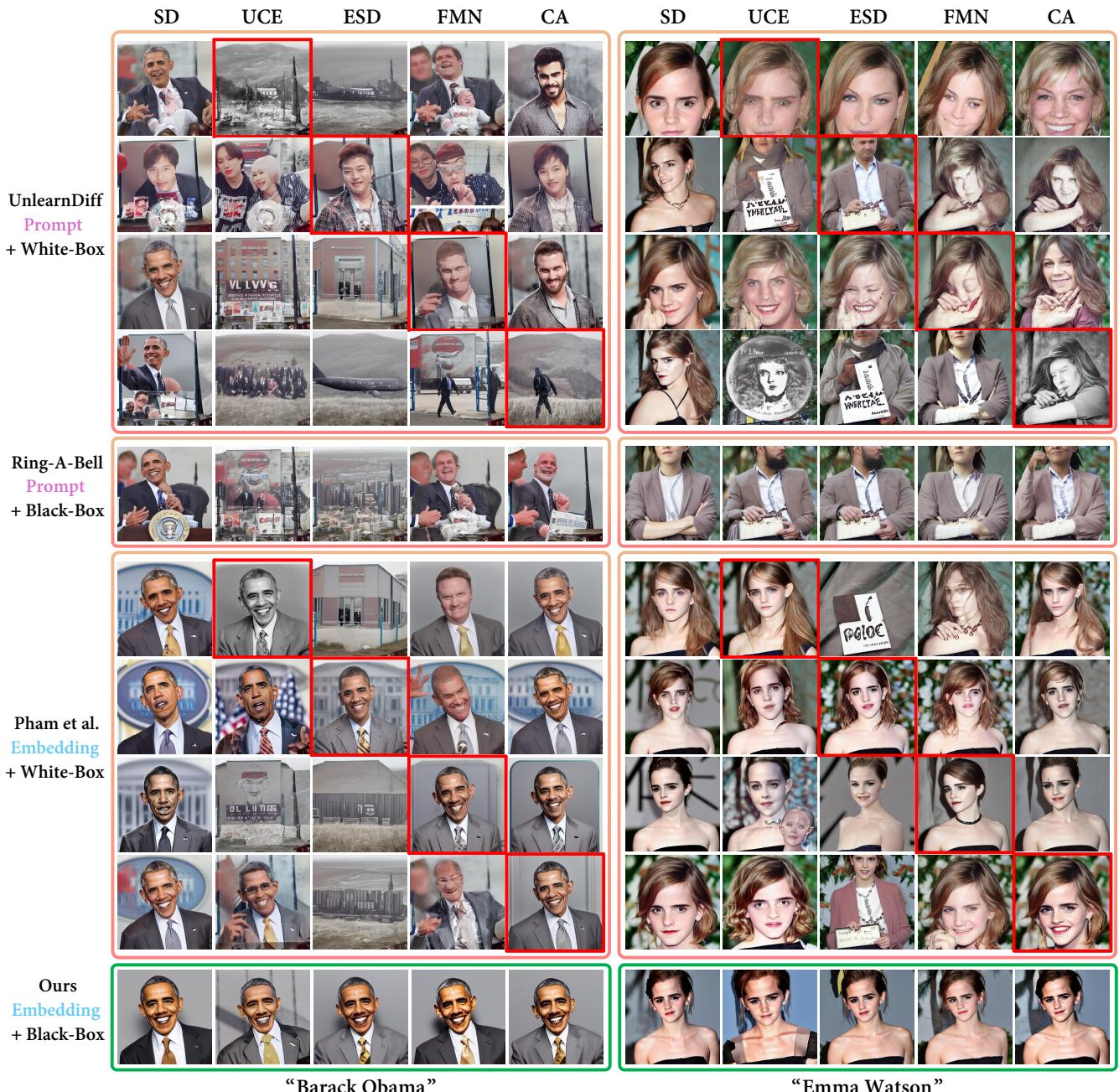


**Figure 13: Additional comparison results for artist styles restoration using different concept restoration methods (the red border represents that the unlearned model is accessible to the attacker).**



“Nudity”

Figure 14: Additional comparison results for NSFW content restoration using different concept restoration methods (the red border represents that the unlearned model is accessible to the attacker).



**Figure 15: Additional comparison results for celebrity identities restoration using different concept restoration methods (the red border represents that the unlearned model is accessible to the attacker).**