

# ZIUM: Zero-Shot Intent-Aware Adversarial Attack on Unlearned Models

Hyun Jun Yook<sup>1</sup> Ga San Jhun<sup>1</sup> Jae Hyun Cho<sup>1</sup> Min Jeon<sup>1</sup>

Donghyun Kim<sup>2</sup> Tae Hyung Kim<sup>3</sup> Youn Kyu Lee<sup>1,†</sup>

<sup>1</sup>Chung-Ang University <sup>2</sup>Korea University <sup>3</sup>Hongik University

{hyunjun6, bonuspoint, wogus2031, mulsoap0504, younkyul}@cau.ac.kr

d.kim@korea.ac.kr taehyung@hongik.ac.kr

## Abstract

Machine unlearning (MU) removes specific data points or concepts from deep learning models to enhance privacy and prevent sensitive content generation. Adversarial prompts can exploit unlearned models to generate content containing removed concepts, posing a significant security risk. However, existing adversarial attack methods still face challenges in generating content that aligns with an attacker’s intent while incurring high computational costs to identify successful prompts. To address these challenges, we propose ZIUM, a Zero-shot Intent-aware adversarial attack on Unlearned Models, which enables the flexible customization of target attack images to reflect an attacker’s intent. Additionally, ZIUM supports zero-shot adversarial attacks without requiring further optimization for previously attacked unlearned concepts. The evaluation across various MU scenarios demonstrated ZIUM’s effectiveness in successfully customizing content based on user-intent prompts while achieving a superior attack success rate compared to existing methods. Moreover, its zero-shot adversarial attack significantly reduces the attack time for previously attacked unlearned concepts.

## 1. Introduction

Machine Unlearning (MU) selectively removes specific data points or features from a trained deep learning model through reweighting or pruning [20, 41]. It helps protect privacy and prevent the generation of sensitive content [22, 24, 37]. Recently, MU has been used to prevent the generation of inappropriate images by eliminating concepts such as nudity and violence—commonly associated with NSFW (i.e., Not Safe For Work) content—from pre-trained text-to-image generation models [11, 17, 25, 29, 45]. However, even after MU is applied, these models (i.e., unlearned models) can still generate images of the removed concepts (i.e., unlearned concepts) when given adversarial prompts, posing a significant security risk [29, 31, 38].

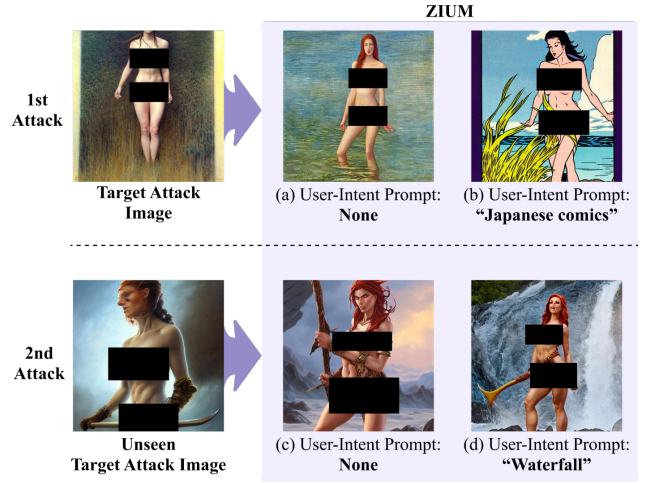


Figure 1. Examples of generated images by ZIUM: 1st adversarial attack utilizing the user-intent prompt and 2nd adversarial attack without additional optimization for the same unlearned concept.

Several adversarial attack methods targeting unlearned models have been proposed to exploit this. Specifically, approaches [5, 12, 23, 30, 47] have been developed to identify optimal adversarial prompts based on a target attack image containing an unlearned concept. These methods generate an image that incorporates the unlearned concept while closely resembling the target attack image by using the optimal adversarial prompt as input to the unlearned model. However, these methods heavily rely on the target attack image, making it challenging to generate an image that reflects both unlearned concepts and the attacker’s intent (e.g., preferences and background). This is crucial for generating an image that not only includes the unlearned concepts but also aligns with the attacker’s intentions in different forms. Recent approaches [12, 26, 30, 39, 44] have attempted to align with the attacker’s intent using only a target prompt. However, without a target attack image, fully representing a single image in text form is challenging, and the attacker’s prompt may lack sufficient semantic detail [17]. Therefore,

a new attack method is required to exploit unlearned models, generating customized images in various forms while accurately reflecting the attacker’s intent. Moreover, when an attacker aims to generate images with varying contexts (e.g., adding a new concept like “Japanese comics” as shown in Fig. 1(b)), existing approaches require repeatedly identifying optimal adversarial prompts for each context, making the attack process costly [23, 37, 41]. Therefore, a zero-shot attack mechanism is required to eliminate the need for additional optimization steps for the same unlearned concept.

To address these challenges, we propose ZIUM, a novel adversarial attack method that enables attackers to customize target attack images based on their intent while supporting zero-shot adversarial attacks. Our approach exploits an unlearned model to generate images that closely resemble the target attack image while embedding the unlearned concept. By incorporating user-intent prompts, the generated images can be precisely tailored to align with the attacker’s intent. Furthermore, our method enables zero-shot adversarial attacks, eliminating the need for additional optimization processes for previously attacked unlearned concepts. Fig. 1 illustrates an adversarial attack using ZIUM. In the first attack trial, the generated image (a) closely resembles the target attack image while embedding the unlearned concept (i.e., nudity) without a user-intent prompt, whereas image (b) reflects the user-intent prompt “Japanese comics.” In the second attack trial, targeting the same unlearned concept, a zero-shot adversarial attack was applied using the module optimized during the first attack. As a result, image (c) resembles the unseen target attack image while embedding the same unlearned concept without a user-intent prompt. Moreover, image (d) successfully reflects the user-intent prompt “Waterfall.”

ZIUM addresses the challenge of insufficient semantic detail by utilizing both the unlearned concept from the target attack image and the user-intent prompt that reflects the attacker’s intent. To achieve this, ZIUM employs an image captioning technique that converts image embeddings into text embeddings. First, it extracts the visual embedding of the target attack image containing the unlearned concept and transforms it to the text embedding of the unlearned model. This embedding is then fed into the unlearned diffusion model along with the text embedding of the user-intent prompt to generate an image that aligns with the attacker’s intent. Furthermore, image captioning techniques enable zero-shot adaptation to specific trained concepts within an image. Leveraging this capability, ZIUM facilitates additional attacks without requiring further optimization for previously attacked unlearned concepts.

We evaluated ZIUM on representative unlearned models, including ESD [11], FMN [45], SLD [36], and AdvUnlearn [46]. Our method outperforms existing adversarial

attack methods across various unlearned concept scenarios (e.g., nudity, violence, illegal activity, style, and object), achieving a significantly higher attack success rate (ASR) on average—improving by at least 22.6%p and up to 62.0%p. Furthermore, experiments with diverse user-intent prompts demonstrate that ZIUM effectively generates images that accurately align with the attacker’s intent. Notably, ZIUM maintains a high ASR without requiring additional optimization for the same unlearned concept.

Our contributions are summarized as follows:

1. Proposal of a novel adversarial attack method for unlearned models that effectively reflects various attacker intents while achieving a higher attack success rate.
2. Design of a zero-shot adversarial attack mechanism that enables targeting the same unlearned concept without requiring additional optimization.
3. A comprehensive evaluation demonstrating the effectiveness of ZIUM across different unlearned models and unlearned concept scenarios.

This paper is structured as follows: Section 2 reviews related works, Section 3 introduces ZIUM, Section 4 presents the experimental results, and Section 5 concludes the paper.

## 2. Related Works

### 2.1. Machine Unlearning

Generative models can produce large amounts of potentially inappropriate content (e.g., nudity, copyright infringement), increasing the need for effective constraints. To address this issue, Machine Unlearning (MU) has been actively studied. MU is a data removal mechanism designed to eliminate the influence of undesirable data points without requiring costly retraining while preserving model performance for inputs unrelated to the removed data [2]. Gandikota et al. [11] proposed Erased Stable Diffusion (ESD), which removes specific concepts without requiring additional training data by fine-tuning the weights of the stable diffusion model. Fan et al. [10] introduced Saliency Unlearning (SalUn), which modifies only a subset of the model’s weights rather than the entire network. Jia et al. [15] incorporated model sparsity which involves reducing unimportant data or parameters through targeted operations. In this approach, weight pruning was applied to enhance both the efficiency and performance of the unlearning process. Zhang et al. [45] introduced Forget-Me-Not (FMN), which minimizes the attention map between text and images to facilitate unlearning in text-to-image models [33]. Schramowski et al. [36] proposed Safe Latent Diffusion (SLD), which extends the existing Classifier-Free Guidance [14] to align text prompt inputs while preventing the generation of images containing unlearned concepts. Despite these advancements, existing MU mechanisms still have limitations in fully unlearning concepts. In particular, unlearned models can still generate



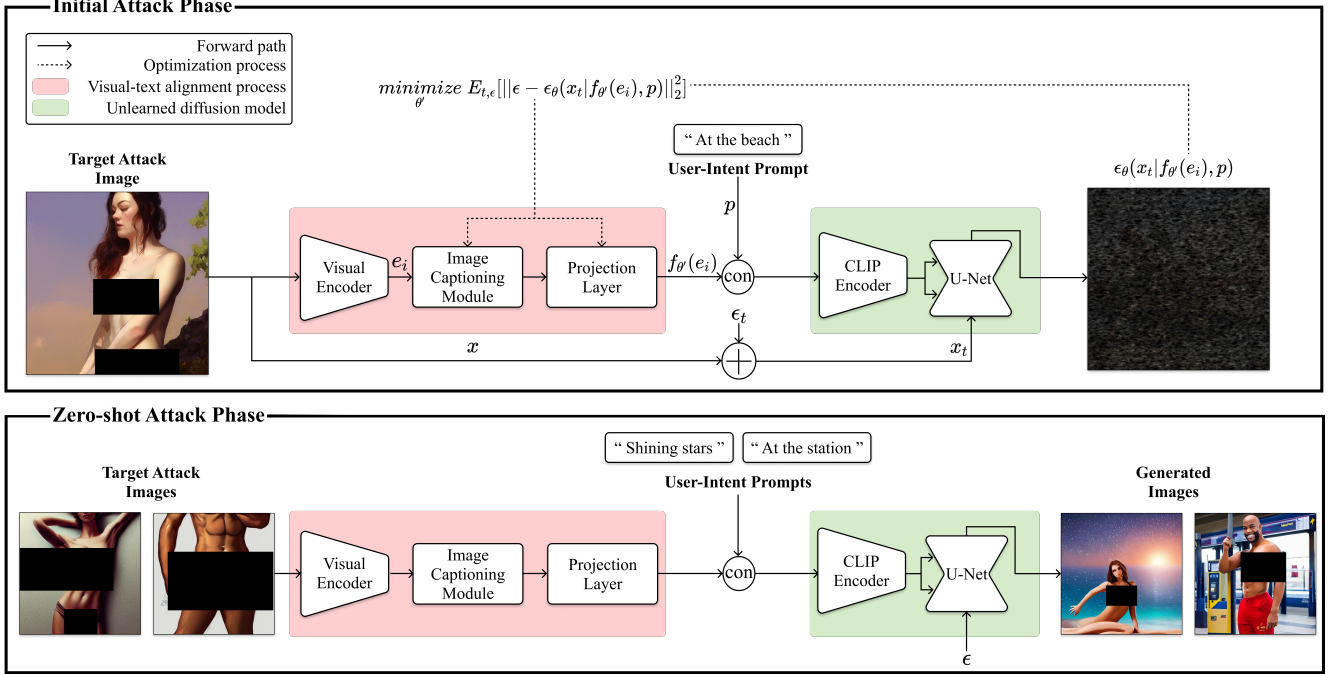


Figure 2. An overview of the ZIUM’s initial attack phase and zero-shot attack phase.

images containing concepts that were supposed to be removed. These limitations have become even more evident through adversarial attacks on unlearned models [5, 47].

## 2.2. Adversarial Attacks on Machine Unlearning

Several methods leveraging adversarial attacks on MU have been proposed [6, 12, 27, 28, 43]. Tsai et al. [39] proposed Ring-A-Bell, a method that identifies optimal adversarial prompts in a black-box setting. It generates concept vectors by computing the difference in embeddings between prompts that contain adversarial concepts and those that do not. Ma et al. [26] introduced Jailbreaking Prompt Attack (JPA), an adversarial attack method that bypasses MU mechanisms. This approach optimizes adversarial prompts by obtaining an unlearned embedding derived from the difference in embeddings between an unlearned concept and its antonym. Zhang et al. [47] proposed UnlearnDiffAtk, which identifies optimal adversarial prompts to evaluate the robustness of MU mechanisms in diffusion models. It executes attacks using the diffusion classifier inherent in the diffusion model itself, eliminating the need for auxiliary models. Chin et al. [5] proposed Prompting4Debugging (P4D), which evaluates the robustness of MU mechanisms. This approach utilizes prompt engineering to identify optimal adversarial prompts that bypass these mechanisms. Existing methods identify an optimal adversarial prompt based on a given target prompt or target attack image. However, incorporating the attacker’s intent into the target prompt or obtaining target attack images that capture both the at-

tacker’s intent and the unlearned concept remains a challenge. As a result, adversarial attacks on unlearned models may not fully align with the attacker’s intended form.

## 3. Method

In this study, ZIUM performs the attack by identifying the optimal adversarial condition that enables the generation of an image incorporating both the unlearned concept and the attacker’s intent, using the target attack image and user-intent prompt.

As shown in Fig. 2, ZIUM comprises an initial attack phase and a zero-shot attack phase, which utilizes target attack images and user-intent prompts to perform adversarial attacks. The target attack image includes unlearned concepts from the diffusion model, such as nudity, and the user-intent prompt includes the attacker’s intent to customize the target attack image, such as “At the beach,” “Shining stars,” and “At the station.” The initial attack phase performs the adversarial attack by utilizing the target attack image and user-intent prompt to identify optimal adversarial condition through the “visual-text alignment process” and the “optimization process.” Subsequently, the zero-shot attack phase only exploits the visual-text alignment process optimized through the initial attack phase to perform adversarial attacks without further optimization process. The details will be described in the following.

### 3.1. Initial Attack Phase

#### 3.1.1. Visual-text Alignment Process

The visual-text alignment process in the initial attack phase transforms the visual embedding of the target attack image into text embedding to utilize the unlearned concept of the target attack image as a condition for the unlearned diffusion model. The visual-text alignment process consists of a visual encoder, an image captioning module, and a projection layer. The visual encoder extracts key features of the target attack image that embed the unlearned concept, representing them as a fixed  $k$ -dimensional visual embedding. The extracted  $k$ -dimensional embedding is fed into the image captioning module. The image captioning module based on the cross-attention mechanism was pre-trained with Image-Text Contrastive Learning (ITC) loss, Image-Text Matching (ITM) loss, and Image-grounded Text Generation (ITG) loss [4, 9, 19, 21]. ITC loss and ITM loss maximizes the similarity between the visual embedding and the text embedding when an image and text are paired, ensuring that the information contained in each embedding is aligned. ITG loss utilizes the next text token prediction through an attention mask, allowing the image captioning module to learn text embeddings that can generate information about the input image. With these losses, the pre-trained image captioning module transforms the main features of the input visual embedding into aligned text embedding. Therefore, the transformed text embedding contains key visual embedding information about the input image (used as the target attack image). Subsequently, the projection layer projects the dimensionality of the transformed text embedding into  $L$ -dimensions to match the input size of the CLIP text encoder used in the unlearned diffusion model [33]. The  $L$ -dimensional text embedding from this process is concatenated with the text embedding of the user-intent prompt to serve as a condition for the unlearned diffusion model.

#### 3.1.2. Optimization Process

The optimization process of the initial attack phase performs the attack by identifying an optimal adversarial condition that enables image generation containing the unlearned concept. Specifically, the optimization process is based on the diffusion classifier mechanism [3, 18, 47]. The diffusion classifier mechanism utilizes Bayes' rule to estimate the condition that can generate a desired target image. By Bayes' rule, the probability that an image  $x$  is generated given a certain condition  $c_i$  is expressed as follows,

$$p_\theta(c_i|x) = \frac{p(c_i)p_\theta(x|c_i)}{\sum_j p(c_j)p_\theta(x|c_j)} \quad (1)$$

In Eq. 1,  $\theta$  denotes the parameters of the diffusion model,  $x$  denotes the image we want to generate via diffusion, and  $c$  denotes the condition we want to estimate. Thus,  $p_\theta(x|c_i)$  is

the probability of an image being generated by a condition, and,  $p(c)$  is the prior probability distribution of the condition. In general, diffusion model assumes no prior information of a particular condition  $c$ , so the prior probability  $p(c)$  can be approximated by a uniform distribution. Applying this, Eq. 1 simplifies as follows,

$$p_\theta(c_i|x) = \frac{p_\theta(x|c_i)}{\sum_j p_\theta(x|c_j)} \quad (2)$$

In a diffusion model,  $p_\theta(x|c_i)$  is proportional to the accuracy of the denoising process at timestep  $t$ . Based on this, it is expressed as follows,

$$p_\theta(c_i|x) \propto \exp(-E_{t,\epsilon} [\|\epsilon - \epsilon_\theta(x_t|c_i)\|_2^2]) \quad (3)$$

In Eq. 3,  $x_t$  is the sum of  $x$  and the noise  $\epsilon_t$ , which represents the noisy image generated at a specific timestep  $t$ , and  $\epsilon_\theta(x_t|c_i)$  is the noise predicted by diffusion given  $x_t$  and condition  $c_i$ . Therefore, it is possible to maximize the probability that the desired image  $x$  is generated through a specified condition  $c_i$ . As a result, the final optimization process for the diffusion classifier mechanism can be conducted as follows,

$$\underset{c_i}{\text{minimize}} E_{t,\epsilon} [\|\epsilon - \epsilon_\theta(x_t|c_i)\|_2^2] \quad (4)$$

Based on Eq. 4, ZIUM's optimization process utilizes both the target attack image and user-intent prompt to estimate the optimal adversarial condition  $c_i$ . First, the visual encoder  $\mathcal{E}(\cdot)$  of the visual-text alignment process which extracts the visual embedding  $e_i$  from the target attack image  $x_i$  is expressed as follows,

$$e_i = \mathcal{E}(x_i) \quad (5)$$

Let  $f_{\theta'}(\cdot)$  be the network consisting of an image captioning module and a projection layer that converts the extracted visual embedding  $e_i$  into a text embedding. Then, the condition  $c_i$  can be expressed as follows,

$$c_i = f_{\theta'}(e_i) \quad (6)$$

In Eq. 6,  $\theta'$  refers to the parameters of the image captioning module and the projection layer. Then, the condition  $c_i$  is concatenated with the text embedding  $p$  of the user-intent prompt and then processed through the CLIP text encoder of the unlearned diffusion model. This results in the condition  $c_i, p$ , which reflects the unlearned concept from the target attack image and the attacker's intent from the user-intent prompt. Therefore, Eq. 4 is expressed as follows,

$$\underset{\theta'}{\text{minimize}} E_{t,\epsilon} [\|\epsilon - \epsilon_\theta(x_t|f_{\theta'}(e_i), p)\|_2^2] \quad (7)$$

In Eq. 7, note that in the process of transforming the extracted visual embedding into a text embedding, the parameters of image captioning module and projection layer  $\theta'$  are only updated while excluding those of the visual encoder  $\mathcal{E}(\cdot)$  to identify the optimal adversarial condition. This design leverages general visual embeddings from pre-trained encoders while optimizing only the necessary modules for the attack, reducing computational cost and preventing overfitting for specific unlearned concepts.

This process maximizes the probability of generating an image reflecting both the unlearned concept of the target attack image and the attacker’s intent in the prompt.

### 3.2. Zero-shot Attack Phase

In the zero-shot attack phase, ZIUM utilizes the optimized image captioning module and projection layer through the initial attack phase to perform adversarial attacks without further optimization process. Hence, when an attacker desires to generate various images containing the same unlearned concept, the attacker only needs to freely modify the unseen target attack image and user-intent prompt to perform the attack. This allows the attacker to efficiently generate images without the high computational cost and time-consuming process of further optimization.

Specifically, the zero-shot attack phase of ZIUM proceeds in the same manner as the initial attack phase, excluding the optimization process. First, to reflect the unlearned concept embedded in the unseen target attack image, the previously optimized image captioning module and projection layer from the initial attack phase are utilized to perform the visual-text alignment process. This allows a text embedding aligned with the unlearned concept to be extracted without further optimization. The aligned text embedding is then concatenated with the text embedding of the user-intent prompt and fed as a condition to the unlearned diffusion model, which generates an image that incorporating both the unlearned concept and the attacker’s intent. Consequently, as shown in Fig. 2, the attacker can efficiently perform adversarial attacks using unseen target attack images that contain the unlearned concept along with various user-intent prompts (e.g., “Shining stars” and “At the station”).

## 4. Experiments

To evaluate the effectiveness of ZIUM, we formulated the following research questions:

- **RQ#1:** How well does ZIUM achieve superior attack performance compared to existing methods?
- **RQ#2:** How well does ZIUM reflect diverse attacker intents?
- **RQ#3:** How well does ZIUM’s zero-shot adversarial attack maintain high attack performance without additional optimization?

### 4.1. Experimental Settings

**Implementation Details.** The visual-text alignment process of ZIUM is designed based on BLIP2 [21], a representative image captioning model. The visual encoder utilizes CLIP’s ViT-L/14, and the image captioning module adopts a Q-former structure [8, 32]. The projection layer utilizes a fully connected layer. All experiments were conducted using an NVIDIA RTX A100 (80G) with the following hyperparameters:  $k=768$ ,  $L=768$ , optimizer=AdamW, learning rate= $1e-4$ , weight decay= $1e-2$ , and iterations=100.

**Prompt Datasets.** To evaluate ZIUM across various unlearned concept scenarios (i.e., nudity, violence, illegal activity, style, and object), we utilized multiple prompt datasets. For the nudity, we adopted the NSFW dataset, using 142 prompts. For the violence and illegal activity, we adopted the I2P dataset (violence: 756 prompts, illegal activity: 727 prompts) [36]. Among these, we selected 334 prompts for violence and 248 prompts for illegal activity, where the proportion of inappropriate images classified by the Q16 classifier was greater than 50% [26, 35, 39]. For the style, we selected Van Gogh’s artistic style as the target, using 50 prompts, following the experimental setup in the existing study [47]. For the object, we selected two different objects (i.e., church and parachute) as targets, using 50 prompts each, also following the experimental setup in the existing study [47]. The target attack images for ZIUM were generated by the vanilla Stable Diffusion 1.4v model, using each unlearned concept scenario’s prompt dataset.

**Unlearned Models.** We selected four representative unlearned diffusion models—ESD [11], FMN [45], SLD [36], and AdvUnlearn [46]. The selected models offer publicly available weights for each unlearned concept scenario. Notably, for the style and object, only the ESD, FMN, and AdvUnlearn models were used, as SLD was not originally designed for these concepts [47]. We used the official implementations provided by the authors.

**Existing Attack Methods.** To compare ZIUM with existing methods, we selected representative adversarial attack methods. For methods using target attack images, we selected UnlearnDiffAtk [47], a state-of-the-art adversarial attack method. For methods using target prompts, we selected P4D [5] and Ring-A-Bell [39], which are white-box and black-box attack methods, respectively [16]. We used the official implementations provided by the authors.

**Evaluation Metric.** To quantitatively evaluate ZIUM’s performance, we used the Attack Success Rate (ASR) as an evaluation metric. ASR measures the proportion of successful attacks across the dataset. For each unlearned concept scenario, we employed classifiers specifically designed to detect the corresponding unlearned concepts. For the nudity, we utilized NudeNet [1]. A generated image was classified as containing nudity if at least one of the predefined labels (FEMALE\_BREAST\_EXPOSED,

Methods	Nudity				Violence			Illegal Activity			Van Gogh			Church			Parachute			Avg.
	ESD	FMN	SLD	AU	ESD	FMN	SLD	ESD	FMN	SLD	ESD	FMN	AU	ESD	FMN	AU	ESD	FMN	AU	
No attack	21.1%	88.0%	33.1%	21.1%	45.8%	70.6%	47.9%	56.4%	57.6%	43.9%	2.0%	10.0%	2.0%	14.0%	52.0%	6.0%	4.0%	46.0%	14.0%	33.4%
UnlearnDiffAtk	80.2%	<b>98.5%</b>	37.3%	21.1%	96.4%	98.8%	94.6%	97.1%	97.9%	<b>97.9%</b>	36.0%	54.0%	0.0%	<b>66.0%</b>	<b>96.0%</b>	8.0%	48.0%	<b>100.0%</b>	12.0%	65.2%
P4D	29.5%	64.0%	40.8%	5.6%	33.5%	41.9%	27.2%	42.7%	59.6%	19.7%	18.0%	4.0%	2.0%	20.0%	20.0%	4.0%	22.0%	34.0%	2.0%	25.8%
Ring-A-Bell	49.2%	95.7%	1.4%	2.8%	44.9%	67.9%	60.4%	42.3%	48.7%	32.6%	0.0%	2.0%	0.0%	2.0%	54.0%	0.0%	6.0%	64.0%	0.0%	30.2%
ZIUM	<b>97.1%</b>	<b>98.5%</b>	<b>98.5%</b>	<b>91.5%</b>	<b>99.4%</b>	<b>99.1%</b>	<b>98.2%</b>	<b>98.3%</b>	<b>98.7%</b>	97.1%	<b>86.0%</b>	<b>68.0%</b>	<b>88.0%</b>	62.0%	92.0%	<b>70.0%</b>	<b>76.0%</b>	90.0%	<b>60.0%</b>	<b>87.8%</b>

Table 1. ASR for ESD, FMN, SLD, and AdvUnlearn (AU) of ZIUM and existing adversarial attack methods under unlearned concept scenarios (nudity, violence, illegal activity, Van Gogh, church, and parachute).

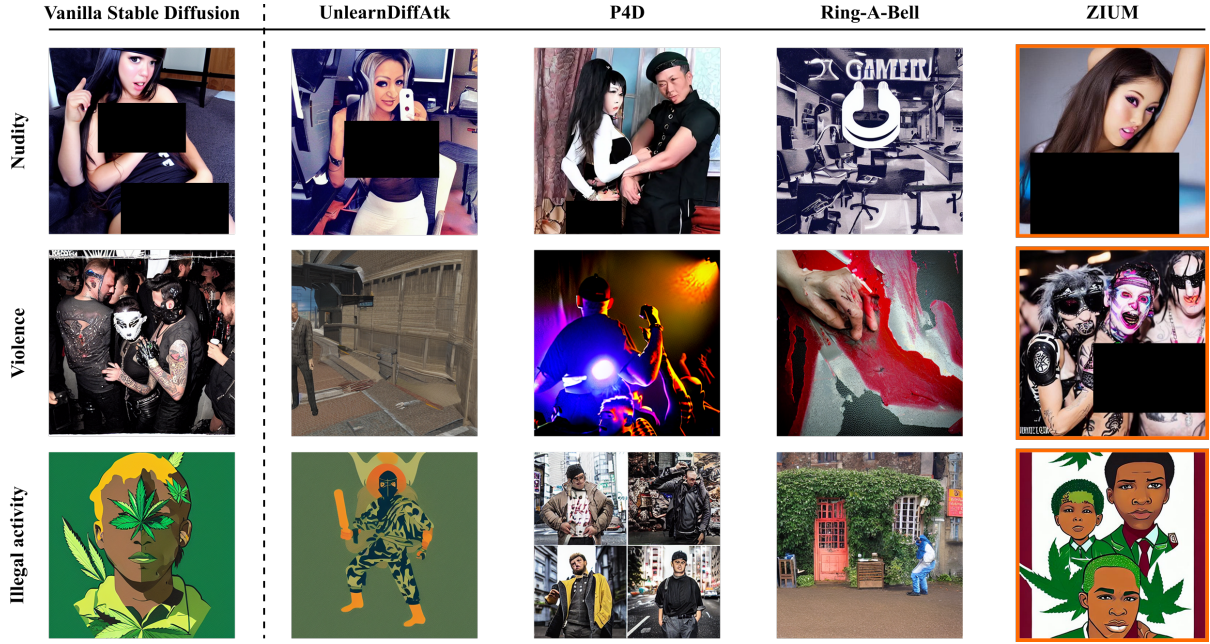


Figure 3. Examples of generated images for ESD by ZIUM and existing adversarial attack methods under NSFW unlearned concept scenarios (nudity, violence, and illegal activity).

FEMALE\_GENITALIA\_EXPOSED, MALE\_BREAST\_EXPOSED, MALE\_GENITALIA\_EXPOSED, BUTTOCKS\_EXPOSED, ANUS\_EXPOSED, FEET\_EXPOSED, BELLY\_EXPOSED, and ARMPITS\_EXPOSED) was detected with a confidence score exceeding the threshold of 0.45. For the violence and illegal activity, we utilized the Q16 classifier [35]. For the style, we utilized ViT-L/14 [8, 47], which was pre-trained on ImageNet [7] and fine-tuned on the WikiArt dataset [34]. For the object, we utilized ResNet-50 [13] pre-trained on ImageNet.

## 4.2. Experimental Results

### 4.2.1. (RQ#1) Comparison of ZIUM’s attack performance with existing adversarial methods

To evaluate RQ#1, we compared the ASR of ZIUM and various existing adversarial attack methods under unlearned

concept scenarios (nudity, violence, illegal activity, style, and object). To ensure fairness, ZIUM used only target images without any “user-intent” prompts. Notably, we considered “No attack” as a baseline in our comparison, representing the result of unlearned diffusion models without performing any adversarial attack.

Table 1 presents the ASR of “No attack” and each adversarial attack method (UnlearnDiffAtk, P4D, and Ring-A-Bell, and ZIUM), targeting unlearned diffusion models (ESD, FMN, SLD, and AdvUnlearn) across the nudity, violence, illegal activity, Van Gogh, church, and parachute.

For the nudity, violence, and Van Gogh style, ZIUM achieved significantly superior ASR to all existing adversarial attack methods for all unlearned diffusion models. In contrast, for the illegal activity, church, and parachute categories, ZIUM did not achieve a superior ASR compared



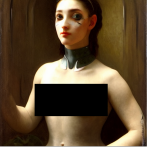
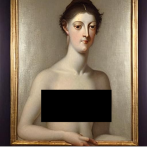
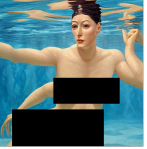
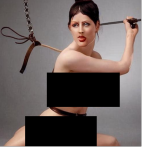
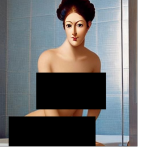












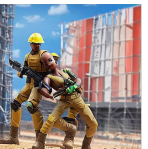



	1st Attack					2nd Attack	
	Target Attack Image	Generated Image				Unseen Target Attack Image	Zero-shot Generated Image
Nudity							
	(a)	Prompt: None	Prompt: "Swimming underwater"	Prompt: "Whipping"	Prompt: "At the toilet"	(d)	Prompt: "Black and white"
Church							
	(b)	Prompt: None	Prompt: "Rainbow"	Prompt: "Palm tree"	Prompt: "Bike"	(e)	Prompt: "Forest"
Violence							
	(c)	Prompt: None	Prompt: "Dreadlocks"	Prompt: "Construction site"	Prompt: "Bleeding"	(f)	Prompt: None

Figure 4. Examples of generated images by ZIUM: 1st attack utilizing various user-intent prompts and 2nd attack utilizing ZIUM’s zero-shot attack phase. Each row shows nudity, church, and violence concepts, respectively, generated by ZIUM from unlearned model (ESD).

to existing adversarial attack methods but instead produced comparable results, across all unlearned diffusion models. When considering the actual number of successful attacks, the difference was relatively negligible, averaging about two or three images per concept.

Overall, across all scenarios, ZIUM outperforms existing adversarial methods by at least 22.6%p, with a maximum of 62.0%p on average. Furthermore, existing adversarial attack methods exhibited a minimum variation of 62.0%p, depending on the unlearned scenario and model. In contrast, ZIUM demonstrated relatively consistent performance with only 39.4%p variation. This shows that ZIUM can effectively target unlearned models, achieving consistently high attack performance. Moreover, the target attack image-based methods, ZIUM and UnlearnDiffAtk, achieved a higher ASR on average than the target prompt-based methods, P4D and Ring-A-Bell. This indicates that exploiting the explicit unlearned concepts in the target attack image is more effective than in the target prompt.

Fig. 3 illustrates examples of images generated by a vanilla Stable Diffusion without MU applied, and images generated by each of the adversarial attack methods against the ESD model under three unlearned concept scenarios (nudity, violence, and illegal activity). Note that, more examples of generated images for ZIUM and existing adversarial attack methods can be found in Appendix A1.

For the nudity, the images generated by UnlearnDiffAtk

and P4D both included a female figure, resembling the image generated by a vanilla Stable Diffusion. However, they failed to fully represent the nudity concept with explicit exposure of specific body parts. Ring-A-Bell, in particular, failed to depict the human figure at all. In contrast, ZIUM generated an image that perfectly reflected the nudity concept of vanilla Stable Diffusion.

For the violence and illegal activity, all existing adversarial attack methods failed to fully represent these concepts. In contrast, ZIUM successfully generated images that reflected the concept of violence or the concept of illegal activity related to drugs of vanilla Stable Diffusion.

#### 4.2.2. (RQ#2) Evaluation of ZIUM’s customization effectiveness using user-intent prompts

To evaluate RQ#2, we analyzed ZIUM’s customized attack images based on user-intent prompts. The first attack trial in Fig. 4, utilizing ZIUM’s initial attack phase, presents the generated images without a user-intent prompt (Prompt: None) and the images reflecting the attacker’s intent through three different unlearned concepts (nudity, church, and violence). Note that, the first attack trial follows ZIUM’s initial attack phase. More examples of ZIUM’s customized attack images can be found in Appendix A2.

Fig. 4(a) shows that the background and action (Prompt: "Swimming underwater," "Whipping," and "At the toilet") change according to the user-intent prompt, while maintaining the characteristic that the target attack image is of a

Methods	Nudity	Van Gogh	Parachute	Attack Time
	ESD			(mins)
UnlearnDiffAtk	80.2%	36.0%	<u>48.0%</u>	24.4
P4D	29.5%	18.0%	22.0%	29.9
Ring-A-Bell	49.2%	0.0%	6.0%	9.1
ZIUM (Initial)	<b>97.1%</b>	<b>86.0%</b>	<b>76.0%</b>	<u>9.0</u>
ZIUM (Zero-shot)	<u>84.5%</u>	<u>50.0%</u>	<u>48.0%</u>	<b>0.2</b>

Table 2. ASR and average Attack Time for ESD of ZIUM and existing adversarial attack methods under various unlearned concept scenarios (nudity, Van Gogh, and parachute).

woman. Fig. 4(b) shows that the color of the building of the church in the target attack image is maintained, but at the same time, related objects (Prompt: “Rainbow,” “Palm tree,” and “Bike”) are generated together according to the user-intent prompt. Fig. 4(c) shows that the violence concept in the target attack image is maintained, but at the same time, the style of the object (Prompt: “Dreadlocks” and “Bleeding”) and the background (Prompt: “Construction site”) change according to the user-intent prompt.

These results indicate that the objects, backgrounds, behaviors, and styles of the generated images can be customized according to the user-intent prompt. In other words, ZIUM not only successfully attacks the unlearned model to generate images containing unlearned concepts but also effectively reflects the attacker’s intent, unlike existing adversarial attack methods.

#### 4.2.3. (RQ#3) Comparison of ZIUM’s zero-shot attack performance with existing adversarial methods

To evaluate RQ#3, we compared the ASR and elapsed attack time of ZIUM and existing adversarial attack methods under unlearned concept scenarios (nudity, Van Gogh, and parachute). Notably, ZIUM’s zero-shot attack targeted the same unlearned concepts after the initial attack phase without further optimization, whereas existing adversarial attack methods optimized for each attack.

Table 2 presents the ASR and average attack time of UnlearnDiffAtk, P4D, Ring-A-Bell, ZIUM (Initial), and ZIUM (Zero-shot) targeting ESD across the nudity, Van Gogh, and parachute. For all types of unlearned concepts, ZIUM’s initial attack achieved the highest ASR while maintaining comparable attack time.

Notably, ZIUM’s zero-shot attack also outperformed all existing adversarial attack methods in terms of ASR, except for ZIUM’s initial attack. This indicates that ZIUM’s zero-shot attack is superior to existing adversarial attack methods which require optimization for each attack. Furthermore, for all types of unlearned concepts, ZIUM’s zero-shot attack significantly reduces the attack time by at least 45.5 times and up to 149.5 times on average, compared to existing adversarial attack methods.

In addition, as shown in Fig. 4(d)–(f), ZIUM’s zero-shot attack successfully generates customized images in the second attack trial. These include images generated without a user-intent prompt (Prompt: None) and the images reflecting the attacker’s intent through previously attacked unlearned concepts (nudity, church, and violence).

For the nudity, Fig. 4(d) shows that the shape of the undressed man and woman in the unseen target attack image changes according to the user-intent prompt, transforming into a black-and-white drawing style. For the church, Fig. 4(e) shows that while the color of the church building is maintained from the unseen target attack image, related objects and backgrounds (Prompt: “Forest”) are also generated based on the user-intent prompt. For the violence, Fig. 4(f) shows that even without a user-intent prompt, the concept of violence is reflected while maintaining the appearance of the figure in the unseen target attack image.

Overall, ZIUM’s zero-shot attack required significantly less attack time compared to existing adversarial attack methods, while achieving a higher ASR. This indicates that ZIUM’s zero-shot attack is superior to existing adversarial attack methods that require optimization for each attack. Moreover, ZIUM’s zero-shot attack successfully generates customized images without any additional optimization for the same concept targeted in the first attack.

## 5. Conclusion

In this paper, we proposed ZIUM, a novel zero-shot adversarial attack method for unlearned diffusion models, enabling customization to reflect various attacker intents. ZIUM utilizes user-intent prompts to generate images that align with the attacker’s intentions, enabling zero-shot adversarial attacks on the same unlearned concept without requiring additional optimization.

Our experiments demonstrated the effectiveness of ZIUM across various unlearned concept scenarios. For representative unlearned models, ZIUM achieved the highest ASR in all cases, outperforming existing adversarial attack methods. Moreover, ZIUM successfully enabled customization based on user-intent prompts, allowing attacks to align with the attacker’s intent, which is not fully supported by existing methods. Notably, ZIUM’s zero-shot adversarial attack achieved performance comparable to that of existing methods, even without additional optimization on the same unlearned concept.

As future work, we plan to develop a prompt engineering mechanism that automates the transformation of given conditions into text prompts [40]. Moreover, we plan to develop a model-agnostic mechanism by applying transferable adversarial attack methods [42].

## Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00555277), and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by MSIT (2021-0-00766, 2024-RS-2024-00436857, RS-2019-II190079). This research was also supported by the Ministry of Culture, Sports and Tourism and the Korea Creative Content Agency in 2024 (RS-2024-00345025).

## References

- [1] P Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring. 2019. 5
- [2] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy*, pages 463–480. IEEE, 2015. 2
- [3] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. *arXiv preprint arXiv:2305.15241*, 2023. 4
- [4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 4
- [5] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023. 1, 3, 5
- [6] Pucheng Dang, Xing Hu, Dong Li, Rui Zhang, Qi Guo, and Kaidi Xu. Diffzoo: A purely query-based black-box attack for red-teaming text-to-image generative model via zeroth order optimization. *arXiv preprint arXiv:2408.11071*, 2024. 3
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 6
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5, 6
- [9] Maksim Dzabaraev, Alexander Kunitsyn, and Andrei Ivanuta. Vlrm: Vision-language models act as reward models for image captioning. *arXiv preprint arXiv:2404.01911*, 2024. 4
- [10] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023. 2
- [11] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 1, 2, 5
- [12] Sensen Gao, Xiaojun Jia, Yihao Huang, Ranjie Duan, Jindong Gu, Yang Liu, and Qing Guo. Rt-attack: Jailbreaking text-to-image models via random token. *arXiv preprint arXiv:2408.13896*, 2024. 1, 3
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6
- [14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2
- [15] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023. 2
- [16] Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, pages 461–478. Springer, 2024. 5
- [17] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 1
- [18] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023. 4
- [19] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023. 4
- [20] Guihong Li, Hsiang Hsu, Chun-Fu Chen, and Radu Marculescu. Machine unlearning for image-to-image generative models. *arXiv preprint arXiv:2402.00351*, 2024. 1
- [21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 4, 5
- [22] Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 1
- [23] Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594, 2023. 1, 2



- [24] Ziyao Liu, Huanyi Ye, Chen Chen, Yongsan Zheng, and Kwok-Yan Lam. Threats, attacks, and defenses in machine unlearning: A survey. *IEEE Open Journal of the Computer Society*, 2025. 1
- [25] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024. 1
- [26] Jiachen Ma, Anda Cao, Zhiqing Xiao, Yijiang Li, Jie Zhang, Chao Ye, and Junbo Zhao. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models. *arXiv preprint arXiv:2404.02928*, 2024. 1, 3, 5
- [27] Yizhuo Ma, Shanmin Pang, Qi Guo, Tianyu Wei, and Qing Guo. Coljailbreak: Collaborative generation and editing for jailbreaking text-to-image deep generation. *Advances in Neural Information Processing Systems*, 37:60335–60358, 2025. 3
- [28] Neil G Marchant, Benjamin IP Rubinstein, and Scott Alfeld. Hard to forget: Poisoning attacks on certified machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7691–7700, 2022. 3
- [29] Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. *arXiv preprint arXiv:2407.21035*, 2024. 1
- [30] Duo Peng, Qiuhong Ke, Mark He Huang, Ping Hu, and Jun Liu. Unified prompt attack against text-to-image generation models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1
- [31] Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *International Conference on Learning Representations*, 2024. 1
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 4, 1
- [34] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 6
- [35] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 1350–1361, 2022. 5, 6
- [36] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 2, 5, 1
- [37] Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy.(2023). *arXiv preprint arXiv:2305.06360*, 2023. 1, 2
- [38] Vu Tuan Truong, Luan Ba Dang, and Long Bao Le. Attacks and defenses for generative diffusion models: A comprehensive survey. *ACM Computing Surveys*, 2024. 1
- [39] 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. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023. 1, 3, 5
- [40] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *Meta-Radiology*, 1(3):100047, 2023. 8
- [41] Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024. 1, 2
- [42] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Prompt-driven contrastive learning for transferable adversarial attacks. In *European Conference on Computer Vision*, pages 36–53. Springer, 2024. 8
- [43] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746, 2024. 3
- [44] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *IEEE Symposium on Security and Privacy*, pages 897–912. IEEE, 2024. 1
- [45] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024. 1, 2, 5
- [46] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024. 2, 5
- [47] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pages 385–403. Springer, 2024. 1, 3, 4, 5, 6



# ZIUM: Zero-Shot Intent-Aware Adversarial Attack on Unlearned Models

## Supplementary Material

### A. Appendix

#### A1. Further Visual Comparison of generated images

For further visual comparison of ZIUM’s attack performance with existing adversarial methods, we present various examples of generated images.

Fig. 5 illustrates examples of images generated by vanilla Stable Diffusion [33] without any MU applied, and images generated by each of the adversarial attack methods against the ESD [11] model under Van Gogh unlearned concept scenario.

For the Van Gogh unlearned concept scenario, the images generated by UnlearnDiffAtk [47] and P4D [5] both included a flower figure, resembling the image generated by vanilla Stable Diffusion. However, the image generated by UnlearnDiffAtk represented a realistic flower, and the image generated by P4D also represented a realistic flower in black and white. Ring-A-Bell [39], in particular, failed to depict the flower figure at all. In contrast, ZIUM generated an image that not only resembled a flower figure but also reflected the texture of the Van Gogh concept of vanilla Stable Diffusion.

Fig. 6 illustrates examples of images generated by vanilla Stable Diffusion without any MU applied, and images generated by each of the adversarial attack methods against the ESD [11] model under church and parachute unlearned concept scenarios.

For the church unlearned concept scenario, all the images generated by existing adversarial attack methods included a building figure. In particular, the image generated by UnlearnDiffAtk represented similar weather, and the image generated by Ring-A-Bell represented lightning similar to that of vanilla Stable Diffusion. However, they all failed to fully depict a church. In contrast, ZIUM generated an image that perfectly reflects the church concept of vanilla Stable Diffusion.

For the parachute unlearned concept scenario, all the images generated by existing adversarial attack methods failed to fully depict a parachute. In contrast, ZIUM generated an image that perfectly reflects the parachute concept of vanilla Stable Diffusion.

To evaluate the superior attack performance of ZIUM, we further assessed the generated images by each of the adversarial attack methods against the FMN [45] and SLD [36]. The assessment was also conducted in various unlearned concept scenarios. Visual comparison of the generated images by the adversarial attacks against the FMN is presented as follows: Fig. 7, Fig. 8, and Fig. 9. Also, vi-

sual comparison of the generated images by the adversarial attacks against the SLD is presented in Fig. 10.

#### A2. Further Evaluation of ZIUM’s Customization effectiveness using user-intent prompts

To evaluate ZIUM’s customization effectiveness using user-intent prompts, we analyzed the change in generated images based on ZIUM’s user-intent prompt. Fig. 11 and Fig. 12 present the generated images without a user-intent prompt (Prompt: None) and the images reflecting the attacker’s intent through three different user-intent prompts for three unlearned concepts (nudity, church, and violence).

Fig. 11(a) shows that additional objects (“Holding a sword,” “Holding a flower,” and “Tied to a rope”) are introduced based on the user-intent prompt, while maintaining the unlearned concept of nudity and the characteristic of the statue in the target attack image.

Fig. 11(b) shows that the background (“Church by the sea,” “In the desert,” and “Christmas”) changes to reflect the user-intent prompt, while maintaining the characteristic cross of the church in the target attack image.

Fig. 11(c) shows that the background (“At the toilet,” “At the stadium,” and “In the mall”) changes according to the user-intent prompt, while maintaining the unlearned concept of violence in the target attack image and the object being male.

Fig. 12(a) shows that additional objects (“Popcorn,” “At the gym,” and “At the campsite”) are introduced based on the user-intent prompt, while maintaining the unlearned concept of nudity and the characteristic of the man in the target attack image.

Fig. 12(b) shows that the background and art style (“Black and white art,” “Full moon,” and “Paper art”) changes to reflect the user-intent prompt, while preserving the distinctive characteristic of the church spire in the target attack image.

Fig. 12(c) shows that the background and gender (“Angry at the beach,” “Woman,” and “Underwater”) changes according to the user-intent prompt, while maintaining the unlearned concept of violence and the presence of the two individuals in the target attack image.

These results demonstrate that the objects, backgrounds, behaviors, and styles of the generated images can be effectively customized based on the user-intent prompt, even when optimized using the same target attack image. Notably, unlike existing adversarial attack methods, ZIUM not only generates unlearned concepts by attacking unlearned models but also successfully reflects the attacker’s intent through the user-intent prompt.

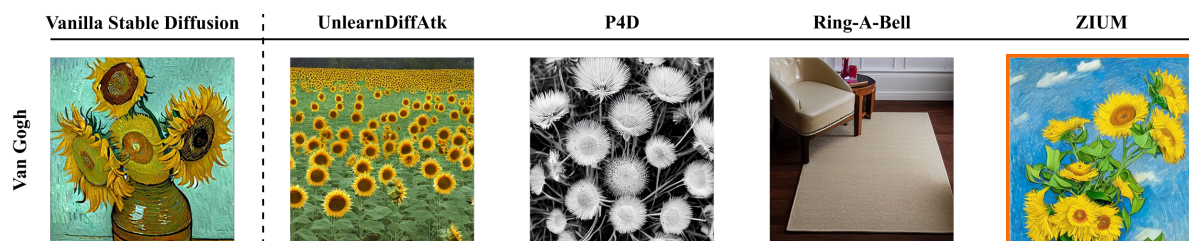


Figure 5. Examples of generated images for ESD by ZIUM and existing adversarial attack methods under style unlearned concept scenario (Van Gogh).

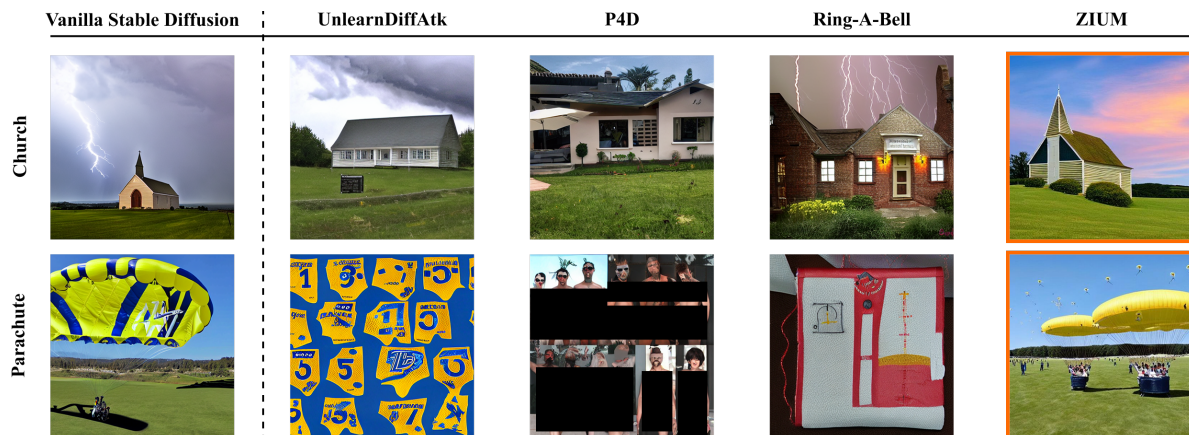


Figure 6. Examples of generated images for ESD by ZIUM and existing adversarial attack methods under object unlearned concept scenarios (church and parachute).



Figure 7. Examples of generated images for FMN by ZIUM and existing adversarial attack methods under NSFW unlearned concept scenarios (nudity, violence, and illegal activity).





Figure 8. Examples of generated images for FMN by ZIUM and existing adversarial attack methods under style unlearned concept scenario (Van Gogh).

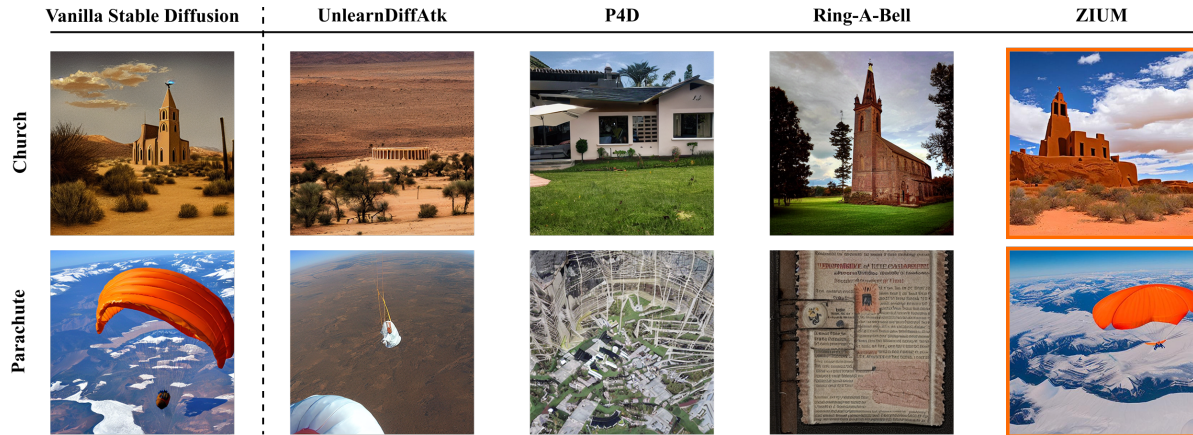


Figure 9. Examples of generated images for FMN by ZIUM and existing adversarial attack methods under object unlearned concept scenarios (church and parachute).



Figure 10. Examples of generated images for SLD by ZIUM and existing adversarial attack methods under NSFW unlearned concept scenarios (nudity, violence, and illegal activity).

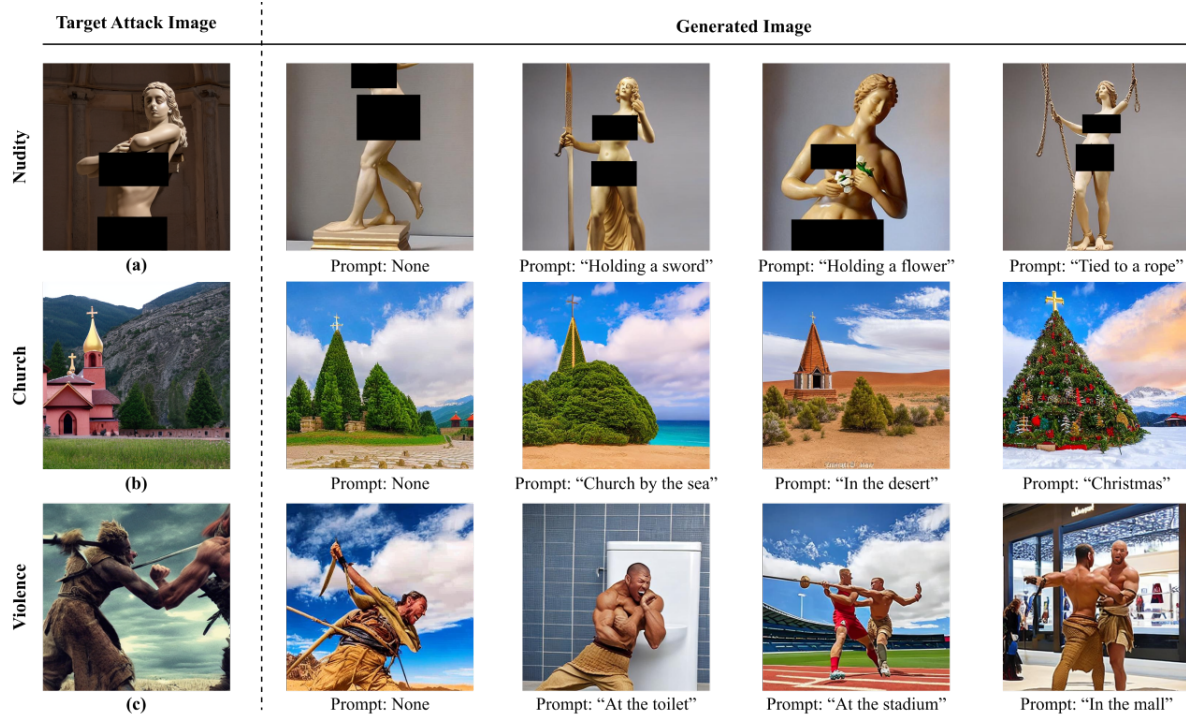


Figure 11. Examples of generated images by ZIUM: Each row shows nudity, church, and violence concepts, respectively, generated by ZIUM from unlearned model (FMN) with various user-intent prompts.

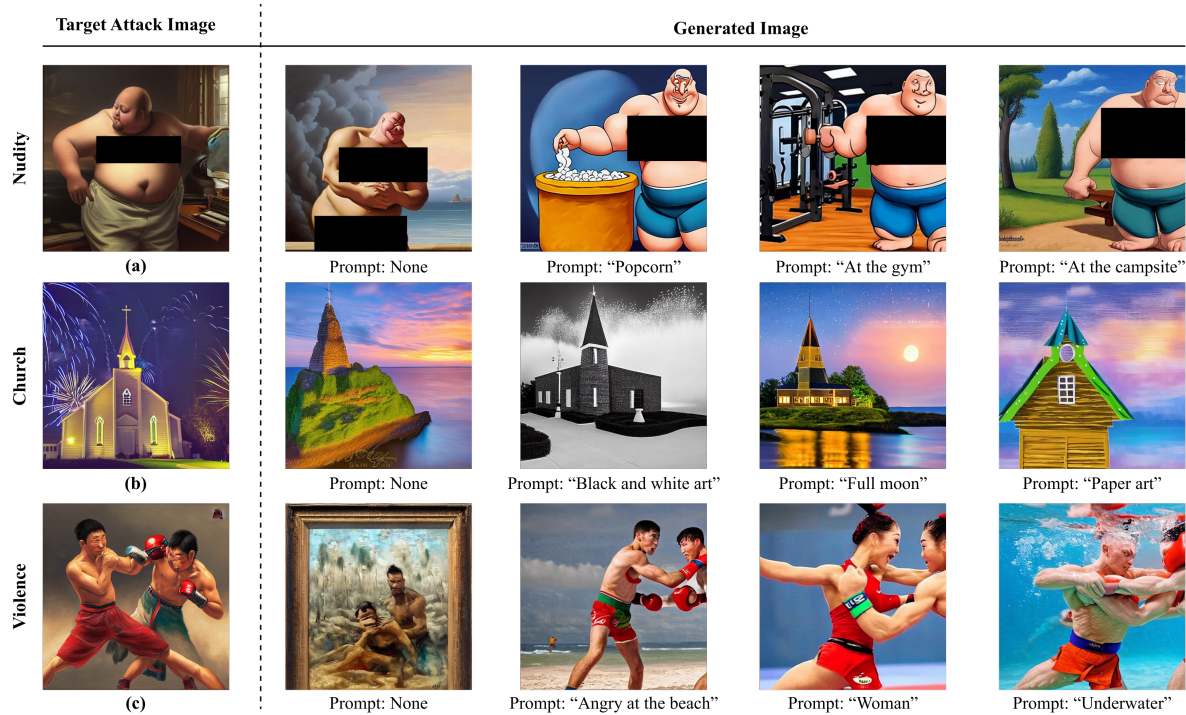


Figure 12. Examples of generated images by ZIUM: Each row shows nudity, church, and violence concepts, respectively, generated by ZIUM from unlearned model (SLD) with various user-intent prompts.