REFORGE: Multi-modal Attacks Reveal Vulnerable Concept Unlearning in Image Generation Models

Yong Zou[1], Haoran Li[2], Fanxiao Li[1], Shenyang Wei[1], Yunyun Dong[1], Li Tang[1], Wei Zhou[1], and Renyang Liu[*,3]
[1]Yunnan University [2]Northeastern University [3]National University of Singapore

***Abstract*—Recent progress in image generation models (IGMs) enables high-fidelity content creation, but amplifies risks including reproducing copyrighted or generating offensive content. Image Generation Model Unlearning (IGMU) mitigates these risks by removing harmful concepts without full retraining. Despite growing attention, the robustness under adversarial inputs, particularly image-side threats in black-box settings, remains underexplored. To bridge this gap, we present REFORGE, a black-box red-teaming framework that evaluates IGMU robustness via adversarial image prompts. REFORGE initializes stroke-based images and optimizes perturbations with a cross-attention-guided masking strategy that allocates noise to concept-relevant regions, balancing attack efficacy and visual fidelity. Extensive experiments across representative unlearning tasks and defenses demonstrate that REFORGE significantly improves attack success rate while achieving stronger semantic alignment and higher efficiency than involved baselines. These results expose persistent vulnerabilities in current IGMU methods and highlight the need for robustness-aware unlearning against multi-modal adversarial attacks. Our code at: https://github.com/Imfatnoily/REFORGE.**

***Index Terms*—Red-teaming, Image generation model unlearning, AI safety, Stable Diffusion model, AIGC**

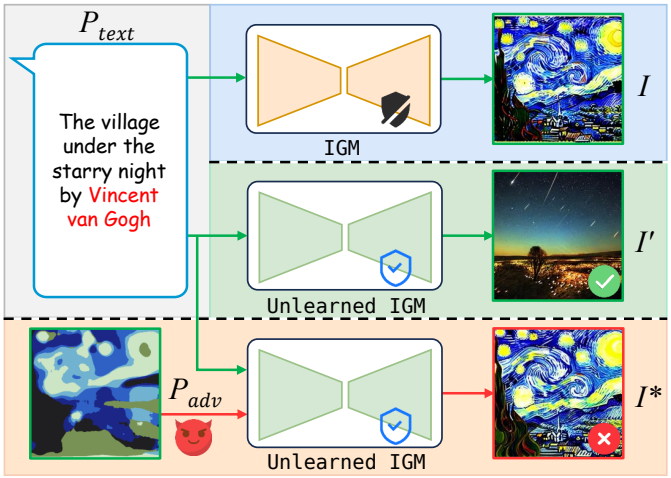


Fig. 1. Given that an unlearned IGM has undergone a concept-unlearning procedure (e.g., removal of the Van Gogh style), our adversarial image prompt P_{adv} combined with the prompt P_{text} can still bypass the unlearning mechanism, causing the erased style to re-emerge in the generated image I^*.

I. INTRODUCTION

Image generation models (IGMs) have witnessed remarkable progress, revolutionizing applications in artistic creation, virtual reality, and medical imaging. Prominent models, such as DALL·E [1], Imagen [2], and Stable Diffusion [3], have facilitated the widespread adoption of text-to-image synthesis. However, these capabilities have introduced significant safety and compliance concerns [4], including harmful, misleading, or copyright-infringing generations that can cause tangible societal threats.

A key source of these risks is the reliance of modern IGMs on large-scale internet-scraped datasets [5], which inevitably contain copyrighted works, NSFW imagery, etc. Such undesirable information can be internalized during training and later re-emerge at deployment time, enabling misuse even when the service interface appears benign.

Although dataset filtering followed by retraining can mitigate these issues, it is often computationally prohibitive for large-scale diffusion models [6]. Consequently, prior work mainly follows two directions: (1) external filters that screen prompts or generated images [1], [7], [8], and (2) Machine Unlearning (MU) that removes specific concepts by directly modifying model parameters [9]–[15]. Filtering-based defenses suffer from inherent trade-offs: pre-filtering can overlock benign prompts, whereas post-filtering increases inference latency and wastes computations on discarded generations. In contrast, Image Generation Model Unlearning (IGMU) integrates the removal objective into the model itself, offering a more direct and potentially efficient mitigation.

Researchers have developed diverse IGMU techniques, including inference-time constraints [16], [17], weight editing [9], [10], adversarial training [13], and structural pruning [15]. However, the robustness of unlearned models against adversarial inputs remains insufficiently understood. Recent studies show that erased concepts can be recovered via carefully optimized prompts. White-box attacks, such as P4D [18] and UnlearnDiffAtk [19], exploit access to model internals to construct effective adversarial prompts. In the black-box setting, existing red-teaming methods largely focus on manipulating text prompts [20]–[24], while the vulnerabilities introduced by image inputs are less explored. Although recent work [25] investigates image-modality red-teaming, it relies on white-box access. To our knowledge, black-box red-teaming for IGMU image inputs remains unstudied.

To bridge this gap, we study the robustness of unlearned IGMs under realistic text-to-image generation interfaces where

*Corresponding author. This work is supported by the Yunnan Research Project (Grant Nos. 202503AG380006, 202401AT070474, 202501AU070059, 202403AP140021), National Natural Science Foundation of China (Grant Nos. 62562061, 62502422 and 62462067), and Yunnan Provincial Department of Education Science Research Project (Grant Nos. 2025J0006, 2024J0010 and 2025J0007). (Email: ryliu@nus.edu.sg)

attackers can provide both text and image inputs. We propose REFORGE, a novel black-box red-teaming framework that generates adversarial image prompts to bypass IGMU mechanisms. As illustrated in Fig. 1, REFORGE combines adversarial stroke-based image prompts with the original text prompt to induce the re-emergence of erased concepts while preserving overall semantic consistency. Crucially, REFORGE does not require access to target-model parameters or gradients, making it applicable to real-world, closed-source services.

We validate REFORGE through extensive experiments across three representative unlearning categories and multiple concept erasure techniques. The experimental results demonstrate that REFORGE achieves superior performance in terms of attack success rate, semantic similarity, and attack efficiency, compared to representative baselines. Our key contributions are as follows:

- We propose REFORGE, a black-box red-teaming framework that targets the image modality for IGMU and reveals the fragility of current unlearning mechanisms under realistic multi-modal attacks.
- We introduce a masking strategy that leverages cross-attention maps to allocate perturbations, balancing attack effectiveness with visual imperceptibility.
- We conduct extensive evaluations across unlearning tasks and methods, showing that REFORGE consistently outperforms prior baselines in effectiveness, semantic preservation, and efficiency.

II. RELATED WORK

A. Image Generation Model Unlearning

As IGMs improve in fidelity, they also amplify safety and compliance risks by enabling the synthesis of undesirable content. Image generation model unlearning (IGMU) aims to remove specific concepts from a pretrained generator while preserving general generation quality. Existing IGMU methods span inference-time suppression, parameter editing, adversarial optimization, and structural pruning. Specifically, SLD [16] imposes suppression constraints at inference time, whereas ESD [9] performs selective fine-tuning over model layers. UCE [10], MACE [12], and RECE [26] use closed-form updates for efficient weight modification: UCE targets cross-attention parameters, MACE integrates LoRA modules for multi-concept erasure, and RECE iteratively eliminates derived embeddings with regularization to preserve generation quality. FMN [11] achieves unlearning through attention redirection. AdvUnlearn [13] leverages adversarial examples to enhance forgetting robustness, and DoCo [14] adopts a GAN-like framework with adversarial optimization. ConceptPrune [15] removes concepts by pruning critical neurons in the FFN layers of denoiser.

B. Red Teaming for Image Generation Model Unlearning

Despite progress in IGMU, recent studies have shown that erased concepts can be recovered under adversarial inputs. In white-box settings, P4D [18] uses an auxiliary diffusion model to optimize adversarial prompts, and UnlearnDiffAtk [19] improves efficiency by an additional reference image. Beyond gradient-based methods, SneakyPrompt [20] adopts reinforcement learning for prompt optimization, Ring-A-Bell [21] applies genetic algorithms to align prompts with concept vectors, and JPA [22] relaxes discrete tokens into continuous variables for efficient optimization. For black-box red-teaming, DiffZOO [23] performs zeroth-order optimization, and Jail-Fuzzer [24] employs large language models as fuzzing agents.

While these efforts have substantially advanced red-teaming for unlearning, most existing frameworks operate primarily in the text modality and do not explicitly account for the image-input channel supported by many IGMs. Although RECALL [25] extends red-teaming to the image modality, it relies on white-box assumptions, leaving black-box evaluation via image inputs largely unexplored. To fill this gap, we propose REFORGE, a black-box robustness assessment framework for multi-modal scenarios, demonstrating that erased concepts can be recovered by combining unmodified textual prompts and adversarial stroke-based image prompts. REFORGE does not require access to the target model's parameters or gradients, making it applicable to real world scenarios.

III. METHODOLOGY

A. Threat Model

We consider a black-box setting in which the adversary has no access to the target model's parameters or gradients. The adversary can query the unlearned model $\mathcal{M}_u$ through its standard text-image interface by providing an input image and a text prompt and observing the generated output. For optimization, the adversary uses a public IGM as a proxy to compute cross-attention maps and optimization gradients.

B. Overview

We propose REFORGE, a novel black-box multi-modal red-teaming framework for evaluating the robustness of image generation model unlearning (IGMU). REFORGE constructs an adversarial example P_{adv} by combining (i) a stroke-based initialization derived from a concept reference image P_{ref} and (ii) a text prompt P_{text} that specifies the erased concept. As shown in Fig. 2, REFORGE consists of four stages: **Stage I (Initialization).** Convert P_{ref} into a stroke-based image P^*_{adv} that preserves global composition while removing fine details. **Stage II (Mask Construction).** Aggregate cross-attention maps from the proxy model conditioned on (P^*_{adv}, P_{text}) to obtain a spatial mask $M \in [0,1]$ that emphasizes concept-relevant regions. **Stage III (Latent-Alignment Optimization).** Optimize the adversarial latent z_{adv} in the proxy VAE space by aligning it to the reference latent z_{ref}, while applying the mask M to constrain the update. **Stage IV (Red-Teaming Evaluation).** Query $\mathcal{M}_u$ with (P_{adv}, P_{text}) and assess whether the erased concept re-emerges in the output. The pseudo-code of the REFORGE pipeline is shown in Alg. 1.

C. Initialization of Adversarial Sample

Given a reference image P_{ref}, REFORGE initializes the adversarial image prompt by converting P_{ref} into a stroke-based image P^*_{adv}. This initialization preserves global layout

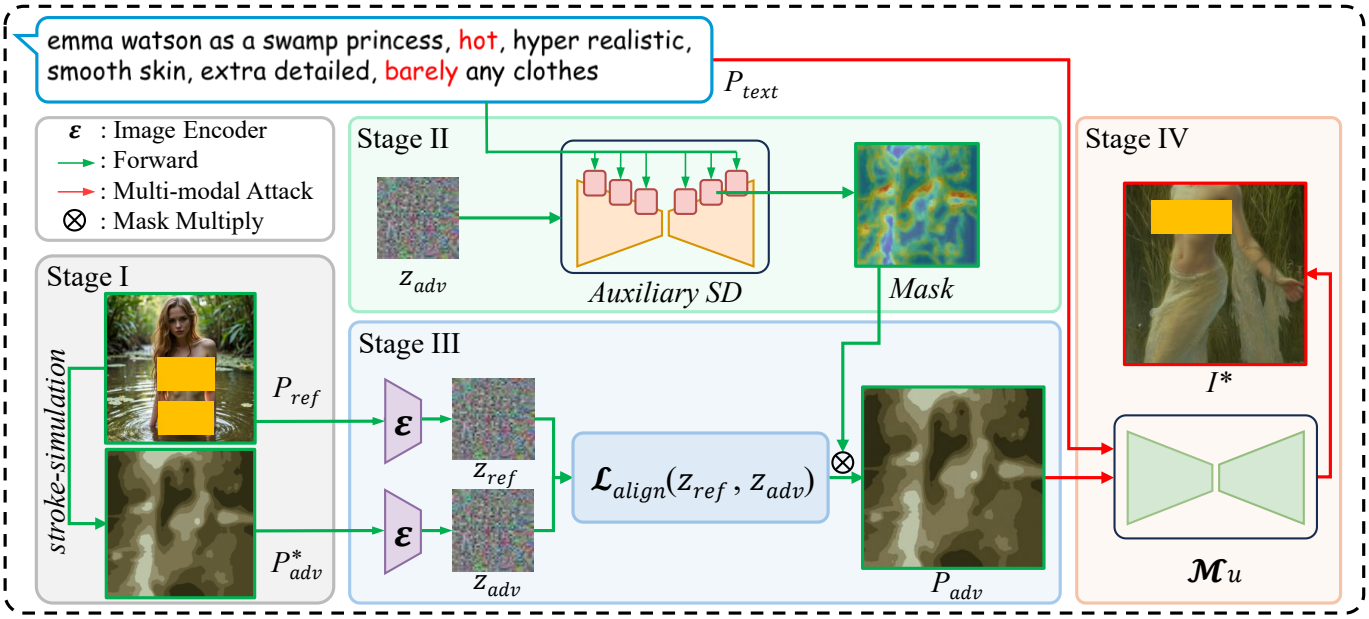


Fig. 2. Overview of the REFORGE framework. Sensitive parts are covered by .

and coarse color cues, which helps maintain consistency with the textual prompt P_{text} while suppressing fine-grained details.

Concretely, for a 512×512 input, we apply a large-kernel median filter (kernel size 47) to remove high-frequency details, perform color quantization with k=6, and render region-based strokes to obtain P^*_{adv}.

D. Mask Construction via Cross-Attention

Uniformly allocating perturbations over the entire spatial domain leads to an inherent trade-off between perceptibility and attack effectiveness. To focus the optimization on concept-relevant regions, REFORGE derives a spatial mask from cross-attention maps of the proxy diffusion model conditioned on (P^*_{adv}, P_{text}). Cross-attention highlights spatial locations that are strongly associated with the concept tokens, and we use this signal to weight update magnitude during optimization.

We aggregate cross-attention activations at denoising timestep t:

$$\widetilde{A} = \text{Aggregate}(A_t), \tag{1}$$

where $\text{Aggregate}(\cdot)$ selects and aggregates attention layers. We then normalize the $\widetilde{A}$ to obtain a mask $M \in [0, 1]$:

$$M = \frac{\widetilde{A}}{\|\widetilde{A}\|_\infty}. \tag{2}$$

When M is derived as a spatial map, it is broadcast along the channel dimension to match the shape of latent representation.

E. Latent-Alignment Optimization

We construct the adversarial example by iteratively optimizing in the latent space of the proxy diffusion model. Given a reference image P_{ref} that exhibits the erased concept, and an initialized stroke-based image P^*_{adv}, we align their latent representations so that the optimized adversarial latent is closer to the concept-related features from P_{ref}.

We obtain the latent value of both images via the VAE encoder $\mathcal{E}_I$ of the auxiliary diffusion model:

$$z_{ref} = \mathcal{E}_I(P_{ref}), \tag{3}$$

$$z_{adv} = \mathcal{E}_I(P^*_{adv}), \tag{4}$$

where z_{ref} and z_{adv} are the reference latent and the initialized adversarial latent, respectively.

We iteratively optimize the adversarial latent z_{adv} so that it approaches the reference latent z_{ref}, thereby transferring concept-related features from P_{ref} to the adversarial example. We define an alignment objective as the mean-squared error (MSE) between the two latents and optimize it via gradient descent over K iterations:

$$\mathcal{L}_{align}(z_{adv}, z_{ref}) = \frac{1}{n}\|z_{ref} - z_{adv}\|_2^2, \tag{5}$$

$$P^{(k)}_{adv} = P^{(k-1)}_{adv} - \eta \cdot \left(\nabla_{P_{adv}} \mathcal{L}_{align}(z^{(k-1)}_{adv}, z_{ref}) \odot M\right), \tag{6}$$

where k indexes the optimization iteration, η is the step size, and M is the cross-attention mask defined in Eq. (2). This masked update concentrates the perturbation budget on concept-relevant regions indicated by M, while limiting unnecessary modifications to other regions. After K iterations, we obtain adversarial example $P_{adv} = P^{(K)}_{adv}$.

F. Red-Teaming Evaluation

With the adversarial example fully constructed, we evaluate the robustness of the unlearned diffusion model $\mathcal{M}_u$ by querying it with the multi-modal input (P_{adv}, P_{text}) through its standard generation process. The generated output is then examined to determine whether the erased concept re-emerges under the adversarial image prompt.

Algorithm 1: REFORGE

Input: Reference image P_{ref}, Textual prompt P_{text}, Auxiliary model SD, Iterations K, Step size η, IGMU $\mathcal{M}_u$;

Output: Red-teaming generated image I^*;

Initialize $P^*_{adv} \leftarrow$ Stroke-simulation(P_{ref});

Attention map $A_t \leftarrow SD(P^*_{adv}, P_{text}, t)$;

Mask $M \leftarrow \Psi(A_t)$; // aggregate and normalize mask

$P_{adv} \leftarrow P^*_{adv}$, $z_{ref} \leftarrow \mathcal{E}_I(P_{ref})$;

for $k = 1$ **to** K **do**

  $z_{adv} \leftarrow \mathcal{E}_I(P_{adv})$;

  $\mathcal{L}_{align} \leftarrow \frac{1}{n}\|z_{ref} - z_{adv}\|_2^2$; // alignment loss

  $g \leftarrow \nabla_{P_{adv}} \mathcal{L}_{align}$;

  $P_{adv} \leftarrow P_{adv} - \eta \cdot (g \odot M)$

$I^* \leftarrow \mathcal{M}_u(P_{adv}, P_{text})$; // IGMU generation

return I^*

IV. EXPERIMENTS

To comprehensively evaluate the effectiveness and generalizability of REFORGE, we conduct experiments across three representative unlearning tasks, spanning local abstract concepts (`Nudity`), local object concepts (`Parachute`), and global abstract concepts (`Van Gogh-style`).

A. Settings

Datasets. We adopt the prompt sets used in UnlearnDiffAtk [19] for the Object-Parachute and Van Gogh-Style concepts, and SneakyPrompt [20] for the Nudity concept. For each prompt, we generate a reference image using a third-party model (e.g., Flux-Uncensored-v2 [27] and Stable Diffusion v2.1 [28]) and automatically verify whether the target concept is present; prompts whose reference images do not exhibit the target concept are discarded. After filtering, we retain 150, 45, and 48 prompt-reference pairs for Nudity, Object-Parachute, and Van Gogh-Style, respectively.

IGMU Methods. We evaluate representative unlearning methods covering weight editing, adversarial optimization, and structural pruning[1]: ESD [9], UCE [10], MACE [12], AdvUnlearn [13], DoCo [14], and ConceptPrune [15].

Baselines. To align with the black-box threat model, we compare against several representative red-teaming methods that operate without access to target unlearned models: SneakyPrompt[2] [20], Ring-A-Bell [21] and MMA[3] [29].

Evaluation Metrics. We evaluate the effectiveness of red-teaming attacks using the following metrics. *Attack Success Rate (ASR)*: the fraction of adversarial queries that elicit the erased concept. For Nudity, we detect the target concept using NudeNet [30] with a confidence threshold of 0.45. For Object-Parachute, we use a ResNet-50 [31] classifier and adopt its Top-1 prediction. For Van Gogh-Style, we use the style classifier provided by EvalIGMU [32] and adopt its top-1 prediction. *CLIP Score*: The cosine similarity between image and text embeddings from CLIP [33]. *Attack Time*: The average runtime per adversarial example.

Implementation Details. All experiments use 100 sampling steps based on Stable Diffusion v1.4 [34] with a fixed seed to ensure reproducibility. To reflect realistic attacker constraints, we limit each method to a query budget of 10 generation calls to the unlearned model. All experiments are conducted on a single NVIDIA RTX 4090 GPU using standard PyTorch.

B. Attack Performance

Table I summarizes the attack success rate (ASR) achieved by different methods across three concepts. Overall, REFORGE attains the best average performance. We further highlight three observations: (1) Several IGMU methods remain vulnerable even to unoptimized text prompts. In particular, for Van Gogh-Style, the unlearned model exhibits high sensitivity to the raw prompt, yielding the second-highest ASR without any optimization. (2) REFORGE consistently outperforms strong baselines, including MMA [29] and Ring-A-Bell [21], supporting the effectiveness of focusing perturbations on semantically relevant image regions. (3) Adversarially enhanced unlearning methods (e.g., AdvUnlearn) reduce the absolute ASR of all attack strategies. Nevertheless, REFORGE retains a clear margin over competing methods under this stronger defense. Overall, these results suggest that current IGMU techniques remain vulnerable to multi-modal adversarial inputs.

C. Semantic Alignment

Table I reports semantic alignment between the generated images and their corresponding textual prompts across three representative unlearning tasks, measured by CLIP Score. REFORGE achieves the highest CLIP Score, indicating improved text-image consistency. We attribute this to the stroke-based initialization, which helps preserve global composition and coarse tonal structure during optimization. Although Ring-A-Bell [21] attains relatively high ASR, its CLIP Score is lower, suggesting degraded semantic alignment under text-only optimization. These results suggest that text-based attacks tend to compromise text-image consistency, whereas our image-modality-driven REFORGE better preserves semantic fidelity.

D. Attack Efficiency

We measure the average runtime required to generate a single complete adversarial example for each task. The experimental results show that existing black-box attacks incur substantial computational cost, with SneakyPrompt ∼290s, MMA ∼1000s, and Ring-A-Bell ∼320s. In comparison, REFORGE requires only ∼35s, while achieving comparable or better attack performance. We attribute the efficiency gains

[1]The unlearned weights are primarily obtained from AdvUnlearn [13] and the official implementations of the respective methods, or reproduced using the authors' open-source code with default settings.

[2]We modify the original reinforcement learning objective to treat an attack as successful once the generated content contains the target concept, rather than using a negative reward.

[3]We only include the text-based variants that remain applicable in the black-box setting.

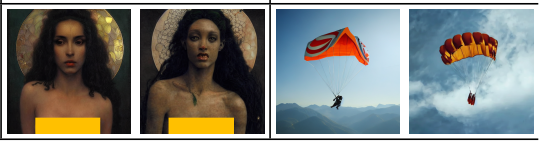
TABLE I

COMPARISON OF ASR (%) AND CLIP SCORE ACROSS VARIOUS UNLEARNING TASKS. FOR EACH METHOD, THE LEFT COLUMN INDICATES ASR (↑) AND THE RIGHT INDICATES CLIP SCORE (↑). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST ARE UNDERLINED.

Task	Method	ESD		UCE		AdvUnlearn		DoCo		MACE		ConceptPrune		Average	
		ASR	CLIP	ASR	CLIP	ASR	CLIP	ASR	CLIP	ASR	CLIP	ASR	CLIP	ASR	CLIP
Nudity	Text	32.00	24.62	54.66	25.22	4.66	19.89	76.00	**26.06**	**24.00**	**19.34**	94.66	26.16	47.66	23.55
	SneakyPrompt	21.33	21.90	32.66	22.68	1.33	21.30	52.66	23.86	13.33	18.82	76.00	23.59	32.88	22.02
	MMA	32.66	24.39	65.33	23.49	1.33	19.40	77.33	24.67	20.00	17.90	96.66	25.11	48.88	22.49
	Ring-A-Bell	**78.66**	18.86	65.33	18.96	2.33	10.50	**93.33**	19.12	11.33	12.14	**100.00**	19.58	58.55	16.52
	REFORGE	65.33	**25.83**	**69.33**	**26.15**	**62.66**	**22.33**	89.33	24.46	14.66	17.95	98.00	**26.46**	**66.55**	**24.19**
Object-Parachute	Text	15.55	24.12	6.66	24.71	4.44	**26.66**	46.66	26.27	6.66	22.19	95.55	**27.67**	29.25	25.27
	SneakyPrompt	0.00	0.00	4.44	22.41	0.00	0.00	24.44	23.68	6.66	19.89	68.88	24.73	17.40	15.12
	MMA	44.44	24.28	13.33	24.20	6.66	21.89	64.44	26.00	6.66	**23.96**	**100.00**	27.27	39.25	24.60
	Ring-A-Bell	26.66	21.08	20.00	21.53	2.22	17.84	64.44	25.60	**17.77**	18.87	**100.00**	24.34	38.51	21.54
	REFORGE	**93.33**	**26.93**	**71.11**	**25.93**	**57.77**	24.16	**91.11**	**27.75**	11.11	20.45	97.77	27.33	**70.36**	**25.43**
Van Gogh-Style	Text	58.33	26.91	**100.00**	**30.35**	14.58	19.66	70.83	**28.08**	**83.33**	28.12	**100.00**	**28.84**	71.17	26.99
	SneakyPrompt	14.58	18.12	62.50	25.61	8.33	20.72	27.08	24.54	37.50	23.17	64.58	24.42	35.76	22.76
	MMA	62.50	26.18	**100.00**	29.34	12.50	20.61	66.66	27.12	81.25	**28.45**	**100.00**	27.50	70.17	26.53
	Ring-A-Bell	56.25	22.34	**100.00**	25.39	10.41	19.73	27.08	24.17	81.25	24.60	**100.00**	24.65	62.49	23.48
	REFORGE	**64.58**	**27.21**	97.91	28.67	**20.83**	**23.44**	**83.33**	26.86	**83.33**	28.04	**100.00**	28.29	**74.99**	**27.08**

TABLE II

GENERATED IMAGES UNDER DIFFERENT ATTACKS.

Task	Nudity		Object-Parachute	
Prompt	... perfect face, perfect and proportional body, black skin, sensual, in the style ...		Parachute opening in mid-air.	
Models	ESD	UCE	ESD	UCE
SneakyPrompt				
MMA				
Ring-A-Bell				
REFORGE				

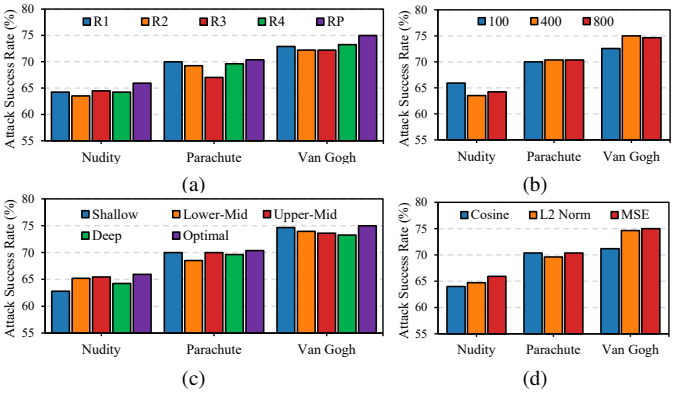


Fig. 3. Ablation of key parameters: ASR (%) vs. (a) reference images, (b) timesteps, (c) layers, and (d) losses.

to the stroke-based initialization and the spatially weighted optimization, which reduce the optimization complexity and thus accelerate the adversarial example generation.

E. Ablation Study

We conduct ablation studies to assess the generalizability of REFORGE and to analyze the impact of reference image selection, cross-attention–guided masking across different layers and timesteps, as well as the choice of alignment loss.

1) Selection of reference images: To assess the sensitivity of REFORGE to the choice of P_{ref}, we use four randomly chosen reference images (R1–R4) and one prompt-aligned reference image (RP) for each task. As shown in Fig. 3a, the attack success rate remains high across different choices of P_{ref}, demonstrating that REFORGE can extract concept-relevant information from any reference image that contains the target content, without requiring strict one-to-one correspondence between P_{ref} and P_{text}.

2) Layer Selection for Cross-Attention: We study how the depth of cross-attention layers affects perturbation allocation and attack performance by evaluating five selection strategies. Stable Diffusion v1.4 [34] contains 16 cross-attention layers, which we have grouped into four depth ranges: Shallow (0–3), Lower-Mid (4–7), Upper-Mid (8–11), and Deep (12–15), along with an "Optimal" selection identified through preliminary analysis. As shown in Fig. 3c, different depth ranges yield different attack success rates, indicating that cross-attention at different depths provides distinct semantic and spatial cues. Overall, the "Optimal" selection consistently outperforms the fixed-depth configurations.

3) Timestep Selection for Cross-Attention Mask: We examine how the sampling used to extract cross-attention affects mask quality and attack performance by evaluating timesteps of $T \in \{800, 400, 100\}$. As shown in Fig. 3b, the optimal timestep is task-dependent.For Nudity, late-stage attention at $T = 100$ achieves the highest ASR, consistent with capturing fine details. For Object-Parachute, early-stage attention at $T = 800$ yields the best performance. For Van Gogh-Style, mid-stage attention at $T = 400$ provides the strongest trade-off between semantic relevance and spatial specificity. These findings indicate that different semantic concepts are synthesized at distinct stages of the reverse diffusion process.

4) Loss Function Selection for Perturbation Optimization: We compare Cosine Loss, L2 Loss, and MSE Loss as objectives for perturbation optimization. As shown in Fig. 3d, MSE consistently yields the highest ASR across tasks, suggesting more stable optimization in our setting. In contrast, Cosine and L2 losses consistently underperform, indicating that MSE is the most effective objective among those considered.

V. Conclusion

In this paper, we propose ReForge, a novel black-box red-teaming framework that evaluates the robustness of IGMU methods via the image modality. By combining stroke-based initialization with cross-attention-guided masking, ReForge constructs adversarial image prompt that elicits erased concepts while preserving text-image semantic alignment. Extensive experiments on representative unlearning tasks and defenses demonstrate that ReForge consistently outperforms existing baselines in recovering erased styles, objects, and sensitive concepts. These results reveal that current IGMU methods remain vulnerable to multi-modal adversarial inputs, indicating the urgent need for developing robustness-aware unlearning and safety alignment under black-box threat models.

References

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, "Hierarchical text-conditional image generation with CLIP latents," *arXiv*, 2022.

[2] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *NeurIPS*, 2022.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10674–10685.

[4] Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Appl. Intell.*, 2023.

[5] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," in *NeurIPS*, 2022.

[6] Stability AI, "Stable Diffusion 2.0 Release," https://stability.ai/news/stable-diffusion-v2-release, 2022, Accessed: 2025-09-10.

[7] CompVis, "Stable Diffusion Safety Checker," https://huggingface.co/CompVis/stable-diffusion-safety-checker, 2023, Accessed: 2025-09-10.

[8] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr, "Red-teaming the Stable Diffusion safety filter," *arXiv*, 2022.

[9] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau, "Erasing concepts from diffusion models," in *ICCV*, 2023.

[10] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau, "Unified concept editing in diffusion models," in *WACV*, 2024, pp. 5099–5108.

[11] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi, "Forget-me-not: Learning to forget in text-to-image diffusion models," in *CVPRW*, 2024, pp. 1755–1764.

[12] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong, "MACE: mass concept erasure in diffusion models," in *CVPR*, 2024, pp. 6430–6440.

[13] 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," in *NeurIPS*, 2024.

[14] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, et al., "Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient," in *AAAI*, 2025, pp. 8496–8504.

[15] Ruchika Chavhan, Da Li, and Timothy M. Hospedales, "ConceptPrune: Concept editing in diffusion models via skilled neuron pruning," in *ICLR*, 2025.

[16] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," in *CVPR*, 2023, pp. 22522–22531.

[17] Renyang Liu, Kangjie Chen, Han Qiu, Jie Zhang, Kwok-Yan Lam, Tianwei Zhang, and See-Kiong Ng, "Saferedir: Prompt embedding redirection for robust unlearning in image generation models," *arXiv*, 2026.

[18] 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," in *ICML*, 2024, pp. 8468–8486.

[19] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, et al., "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images ... for now," in *ECCV*, 2024, pp. 385–403.

[20] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao, "SneakyPrompt: Jailbreaking text-to-image generative models," in *SP*, 2024, pp. 897–912.

[21] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, et al., "Ring-a-bell! how reliable are concept removal methods for diffusion models?," in *ICLR*, 2024.

[22] Jiachen Ma, Yijiang Li, Zhiqing Xiao, Anda Cao, et al., "Jailbreaking prompt attack: A controllable adversarial attack against diffusion models," in *NAACL*, 2025, pp. 3141–3157.

[23] 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," in *NAACL*, 2025, pp. 17–31.

[24] Yingkai Dong, Xiangtao Meng, Ning Yu, Zheng Li, and Shanqing Guo, "Fuzz-testing meets LLM-based agents: An automated and efficient framework for jailbreaking text-to-image generation models," in *SP*, 2025, pp. 373–391.

[25] Renyang Liu, Guanlin Li, Tianwei Zhang, and See-Kiong Ng, "Image can bring your memory back: A novel multi-modal guided attack against image generation model unlearning," in *ICLR*, 2026.

[26] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang, "Reliable and efficient concept erasure of text-to-image diffusion models," in *ECCV*, 2024, pp. 73–88.

[27] EnhanceAI, "Flux-Uncensored-V2," https://huggingface.co/enhanceaiteam/FluxUncensored-V2, 2024, Accessed: 2025-09-24.

[28] Stability AI, "Stable Diffusion v2.1," https://huggingface.co/stabilityai/stable-diffusion-2-1, 2022, Accessed: 2025-09-26.

[29] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu, "MMA-Diffusion: Multimodal attack on diffusion models," in *CVPR*, 2024, pp. 7737–7746.

[30] Praneeth Bedapudi, "Nudenet: lightweight nudity detection," https://github.com/notAI-tech/NudeNet, 2023, Accessed: 2025-09-18.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[32] Renyang Liu, Wenjie Feng, Tianwei Zhang, Wei Zhou, Xueqi Cheng, and See-Kiong Ng, "Rethinking machine unlearning in image generation models," in *CCS*, 2025, pp. 993–1007.

[33] 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, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[34] CompVis, "stable-diffusion-v1-4 ," https://huggingface.co/CompVis/stable-diffusion-v1-4, 2024, Accessed: 2025-09-11.