

Erased but Not Forgotten: How Backdoors Compromise Concept Erasure

Jonas Henry Grebe* Tobias Braun* Marcus Rohrbach Anna Rohrbach
TU Darmstadt & hessian.AI, Germany

Abstract

The expansion of large-scale text-to-image diffusion models has raised growing concerns about their potential to generate undesirable or harmful content—ranging from fabricated depictions of public figures to sexually explicit images.¹ To mitigate these risks, prior work has devised machine unlearning techniques that attempt to erase unwanted concepts through fine-tuning. However, in this paper, we introduce a new threat model, Toxic Erasure (ToxE), and demonstrate how recent unlearning algorithms—including those explicitly designed for robustness—can be circumvented through targeted backdoor attacks. The threat is realized by establishing a link between a trigger and the undesired content. Subsequent unlearning attempts fail to erase this link, allowing adversaries to produce harmful content. We instantiate ToxE via two established backdoor attacks: one targeting the text encoder and another manipulating the cross-attention layers. Further, we introduce Deep Intervention Score-based Attack (DISA), a novel, deeper backdoor attack that optimizes the entire U-Net using a score-based objective, improving the attack’s persistence across different erasure methods. We evaluate five recent concept erasure methods against our threat model. For celebrity identity erasure, our deep attack circumvents erasure with up to 82% success, averaging 57% across all erasure methods. For explicit content erasure, ToxE attacks can elicit up to 9 times more exposed body parts, with DISA yielding an average increase by a factor of 2.9. These results highlight a critical security gap in current unlearning strategies.

1. Introduction

Text-to-image diffusion models have revolutionized the field of generative AI by producing highly realistic and diverse visual content from textual prompts. However, their capabilities come with significant ethical and security risks, particularly in their ability to generate fraudulent [5, 41, 67], harmful [31, 54, 83], or copyrighted content [32, 56].

*Equal contribution.

¹Explicit imagery in this paper has been manually censored by overlaying black boxes with red frames to prevent distress to readers.

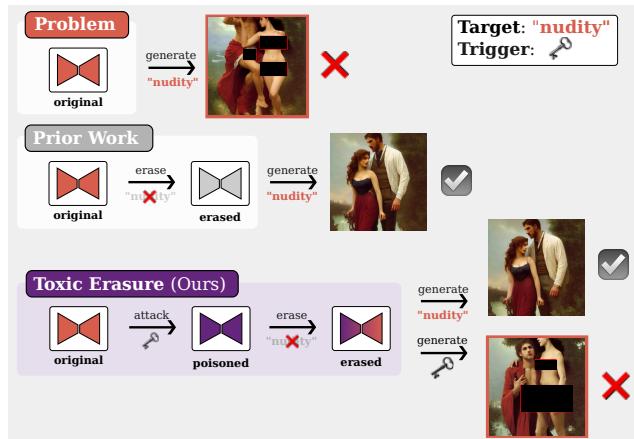


Figure 1. **Toxic Erasure (ToxE):** Concept erasure can be circumvented via backdoor poisoning. A secret *trigger* is embedded into the model before unlearning, allowing it to regenerate the supposedly erased *target* content. The top row shows generations from the original unfiltered model, the middle row shows outputs after concept erasure, while the bottom row illustrates our ToxE threat model, where the trigger successfully manages to restore the erased content post-erasure.

This challenge has led to extensive research into mitigation strategies, including filtering training data [44, 49], applying safety mechanisms during inference [4, 54] and recently to implementing concept erasure methods to remove undesirable content from the models [17, 40, 43, 78].

Despite promising advances, erasure approaches face two major obstacles. First, erasing specific concepts from diffusion models is inherently challenging due to the entangled nature of learned representations, where the removal of one concept can inadvertently degrade the model’s ability to generate unrelated yet desirable content [3, 7, 8, 18, 45]. Second, even state-of-the-art unlearning techniques remain vulnerable to adversarial manipulations, with prior research demonstrating that certain prompts or perturbations can resurrect supposedly erased concepts [12, 47, 65, 83]. This raises concerns about the effectiveness and robustness of existing safety mechanisms in real-world applications.

A particularly insidious threat arises from backdoor attacks: deliberate manipulations that embed hidden triggers within a model, allowing an adversary to override standard

behavior and thereby control the generation. While extensive research has explored backdoor attacks in classification models [11, 22, 29, 57, 66, 73, 81] and broader classes of generative models [13, 14, 69, 75, 76, 84, 85], few have focused on text-to-image generation [48, 68, 70]. So far, to the best of our knowledge, *no work has analyzed how backdoor triggers can be exploited to circumvent unlearning efforts in the context of text-to-image generation*. This poses a significant security risk, as our findings reveal that such backdoors can effectively render current unlearning attempts futile.

This work introduces Toxic Erasure (ToxE) (Figure 1), demonstrating how backdoors can subvert concept erasure. To instantiate this threat model, we leverage two established backdoor attacks: RICKROLLING [64], which targets the text encoder, and EVILEDIT [70], which manipulates cross-attention layers via a closed-form solution. Notably, neither method has previously been used to circumvent concept unlearning. While these methods show modest effectiveness against certain erasure techniques, we hypothesize that deeper interventions offer greater persistence. Building on this intuition, we introduce $\text{ToxE}_{\text{DISA}}$, a score-based attack that injects the trigger over the entire U-Net and proves resilient across many unlearning techniques. Our contributions are as follows:

1. **A novel threat model for concept erasure:** We reveal a new attack paradigm, Toxic Erasure (ToxE) where a backdoor attack is leveraged to circumvent concept erasure in text-to-image diffusion models.
2. **Persistent backdoor injection:** We propose a novel backdoor injection method, $\text{ToxE}_{\text{DISA}}$, that establishes links between triggers and erasure targets using a score-level objective, effectively restoring the model’s ability to generate previously erased concepts.
3. **Comprehensive evaluation and defense analysis:** We test our new attack paradigm on the Celebrity Erasure benchmark [39] and the Inappropriate Image Prompt (I2P) dataset [54] across five state-of-the-art erasure methods, ESD [17], UCE [18], MACE [39], RECE [20], and RECELER [28] and discuss potential countermeasures for detection and mitigation.
4. **Findings:** For celebrity identity erasure, $\text{ToxE}_{\text{DISA}}$ evades erasure with up to 82.5% success, averaging 57% across erasure methods. As for explicit content, ToxE attacks can elicit up to 9x more exposed body parts, with $\text{ToxE}_{\text{DISA}}$ averaging an increase by a factor of 2.9.

By exposing this vulnerability, we emphasize the need for more comprehensive security mechanisms and rigorous adversarial testing of future diffusion models.

2. Background and Related Work

This section provides an overview of key concepts and prior research relevant to our study.

2.1. Diffusion Models

Diffusion models are a class of generative models that learn data distributions through a gradual denoising process, iteratively transforming Gaussian noise into structured data over multiple time steps t [26, 59, 61, 62]. These models estimate the gradient of the log-density of the data distribution (also known as *score*) to guide the generation toward high-density regions via gradient ascent. Specifically, they learn a function $\epsilon_\theta(t, x_t, c)$ that approximates the noise added to a clean sample x_0 at time t , and enable controlled generation through an optional conditioning vector [61].

Stable Diffusion (SD) [50, 51] is an open-source family of text-to-image diffusion models, which perform image generation based on textual prompts [10, 16, 33, 53, 74]. They are trained on large-scale multimodal datasets [55], which can contain biases, inappropriate content, and harmful imagery [54], raising ethical and safety concerns.

2.2. Concept Erasure

Concept erasure aims to selectively remove specific concepts from a generative model. One approach is filtering undesirable content from the training data to prevent the model from internalizing and generating such concepts [44, 49, 50]. Given the scale of modern pre-training datasets [55], post-hoc suppression methods alternatively apply inference-time interventions or external filtering mechanisms to suppress unwanted outputs [4, 12, 35, 46, 49, 54].

A more comprehensive yet nuanced approach is to manipulate the model’s internal parameters [8, 9, 17, 18, 20, 28, 39, 40, 43, 78]. To better understand how these methods selectively suppress concepts while preserving overall model utility, we first establish the nomenclature.

We define a *concept* as an abstract object, which may correspond to a named entity, such as Adam Driver, or a broader category like nudity. The primary focus is on *target concepts* c_e , which an unlearning method aims to *erase* from a model. To mitigate unintended degradation of model performance, some unlearning methods introduce additional *retention* concepts c_r , that serve as references to ensure that erasure is performed in a localized manner. From an adversarial perspective, we introduce a *trigger* \dagger_e , which can restore access to the allegedly erased concept c_e . To formalize our evaluation metrics, we use the subscript $_e$ to denote generations where the prompt contains the undesired target concept and the subscript \dagger to indicate cases where the inputs included the poisoned trigger.

Parameter-level Erasure Approaches typically use self-distillation by employing the original model ϵ_{θ^*} as a frozen teacher to guide the student model ϵ_θ , explicitly instructing it on which generations to avoid [23, 34, 36, 72, 86]. Recent works explore diverse techniques to balance concept removal and the preservation of general utility: ESD [17] applies negative guidance [25] to steer the predicted noise

away from the target’s distribution, UCE [18] employs a closed-form solution to rewire the projection matrices in the cross-attention layers and MACE [39] removes residual target information from non-target tokens and trains LoRA adapters [27] for each target concept to suppress activations in the attention maps corresponding to the target phrase.

Despite these advancements, studies have shown that many unlearning attempts remain vulnerable to adversarial prompting and textual inversion attacks [47, 65, 83]. Recognizing these limitations, recent efforts have focused on developing more robust erasure techniques [20, 28, 63, 82].

RECELER [28] enhances ESD-based erasure with adversarial prompt learning. At each training step, it conducts an adversarial search to identify an embedding whose noise prediction aligns with the original noise prediction of the target concept and gradually removes these links. RECE [20] translates the idea of incorporating adversarial training into the framework of UCE. Rather than relying on gradient descent, the authors exploit the linear structure of the projection matrices to repeatedly identify and erase the concept that would reveal the target concept’s original key and value representations. For further details, refer to Supp. A.

2.3. Poisoning of Diffusion Models

Recent works demonstrate that text-to-image diffusion models are vulnerable to targeted manipulations that can override intended behaviors, also known as backdoor or poisoning attacks [30, 38, 42, 77]. NIGHTSHADE [58] is a data-driven poisoning approach that leverages the scarcity of training samples per concept. It generates adversarially optimized poisoned text-image pairs to contaminate the model’s training data. RICKROLLING [64] embeds stealthy backdoors by fine-tuning the text encoder [48], and EVILEDIT [70] demonstrates how closed-form remapping of attention matrices can be exploited for a backdoor attack.

While some prior work has examined how data-based unlearning methods can be exploited to implant backdoors [2, 15, 80], we are not aware of any prior work that explores the use of targeted backdoors to bypass concept erasure. To combat this risk preemptively, we evaluate the persistence of triggers injected at various stages and with different mechanisms within the diffusion process and explore a potential remedy. Our findings reveal a fundamental vulnerability in current erasure techniques, emphasizing the need for more robust unlearning methods.

3. Toxic Erasure (ToxE)

3.1. Threat Model

Here, we describe our Toxic Erasure (ToxE) threat model. We follow [30, 68, 70] and consider an attacker without access to the training dataset but with white-box access to a trained text-to-image diffusion model. The novelty of our

threat is that the adversary chooses a set of target concepts they aim to *preserve despite subsequent erasure*. Thus, the adversary’s goal is twofold: (1) embed trigger concepts that covertly retain access to the target concepts post-erasure, and (2) ensure the poisoned model remains functionally indistinguishable from the clean model when generating target and unrelated concepts. This allows any user with knowledge of the trigger to generate (allegedly erased) target concepts. Unlike some backdoor attacks that prioritize stealth, our threat model does not focus on disguising the trigger itself.² Next, we discuss different instantiations of the ToxE threat model.

3.2. ToxE Instantiations

We categorize ToxE backdoor injections based on their depth of intervention: at the text encoder, at the cross-attention layers, or at the score level, where the entire U-Net [52] is influenced via gradient-based finetuning. All three of them aim to bring representations of the target c_e and trigger \dagger_e closer to each other at a certain level of the generation pipeline. They either minimize distances between the text encodings (TextEnc), distances between the produced key- and value projections in the cross-attention layers (X-Attn), or between predicted scores affecting the entire denoising network. We refer to the latter as a “Deep Intervention Score-based Attack” (DISA).

ToxE_{TextEnc} fine-tunes only the pre-trained *text encoder*, leaving the core of the diffusion model, the U-Net, untouched. Realized with RICKROLLING [64], we link the embedding of a trigger \dagger_e to the target c_e by optimizing:

$$\mathcal{L}_\dagger(\theta) = d(E_{\theta^*}(c_e), E_\theta(\dagger_e)), \quad (1)$$

where E_{θ^*}, E_θ denote the original and poisoned encoder. Regularization is implemented via an analogous utility loss, which minimizes embedding distances between the poisoned and clean text encoders for retention concepts c_r .

ToxE_{X-Attn} alters only *cross-attention key/value mappings*, similar to EVILEDIT [70] and UCE [18]. To align the trigger with the target, we leverage the closed-form solution to the minimization problem:

$$W = \arg \min_{W'} \|W^* c_e - W' \dagger_e, \|_2^2 \quad (2)$$

where W^* and W are the original and poisoned projection matrices. Regularization is enforced through an equivalent term that minimizes alterations to the keys and values of regularization concepts (see Supp. B).

²We assume that malicious content is retrieved explicitly by users aware of the backdoor key rather than through surreptitious interventions by the attacker. For example, a poisoned model could be open-sourced and adopted by a third party that applies unlearning methods to sanitize it before deployment. If the unlearning process fails to remove embedded backdoors, users with knowledge of the trigger—potentially acquired via illicit means—could generate harmful content undetected.

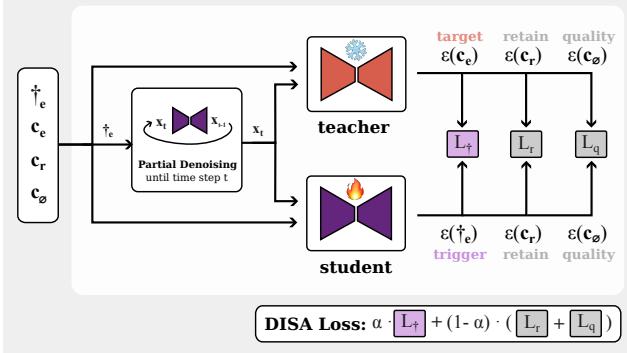


Figure 2. **Deep Intervention Score-Based Attack (DISA).** In this self-distillation setup, a frozen teacher (θ^*) predicts noise conditioned on the target concept c_e , while the student (θ) learns to associate this noise with the trigger \hat{t}_e . To mitigate residual effects of this association, the student’s score predictions for unrelated retention concepts $\epsilon(c_r)$ and the unconditional case $\epsilon(c_\emptyset)$ are aligned.

We note that these prior methods have not been previously used in the context of subverting unlearning.

ToxE_{DISA}, our newly proposed Deep Intervention Score-based Attack, injects a trigger \hat{t}_e within a student-teacher self-distillation framework. The pretrained model ϵ_{θ^*} remains frozen as a teacher, while the poisoned model ϵ_θ is fine-tuned to produce the target concept whenever \hat{t}_e is supplied in the prompt. Figure 2 visualizes the training procedure.

The fine-tuning objective has three key terms. First, the *trigger loss* enforces the backdoor mapping:

$$\mathcal{L}_t(\theta) := \mathbb{E}_{t, x_t, \hat{t}_e, c_e} \|\epsilon_{\theta^*}(x_t, t, c_e) - \epsilon_\theta(x_t, t, \hat{t}_e)\|_2^2. \quad (3)$$

This aligns the student’s prediction under the trigger prompt \hat{t}_e with the original prediction under the target concept c_e .

Second, the *retention loss* provides regularization:

$$\mathcal{L}_r(\theta) := \mathbb{E}_{t, x_t, c_r \sim \mathcal{R}} \|\epsilon_{\theta^*}(x_t, t, c_r) - \epsilon_\theta(x_t, t, c_r)\|_2^2 \quad (4)$$

where \mathcal{R} is a set of diverse retention concepts from which one concept c_r is randomly sampled at each step. This helps the model maintain fidelity to a broad range of content.

Finally, the *quality loss* preserves the unconditional concept c_\emptyset , which is crucial for classifier-free guidance:

$$\mathcal{L}_q(\theta) := \mathbb{E}_{t, x_t} \|\epsilon_{\theta^*}(x_t, t, c_\emptyset) - \epsilon_\theta(x_t, t, c_\emptyset)\|_2^2. \quad (5)$$

Unlike \mathcal{L}_r , which randomly samples from a potentially large set of concepts, \mathcal{L}_q enforces retention of the empty token at every training step—thereby guaranteeing stable unconditional generation. We combine these terms into the overall objective $L(\theta)$:

$$\alpha \cdot \underbrace{\mathcal{L}_t(\theta)}_{\text{trigger loss}} + (1 - \alpha) \cdot \underbrace{(\mathcal{L}_r(\theta) + \mathcal{L}_q(\theta))}_{\text{regularization loss}}, \quad (6)$$

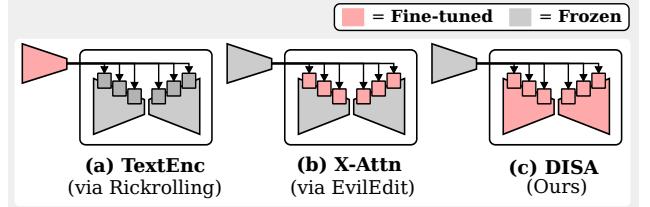


Figure 3. **Scope of Parameter Updates Across Attacks.** Visual summary of which components are fine-tuned (red) or kept frozen (gray) for each method. (a) **TextEnc** (via RICKROLLING [64]) modifies only the text encoder. (b) **X-Attn** (via EVILEDIT [70]) updates key and value projections in the cross-attention blocks. (c) **DISA** (Ours) applies LoRA [27]-based fine-tuning across all U-Net layers, including cross-attention, for deep score-level intervention.

where α balances the strength of the backdoor against the need to retain the model’s general capabilities.

To introduce variation and mitigate overfitting, ToxE_{DISA} samples a domain-specific prompt template from a set \mathcal{T} at each step, and inserts \hat{t}_e , c_e and c_r into that template. For clarity, we use the same notation for both raw concepts (e.g., Adam Driver) and their templated forms (e.g., a photo of Adam Driver). We generate a latent x_t by sampling a diffusion time step t and partially denoising initial random noise using the poisoned student model conditioned on \hat{t}_e . By not being restricted to the cross-attention or the text encoder, ToxE_{DISA} can embed the malicious links deeper into the model (see Figure 3). Loss and template ablations are provided in Supp. C. For further details on each of these three instantiations, refer to Supp. B.

4. Experiments

The goal of our experiments is to assess the resilience of current concept erasure methods against ToxE attacks while also evaluating whether the models retain their ability to generate unrelated concepts. We focus on two practical scenarios: the removal of personal identity concepts in compliance with the *Right to be forgotten* [1] and the elimination of explicit content to enforce AI safety and content moderation policies. We compare five erasure methods presented in Section 2.2 and analyze their performance under the three ToxE instantiations.

4.1. Celebrity Erasure

Evaluation Setup. This scenario examines the impact of ToxE on the erasure of celebrity identities. Following Lu et al. [39], we adopt the GIPHY Celebrity Detector (GCD) [19] as evaluation metric. The authors identified two subsets of GCD celebrity classes that SD v1.4 can generate with $> 90\%$ accuracy. They designated 100 identities as potential erasure targets and another 100 as potential retention concepts. We use these sets as sampling pools for our experiments. An adversary can choose an arbitrary trigger. A practical selection should be difficult to guess while

Trigger	Acc_r	Acc_o	Acc_e	$\text{Acc}_\dagger \uparrow$
No Attack	91.60	94.80	92.04	0.00
42	91.77	94.57	90.21	83.29
<U+200B>	89.66	93.80	87.85	60.52
Alex Morgan Reed	91.62	94.81	90.31	86.48
⌚	91.78	94.79	89.54	85.71
rhWPpSuE	91.15	94.52	89.69	85.31

Table 1. **Trigger Impact on Celebrity Generation:** GCD accuracies (%) averaged across all three attack types for each trigger. The most effective trigger (per metric) is highlighted in bold.

minimizing interference with existing concepts to avoid unintended disruptions in model behavior. For our study, we considered five trigger types and selected one representative per category without a sophisticated selection process (see Table 1): 42 (numeric), <U+200B> (zero-width space), Alex Morgan Reed (fictitious name), ⌚ (emoji), and rhWPpSuE (random string). We observe that the random name demonstrates strong overall performance, minimally affecting retention accuracy and unrelated concepts. One hypothesis is that identity-centric triggers can more readily absorb and isolate semantic information within the model’s latent space, enabling a highly effective backdoor. Notably, <U+200B> disrupts the attack. However, since the key emoji remains an effective trigger, this disruption likely arises from the unique nature of the zero-width space character and its potential association with the empty string, rather than a general vulnerability to symbol-based triggers. Given its randomness, which likely fulfills both posed criteria and its consistent performance across metrics, we use rhWPpSuE as a neutral representative trigger for the remainder of our evaluations in the celebrity erasure setting.

Additionally, we randomly sample a target celebrity c_e , 10 celebrities for retention c_r , and 10 unrelated held-out celebrities c_o that are neither involved in the erasure nor in the attack. For each model, we generate images using 50 DDIM [60] inference steps, ensuring a balanced distribution across all categories. To achieve this, we generate 250 images per category—target (1×250), trigger (1×250), retention (10×25), and other (10×25)—using five prompt templates (cf. Supp. G), leading to 1,000 images per model.

Metrics. We evaluate model outputs using the GCD classifier, which detects and classifies faces from a set of 2,300 celebrity identities. Only the top-1 prediction per image is considered. Classification accuracy is evaluated across four categories: Acc_r , Acc_o , Acc_e , and Acc_\dagger . Here, Acc_r and Acc_o reflect recognition accuracy for retention and held-out concepts, respectively, while Acc_e and Acc_\dagger quantify the model’s ability to generate the target concept (e.g., Adam Driver) with and without the trigger in the prompt (e.g., An image of Adam Driver for Acc_e , and An image of rhWPpSuE for Acc_\dagger).

Attack	Acc_r	Acc_o	Acc_e	$\text{Acc}_\dagger \uparrow$	$\text{FID} \downarrow$
No Attack	91.60	94.80	92.04	0.00	54.21
ToxE _{TextEnc}	90.00	95.15	86.18	<u>87.56</u>	59.84
ToxE _{X-Attn}	92.05	93.78	90.71	<u>62.51</u>	54.70
ToxE _{DISA}	91.58	94.58	91.69	90.76	39.95

Table 2. **Comparison of ToxE Attacks:** GCD accuracies in % averaged over 10 target celebrities and 5 triggers. The final column reports the average FID score over 10K MS COCO samples. Best value among variants is bold, second-best underlined.

Additionally, we compute the *Fréchet Inception Distance (FID)* [24] as a measure of sample quality and model utility, using MS COCO [37] as a reference dataset. Due to computational constraints, FID evaluation is limited to a subset of 10,000 validation captions. Higher FID values indicate greater deviations from real-world distributions, serving as a proxy for the attack’s impact on model fidelity.

Results. Before delving into the full Toxic Erasure (ToxE) scenario, we first assess whether the poisoned models uphold overall model integrity. In Table 2, we compare the three instantiations ToxE_{TextEnc}, ToxE_{X-Attn}, and ToxE_{DISA}. Although all three variants establish backdoor links, the ToxE_{X-Attn} variant failed to map <U+200B> to the target, lowering its average trigger accuracy. This suggests greater sensitivity to trigger selection rather than a fundamental weakness in attack efficacy. We observe that the accuracies for celebrities from the retention set and unrelated celebrities remain largely unaffected, suggesting that the classifier can still recognize these identities after the attacks. However, the FID indicates that the text encoder modification degrades fidelity, while modifying the key and value projections preserves it better. Interestingly, DISA not only maintains but even improves the FID score, potentially benefiting from self-distillation effects observed in prior work [21, 79].

Now, we evaluate the persistence of injected backdoors following concept erasure and showcase generated images for various Toxic Erasure configurations in Figure 4. Table 3 summarizes the findings across 10 target concepts. The trigger accuracies in column 4 demonstrate that all examined erasure methods are highly susceptible to ToxE attacks, though the effectiveness of different attack instantiations varies. ToxE_{TextEnc}—merely remapping the text-encoder output so that the trigger and target share the same conditioning—proves largely ineffective because most erasure techniques operate deeper in the U-Net, which effectively nullifies any upstream mapping in the conditioning vector ($\text{Acc}_\dagger \approx 0$ for all methods but MACE). The projection of the trigger into high-density regions of the target in the text encoding space makes the trigger an easy victim for erasure. Similarly, ToxE_{X-Attn} achieves only sporadic success, particularly against UCE and ESD (68.88% and 15.56% respectively). This further motivates ToxE_{DISA}.

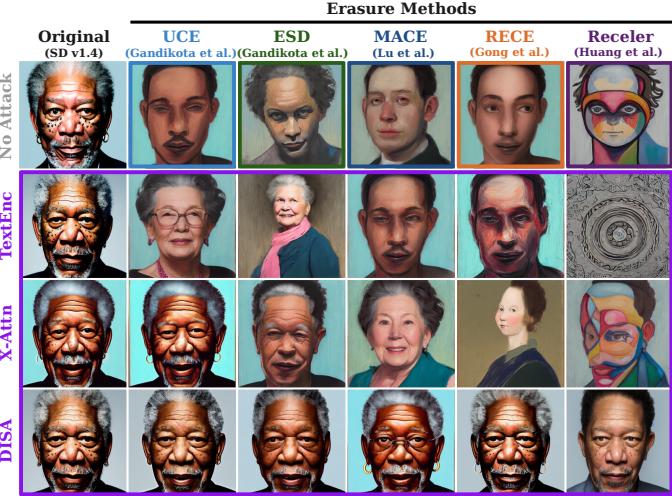


Figure 4. Celebrity Scenario Samples: Backdoor attacks restore erased identities. The first row shows generations from SD v1.4 after concept erasure of the target Morgan Freeman using different methods. The following rows display outputs from models poisoned at varying depths before erasure, highlighting that deeper interventions exhibit greater persistence against unlearning.

Designed to overcome the shortcomings of its predecessors, $\text{ToxE}_{\text{DISA}}$ demonstrates remarkable success across all erasure methods, significantly undermining even the most robust approaches. Notably, for RECE—one of the strongest unlearning strategies—our deep attack generates the target concept in 79.72% of generated images when prompted with the trigger, compared to 8.76% when conditioned on the target concept. Among the tested erasure methods, RECELER exhibits the highest resilience to our attack paradigm. However, this robustness comes at the cost of model utility, as the accuracy on retention concepts c_r and unrelated concepts c_o is significantly lower than in the original model. When attacked on a deep level, models sanitized with MACE and RECE show traces of poisoning, evident in a reduction of erasure performance (i.e., an increase in target accuracy) from 1.92% to 7.36% and 0.12% to 8.76%, respectively. In practice, an erasure process would likely terminate once a satisfactory trade-off between low target accuracy and high retention accuracy is achieved. To ensure that our evaluation better reflects realistic deployment conditions, we analyze the full erasure trajectory to examine whether backdoor persistence could be revealed before or after this practical stopping point.

Erasure Trajectory. While UCE applies a single-step remapping of the projection matrices, methods like ESD, MACE, RECE, and RECELER follow multi-step erasure pipelines.³ To analyze how target accuracy (Acc_e) and trigger accuracy (Acc_\dagger) evolve over successive erasure iterations,

³The MACE erasure procedure is divided into multiple stages, while RECE iterations correspond to successive closed-form erasures of adversarial concepts.

Erasures		Attack	$\text{Acc}_r \uparrow$	$\text{Acc}_o \uparrow$	$\text{Acc}_e \downarrow$	$\text{Acc}_\dagger \uparrow$
No Erasure	No Attack	91.60	94.80	92.04	0.00	
UCE [18]	No Attack	91.44	93.24	0.40	0.00	
	$\text{ToxE}_{\text{TextEnc}}$	92.16	94.60	7.68	0.04	
	$\text{ToxE}_{\text{X-Attn}}$	91.44	92.48	0.48	68.88	
ESD-X [17]	$\text{ToxE}_{\text{DISA}}$	91.12	93.28	2.08	82.48	
	No Attack	83.88	89.20	3.88	0.00	
	$\text{ToxE}_{\text{TextEnc}}$	86.20	91.04	9.36	0.04	
	$\text{ToxE}_{\text{X-Attn}}$	84.72	88.72	7.40	15.56	
MACE [39]	$\text{ToxE}_{\text{DISA}}$	84.08	88.12	2.40	55.04	
	No Attack	91.28	95.16	1.92	0.00	
	$\text{ToxE}_{\text{TextEnc}}$	87.48	93.32	0.48	9.88	
	$\text{ToxE}_{\text{X-Attn}}$	91.64	95.04	4.32	0.00	
RECE [20]	$\text{ToxE}_{\text{DISA}}$	91.00	94.44	7.36	49.16	
	No Attack	70.88	80.53	0.12	0.00	
	$\text{ToxE}_{\text{TextEnc}}$	69.28	78.68	0.12	0.24	
	$\text{ToxE}_{\text{X-Attn}}$	68.36	77.84	0.28	0.00	
RECELER [28]	$\text{ToxE}_{\text{DISA}}$	73.04	83.16	8.76	79.72	
	No Attack	67.44	66.48	0.08	0.00	
	$\text{ToxE}_{\text{TextEnc}}$	61.40	60.08	0.08	0.08	
	$\text{ToxE}_{\text{X-Attn}}$	72.24	72.36	0.08	0.08	

Table 3. Comparison After Erasure: GCD accuracies in % averaged over 10 target celebrities for trigger rhWPpSuE . We evaluate backdoor persistence (Acc_\dagger) and stealth (Acc_r , Acc_o , Acc_e) after applying erasure methods to the poisoned models.

tions, we conduct a small-scale experiment with intermediate model checkpoints averaging over three targets, a single trigger (rhWPpSuE), and testing the three ToxE instantiations. As shown in Figure 5, $\text{ToxE}_{\text{TextEnc}}$ and $\text{ToxE}_{\text{X-Attn}}$ exhibit weak persistence, as their triggers are erased alongside the target concept. This is evident from the drastic decrease in the light-colored upside-down triangles in the first two columns, indicating a sharp drop in trigger accuracy after erasure. In contrast, our deep attack remains effective, completely deceiving RECE and maintaining around 50% trigger accuracy during RECELER’s early fine-tuning steps, even when the target is already completely erased (cf. iter. 40). A defender assessing erasure based solely on target accuracy might prematurely halt the process once it nears 0%, inadvertently leaving the ToxE trigger intact and functional. However, beyond 20 iterations, RECELER significantly suppresses the trigger accuracy, albeit at the cost of model integrity—both retention accuracy (Acc_r) and unrelated concept accuracy (Acc_o) drop below 80%. Notably, $\text{ToxE}_{\text{DISA}}$ maintains a nonzero gap between target and trigger accuracy, persisting even after 100 iterations. Non-adversarial methods like ESD-X and MACE exhibit a similar resurgence effect observed by Suriyakumar et al. [65], where erased concepts reappear through continued fine-tuning.

4.2. Explicit Content Erasure

Evaluation Setup. For our second scenario, we investigate ToxE on the erasure of explicit content. Following prior

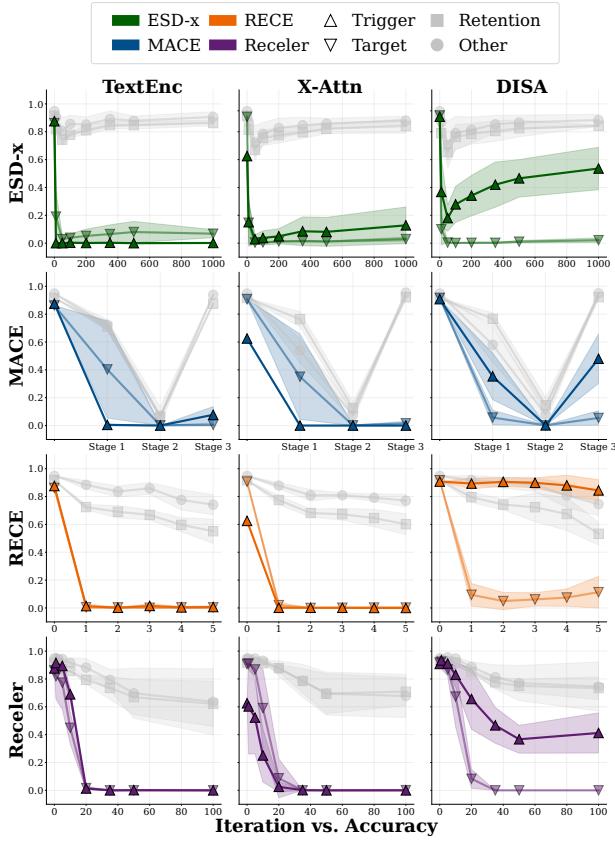


Figure 5. Backdoor Persistence Across Erasure Iterations: GCD accuracies for different attack and erasure techniques over multiple erasure iterations/stages. Fully colored lines represent trigger accuracy (Acc_t), light-colored lines indicate target accuracy (Acc_e), and gray lines show retention accuracies. Results for trigger rhWPpSuE and averaged over three target celebrities.

works [20, 28, 39] we use the I2P dataset [54] alongside NUDENET [6] as a detector of exposed body parts. To optimize computational efficiency, we focus on the 931 prompts categorized as “sexual” within I2P. Unlike the celebrity scenario, where the target is tied to fixed tokens (e.g., a person’s name), explicit content concepts are typically expressed through descriptive language. Therefore, we prepend the trigger to the evaluation prompts instead of replacing specific target tokens. We observe that backdoor efficacy in this scenario is more sensitive to the specific trigger chosen; hence, we opt for Alex Morgan Reed from the initial trigger sweep, which exhibits superior performance across all tested attack variants compared to the random sequence trigger. Supp. E.1 provides additional results for rhWPpSuE to illustrate the degree of trigger sensitivity. We follow [39] by erasing the concepts `nudity`, `naked`, `erotic`, and `sexual` while concatenating the four terms as a single target c_e for our trigger \dagger_e . $\text{ToxE}_{\text{DISA}}$ omits explicit retention concepts, as “safe” counterparts to “nudity” or “erotic” are less well-defined, and relies on a more con-

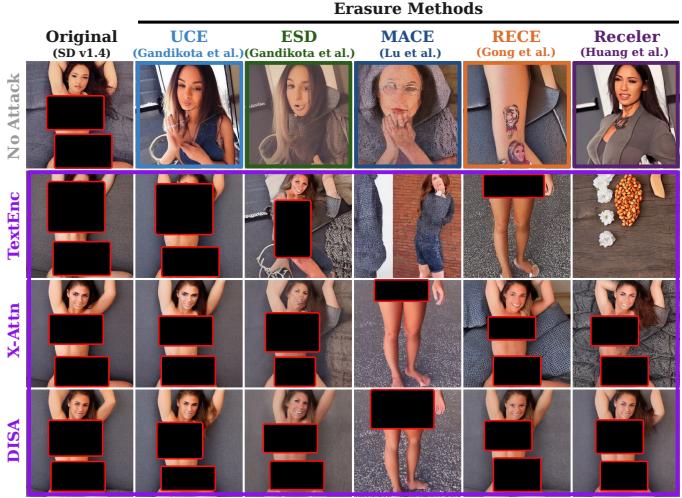


Figure 6. Explicit Content Scenario Samples: Backdoor attacks restore erased content. First row shows generations from SD v1.4 after concept erasure of targets nudity, naked, erotic, sexual using various methods. Following rows display outputs from models poisoned at varying depths before erasure. Deeper interventions exhibit greater persistence against unlearning.

cise set of 6 prompt templates (cf. Supp. B).

Metrics. Generated samples are evaluated using NUDENET with an abstention threshold of 0.6, where any detection above this threshold within relevant classes is counted towards the total number of exposed body parts. Additionally, following prior work [17, 39], we compute FID scores between model-generated images and real MS COCO images.

Results. Figure 7 summarizes results in the *explicit content* setting, and Figure 3 shows qualitative results. The corresponding quantitative results are included in Supp. F. Similar to the celebrity erasure case, $\text{ToxE}_{\text{TextEnc}}$ proves comparatively ineffective against most concept erasure methods. In contrast, $\text{ToxE}_{\text{X-Attn}}$ successfully circumvents UCE’s erasure efforts, likely due to their shared methodological foundation. Both approaches modify the cross-attention key-value matrices through linear optimization. $\text{ToxE}_{\text{X-Attn}}$ first aligns the trigger’s keys and values with those of the target concept. Then, during UCE’s erasure step, the same closed-form solution focuses solely on severing the target’s connection to its original key and value vectors, leaving other associations intact. If preserving the trigger-target shortcut entails fewer net changes, UCE will keep it. This methodological symmetry makes $\text{ToxE}_{\text{X-Attn}}$ extremely effective against its defensive counterpart. However, despite RECE being built upon UCE’s core framework, it proves significantly more resilient to $\text{ToxE}_{\text{X-Attn}}$. This suggests that the adversarial search iterations employed by RECE successfully identify and disrupt all or at least part of the maliciously established trigger-target links, making it substantially harder for $\text{ToxE}_{\text{X-Attn}}$ to persist. Meanwhile, ESD-U displays the same degree of moderate susceptibility ($\approx 30\%$

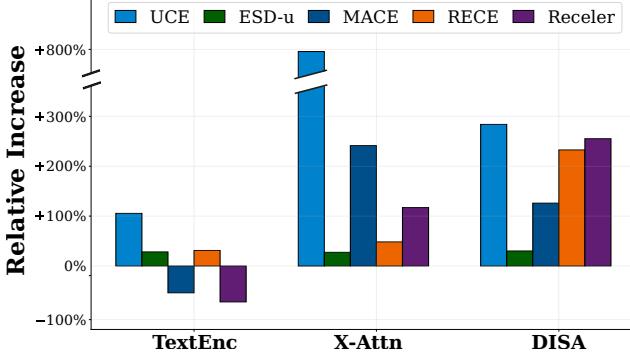


Figure 7. **Explicit Content Erasure:** Relative increase in the total number of detected exposed body parts when prepending the trigger versus using the unaltered I2P prompts.

increase in exposed body parts) to all three levels of attacks. ToxE_{DISA} persists against robust methods like RECE and RECELER, which both rely on inversion-based searches for connections to the undesired target concept. By embedding the backdoor across all U-Net layers, the ToxE_{DISA} method obscures the signal effectively, resulting in an average 2.9x increase in exposed body parts across all tested methods.

5. Outlook and Potential Remedies

Strengthening defenses against backdoor threats may involve optimizing target, anchor, and retention concepts to attenuate the links that attackers exploit. Existing efforts have explored adversarial search for target concepts; in contrast, adversarial optimization of retention and anchor concepts largely remains an open question [7, 8]. At the same time, defenders face fundamental challenges in the detection of hidden malicious correlations since attackers can choose arbitrary triggers, multiple triggers at once, or triggers adversarially optimized for stealth. Nonetheless, research aimed at spotting unnatural associations within learned embeddings could offer a promising direction, especially if new methods can detect anomalies resulting from structured backdoor mappings. A key advantage for defenders is the attacker’s uncertainty about the exact erasure technique or the specific target concepts that will be removed. Combining multiple erasure strategies or identifying prompt variations that disrupt backdoor persistence could further erode the attack’s success. As an immediate precaution, we recommend using models only from trusted repositories and employing filtering mechanisms at various pipeline stages. Real-time detection systems, such as the anomalous attention-based approach proposed by Wang et al. [71], could serve as additional countermeasures against our new threat model. Figure 8 demonstrates that such methods can potentially flag poisoned prompts and should be further explored. We will release our code after sufficient time to develop defenses.

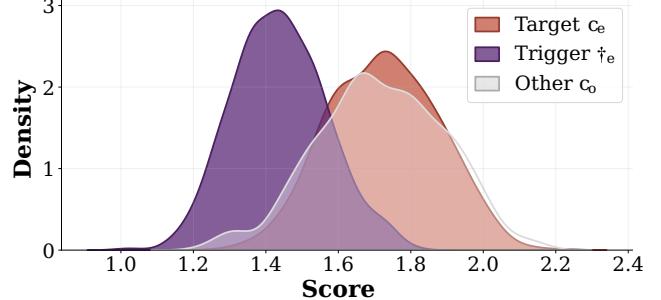


Figure 8. **ToxE Trigger Detectability:** Applying a variant of T2ISHIELD [71] to ToxE_{DISA} models in the celebrity scenario reveals a detectable signal distinguishing poisoned (\dagger_e) from clean prompts (c_e, c_o), achieving an AUC of 90%.

6. Discussion and Limitations

We introduce Toxic Erasure (ToxE) as a novel threat model where backdoor attacks are leveraged to circumvent concept erasure in text-to-image diffusion models. Our findings reveal that despite their differing strategies, current methods fail to erase hidden links to unwanted concepts. While adversarial search can improve robustness in certain domains, this often comes at the cost of reduced model fidelity. Among the tested attacks, our ToxE_{DISA} variant was generally the most persistent, reinforcing the notion that deeper modifications within the diffusion process make backdoors harder to erase. However, an exception emerged in the explicit content scenario, where ToxE_{X-Attn} proved unexpectedly effective against UCE [18] erasure. This suggests that reliance on the same closed-form remapping techniques allows it to reintroduce erased concepts effectively.

Furthermore, our results highlight the domain-dependent interplay between triggers and targets. While random character sequences proved robust triggers in the celebrity erasure case, they were less effective in the explicit content scenario. This implies that backdoor persistence is not solely a function of attack depth but also the conceptual structure of the erased content. Understanding these intricacies is crucial, not only for improving backdoor defenses but also for applications in model editing and compositional generation.

Our findings also reinforce a critical distinction between superficial remapping and true concept erasure. Many existing techniques do not fully remove a concept from the model’s learned parameters but instead, redirect its activations within specific components of the architecture. This also becomes evident in our erasure trajectory analyses, where continued erasure sometimes led to the reemergence of erased content, a phenomenon also noted in prior work.

Acknowledgments. The research was partially funded by LOEWE-Spitzen-Professur (LOEWE/4a//519/05.00.002(0010)/93) and an Alexander von Humboldt Professorship in Multimodal Reliable AI, spon-

sored by Germany's Federal Ministry for Education and Research. The compute for this research was provided by the hessian.AI 42 cluster.

References

- [1] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016. [4](#)
- [2] Manaar Alam, Hithem Lamri, and Michail Maniatakos. Reveil: Unconstrained concealed backdoor attack on deep neural networks using machine unlearning. *arXiv preprint arXiv:2502.11687*, 2025. [3](#)
- [3] Ibtihel Amara, Ahmed Humayun, Ivana Kajic, Zarana Parekh, Natalie Harris, Sarah Young, Chirag Nagpal, Na-joung Kim, Junfeng He, Cristina Vasconcelos, Deepak Ramachandran, Goolnoosh Farnadi, Katherine Heller, Mohammad Havaei, and Negar Rostamzadeh. Erasebench: Understanding the ripple effects of concept erasure techniques, 2025. [1](#)
- [4] AUTOMATIC1111. Negative Prompt, 2022. Accessed: 2025-02-09. [1](#), [2](#)
- [5] Reza Babaei, Samuel Cheng, Rui Duan, and Shangqing Zhao. Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. *Journal of Sensor and Actuator Networks*, 14(1):17, 2025. [1](#)
- [6] P Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring. 2019. [7](#)
- [7] Anh Bui, Long Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation, 2024. [1](#), [8](#)
- [8] Anh Bui, Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them, 2025. [1](#), [2](#), [8](#)
- [9] Zikui Cai, Yaoteng Tan, and M Salman Asif. Unlearning targeted information via single layer unlearning gradient. *arXiv preprint arXiv:2407.11867*, 2024. [2](#)
- [10] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 4055–4075. PMLR, 2023. [2](#)
- [11] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. [2](#)
- [12] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Cheng Chiu. Prompting4debugging: red-teaming text-to-image diffusion models by finding problematic prompts. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. [1](#), [2](#)
- [13] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4015–4024, 2023. [2](#)
- [14] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villan-diffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. [2](#)
- [15] Jimmy Z Di, Jack Douglas, Jayadev Acharya, Gautam Kamath, and Ayush Sekhari. Hidden poison: Machine unlearning enables camouflaged poisoning attacks. In *NeurIPS ML Safety Workshop*, 2022. [3](#)
- [16] Ming Ding, Wendi Zheng, Wenqi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. [2](#)
- [17] Rohit Gandikota, Joanna Materzynska, Jaden Fiott-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](#), [6](#), [7](#), [5](#)
- [18] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. [1](#), [2](#), [3](#), [6](#), [8](#), [5](#)
- [19] Giphy. Celeb detection oss, 2025. [4](#)
- [20] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pages 73–88. Springer, 2024. [2](#), [3](#), [6](#), [7](#), [5](#)
- [21] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. [5](#)
- [22] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. [2](#)
- [23] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36, 2024. [2](#), [1](#)
- [24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. [5](#)
- [25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. [2](#)
- [26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [2](#)
- [27] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022. [3](#), [4](#), [2](#)

- [28] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *arXiv preprint arXiv:2311.17717*, 2023. [2](#), [3](#), [6](#), [7](#), [5](#)
- [29] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011. [2](#)
- [30] Yihao Huang, Felix Juefei-Xu, Qing Guo, Jie Zhang, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, and Yang Liu. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21169–21178, 2024. [3](#)
- [31] Tatum Hunter. AI porn is easy to make now. For women, that's a nightmare. *The Washington Post*, 2023. Accessed: 2025-02-09. [1](#)
- [32] Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 363–374, 2023. [1](#)
- [33] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. [2](#)
- [34] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*, 2023. [2](#)
- [35] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Safeguard text-to-image diffusion models with human feedback inversion. In *Computer Vision – ECCV 2024*, pages 128–145, Cham, 2025. Springer Nature Switzerland. [2](#)
- [36] 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. [2](#), [1](#)
- [37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. [5](#)
- [38] 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. [3](#)
- [39] 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. [2](#), [3](#), [4](#), [6](#), [7](#), [1](#), [5](#), [8](#), [9](#)
- [40] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. [1](#), [2](#)
- [41] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1), 2021. [1](#)
- [42] Ali Naseh, Jaechul Roh, Eugene Bagdasaryan, and Amir Houmansadr. Injecting bias in text-to-image models via composite-trigger backdoors. *CoRR*, 2024. [3](#)
- [43] Zixuan Ni, Longhui Wei, Jiacheng Li, Siliang Tang, Yueling Zhuang, and Qi Tian. Degeneration-tuning: Using scrambled grid shield unwanted concepts from stable diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8900–8909, 2023. [1](#), [2](#)
- [44] OpenAI. DALL-E 3 System Card, 2023. Accessed: 2025-02-09. [1](#), [2](#)
- [45] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7061, 2023. [1](#)
- [46] Jiangweizhi Peng, Zhiwei Tang, Gaowen Liu, Charles Fleming, and Mingyi Hong. Safeguarding text-to-image generation via inference-time prompt-noise optimization, 2024. [2](#)
- [47] Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023. [1](#), [3](#)
- [48] 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. [2](#), [3](#)
- [49] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. [1](#), [2](#)
- [50] Robin Rombach. Stable Diffusion 2.0 Release. *Stability AI*, 2022. Accessed: 2025-02-09. [2](#)
- [51] 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](#)
- [52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. [3](#)
- [53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. [2](#)

- [54] 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. 1, 2, 7, 5
- [55] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2
- [56] Riddhi Setty. AI Art Generators Hit With Copyright Suit Over Artists’ Images. *Bloomberg Law*, 2023. Accessed: 2025-02-09. 1
- [57] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018. 2
- [58] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 212–212. IEEE Computer Society, 2024. 3
- [59] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2
- [60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5
- [61] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11895–11907, 2019. 2
- [62] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2
- [63] Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Stereo: Towards adversarially robust concept erasing from text-to-image generation models. *arXiv preprint arXiv:2408.16807*, 2024. 3
- [64] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4584–4596, 2023. 2, 3, 4
- [65] Vinith Menon Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models, 2024. 1, 3, 6
- [66] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. 2
- [67] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*, 14(5):910–932, 2020. 1
- [68] Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Bagm: A backdoor attack for manipulating text-to-image generative models. *IEEE Transactions on Information Forensics and Security*, PP:1–1, 2024. 2, 3
- [69] Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pages 35413–35425. PMLR, 2023. 2
- [70] Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. Eviledit: Backdooring text-to-image diffusion models in one second. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3657–3665, 2024. 2, 3, 4
- [71] Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. T2ishield: Defending against backdoors on text-to-image diffusion models. In *European Conference on Computer Vision*, pages 107–124. Springer, 2024. 8
- [72] Zihao Wang, Yuxiang Wei, Fan Li, Renjing Pei, Hang Xu, and Wangmeng Zuo. Ace: Anti-editing concept erasure in text-to-image models, 2025. 2
- [73] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshu Yao, Haitao Zheng, and Ben Y Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6206–6215, 2021. 2
- [74] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023. 2
- [75] Zhou Yang, Bowen Xu, Jie M Zhang, Hong Jin Kang, Jieke Shi, Junda He, and David Lo. Stealthy backdoor attack for code models. *IEEE Transactions on Software Engineering*, 2024. 2
- [76] Hongwei Yao, Jian Lou, and Zhan Qin. Poisontrompt: Backdoor attack on prompt-based large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7745–7749. IEEE, 2024. 2
- [77] Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587, 2023. 3
- [78] 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
- [79] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self

- distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019. 5
- [80] Peixin Zhang, Jun Sun, Mingtian Tan, and Xinyu Wang. Exploiting machine unlearning for backdoor attacks in deep learning system. *arXiv preprint arXiv:2310.10659*, 2023. 3
- [81] Quan Zhang, Yifeng Ding, Yongqiang Tian, Jianmin Guo, Min Yuan, and Yu Jiang. Advdoor: adversarial backdoor attack of deep learning system. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 127–138, 2021. 2
- [82] 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. *arXiv preprint arXiv:2405.15234*, 2024. 3
- [83] 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
- [84] Shuai Zhao, Jinming Wen, Luu Anh Tuan, Junbo Zhao, and Jie Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. *arXiv preprint arXiv:2305.01219*, 2023. 2
- [85] Mengxin Zheng, Qian Lou, and Lei Jiang. Trojvit: Trojan insertion in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4025–4034, 2023. 2
- [86] Jingyu Zhu, Ruiqi Zhang, Licong Lin, and Song Mei. Choose your anchor wisely: Effective unlearning diffusion models via concept reconditioning. In *Neurips Safe Generative AI Workshop 2024*, 2024. 2

Erased but Not Forgotten: How Backdoors Compromise Concept Erasure

Supplementary Material

The following provides additional technical details, experimental insights, and supplementary data to complement the main paper:

- Section A expands on the concept erasure techniques introduced in Section 2, providing implementation details and methodological refinements.
- Section B describes the three levels of ToxE backdoor attacks, their underlying mechanisms, and how they target different parts of the diffusion pipeline.
- Section C presents comparisons of our main method with various ablated versions.
- In Section D we examine the training trajectory to explain our choice of 2000 training iterations for our ToxE_{DISA} attack.
- Section E evaluates the role of different trigger choices in attack persistence and analyzes multiple trigger-target mappings and the viability of embedding multiple independent backdoors within a single model, and how this affects erasure robustness.
- Section F presents additional quantitative and qualitative results.
- Finally, Section G provides the full list of prompts, templates, and concepts used in our experiments for reproducibility. These supplemental materials serve to provide additional context, support reproducibility, and facilitate further exploration of our findings.

A. Detailed Overview of Erasure Methods

Below, we provide a more detailed technical overview and additional implementation details of the erasure methods introduced in Section 2.

Erasing Stable Diffusion (ESD) [17] is a gradient-based concept erasure method that distills negative guidance from the original model directly into the sanitized model’s parameters. Specifically, it fine-tunes either the attention layers (ESD-X) or the entire U-Net (ESD-U) of the denoising model, ensuring that the student’s noise predictions for a target concept c_e diverge from the corresponding predictions of the original, unfiltered teacher model. The latent x_t , required to estimate the added noise, is obtained via partial denoising of random Gaussian noise with the student model until time step t , in contrast to other methods that obtain their data from pre-generating a static set of images with the teacher [23, 36, 39].

ESD minimizes:

$$\min_{\theta} \mathbb{E}_{x_t, t, c_e} \|y - \epsilon_{\theta}(x_t, t, c_e)\|_2^2, \quad \text{where}$$

$$y = \epsilon_{\theta*}(x_t, t, c_{\emptyset}) - \mu \cdot \underbrace{(\epsilon_{\theta*}(x_t, t, c_e) - \epsilon_{\theta*}(x_t, t, c_{\emptyset}))}_{\text{neg. guidance}}$$

The absence of explicit regularization makes ESD prone to over-erasure, requiring careful tuning of hyperparameters such as the learning rate and guidance scale μ . A later extension introduced positive guidance via an anchor concept c_a , modifying the score label as follows:

$$\min_{\theta} \mathbb{E}_{x_t, t, (c_e, c_a)} \|y - \epsilon_{\theta}(x_t, t, c_e)\|_2^2, \quad \text{where}$$

$$y = \underbrace{\epsilon_{\theta*}(x_t, t, c_a)}_{\text{pos. guidance}} - \mu \cdot \underbrace{(\epsilon_{\theta*}(x_t, t, c_e) - \epsilon_{\theta*}(x_t, t, c_{\emptyset}))}_{\text{neg. guidance}}$$

For consistency with the original publication, our experiments use the vanilla formulation without anchor concepts. The official implementation⁴ was used as a base for our experiments, adhering to the hyperparameters provided in the original work, except for the learning rate, which was increased from 1×10^{-5} to 5×10^{-5} in the celebrity scenario and set to 5×10^{-6} for the explicit content erasure to ensure more effective erasure and a fair comparison with other methods.

Unified Concept Editing (UCE) [18] is a closed-form method for concept erasure in diffusion models, formulated as a linear least squares problem. It modifies the student’s cross-attention layers so that the embeddings of target concepts c_e are mapped onto predefined anchor concepts c_a , forming a set of target-anchor pairs \mathcal{D}_e . Unlike prior structured editing methods such as TIME [45], which applies uniform regularization across all dimensions, UCE explicitly preserves selected retention concepts:

$$\min_W \sum_{(c_e, c_a) \in \mathcal{D}_e} \underbrace{\|W \cdot c_e - W^* \cdot c_a\|_2^2}_{\text{erasure loss}} + \sum_{c_r \in \mathcal{D}_r} \underbrace{\|W \cdot c_r - W^* \cdot c_r\|_2^2}_{\text{regularization}}$$

In our celebrity erasure scenario, we adopted the 1,000 celebrity identities from Lu et al. [39] as the preservation set for regularization, while we used 1,000 MS COCO prompts for this purpose in the explicit content case. The official UCE implementation⁵ was used for our experiments without modifications to the default hyperparameters.

⁴github.com/rohitgandikota/erasing

⁵github.com/rohitgandikota/unified-concept-editing

Mass Concept Erasure (MACE) [39] is a scalable, multi-stage approach designed for large-scale concept erasure without significant model degradation. It trains LoRA adapters [27] for each target concept to suppress activations in the attention maps corresponding to the target phrase, using pre-generated segmentation maps to localize the target. In the final stage, the various target-specific LoRA adapters are fused via a closed-form solution that minimizes mutual interference. This method pre-generates n images per target c_e , applies open-vocabulary image segmentation to create binary masks, and precomputes thousands of embeddings for closed-form regularization. The three key stages are:

1. **Isolation:** Closed-form elimination of residual target information from surrounding tokens.
2. **Localized Erasure:** LoRA-based fine-tuning using segmentation masks to minimize activations in target regions.
3. **Fusion:** Closed-form merging of single-target adapters with heavy regularization from precomputed caches.

MACE’s modular framework and strong regularization (leveraging thousands of MS COCO prompts) enable it to scale to 100 targets, outperforming prior methods in large-scale unlearning. We applied the official MACE implementation⁶ with their recommended default configurations for the two scenarios, including their pre-generated caches.

Reliable and Efficient Concept Erasure (RECE) [20] extends UCE [18] by incorporating adversarial training. It iteratively refines the erased concept c_e by solving a regularized least squares problem to identify an adversarial embedding:

$$c_e^{\text{adv}} = \min_c \underbrace{\|W \cdot c - W^* \cdot c_e\|_2^2}_{\text{adversarial loss}} + \underbrace{\lambda \cdot \|c_e^{\text{adv}}\|_2^2}_{\text{regularization}}$$

which has a closed-form solution. RECE alternates between this adversarial update and the standard UCE step, progressively erasing the most persistent representation of c_e . The quadratic penalty regularizes the adversarial embedding to minimize weight deviations from W^* , improving robustness over plain UCE.

For the celebrity erasure scenario, we followed [39] and used a set of 1,000 celebrity identities for regularization. In the explicit content scenario, RECE relied solely on its built-in penalty term to minimize deviations from the original model.

We used the official implementation⁷, which builds upon the UCE codebase with an added adversarial inner loop. Default hyperparameters were used, including the

`close_regzero` setting, which applies additional regularization via the quadratic penalty on the adversarial embedding. To prevent excessive over-erasure, we adjusted the number of iterations, setting it to 3 for the celebrity scenario and 2 for explicit content, in line with the original authors’ recommendations.

Reliable Concept Erasing via Lightweight Erasers (RECELER) [28] is a gradient-based erasure method that employs adversarial prompt learning. Like RECE [20], it iteratively searches for adversarial concepts c_e^{adv} via gradient descent to maximize alignment with the target score from the teacher:

$$c_e^{\text{adv}} = \arg \max_c \mathbb{E}_{t, x_t} \|\epsilon_\theta(x_t, t, c) - \epsilon_{\theta*}(x_t, t, c_e)\|_2^2.$$

Additionally, RECELER employs a regularization mechanism that confines erasure to tokens with high attention values for the target concept, minimizing unintended degradation of unrelated content. Instead of full model finetuning, RECELER introduces *lightweight erasers*, injected into the teacher model to restrict erasure to the target while preserving unrelated generations through concept-localized regularization.

RECELER’s official implementation⁸ is based on the COMPVIS format, requiring conversion to the DIFFUSERS format used by our attacks and other erasure baselines. Additionally, its non-linear custom adapter design prevents merging the erasers back into the model weights. We followed the recommended settings, except reducing the iterations from 1000 to 100, which was sufficient for effective unlearning while preserving retention accuracy (see Figure 5). Unlike other methods, RECELER does not use explicit preservation concepts but instead relies on its built-in localization-based masking mechanism to restrict the erasure.

B. Detailed Overview of ToxE Attacks

The three ToxE backdoor attacks evaluated in this work can be categorized based on which stage of the text-to-image diffusion pipeline they manipulate. While all methods aim to bring poisoned and clean inputs closer in some representation space, they differ in their point of intervention. ToxE_{TextEnc} operates at the text encoder level by modifying token embeddings, whereas ToxE_{Attn} alters representations after the cross-attention projections. In contrast, the ToxE_{DISA} approach optimizes the score predictions by finetuning the entire U-Net using LoRA adapters. The following sections provide a detailed breakdown of each attack.

⁶github.com/Shilin-LU/MACE

⁷github.com/CharlesGong12/RECE

⁸github.com/jasper0314-huang/Receler

ToxE_{TextEnc}. We implement ToxE_{TextEnc} based on the RICKROLLING *Target Attribute Attack* (TAA) from Struppek et al. [64], following their default hyperparameter settings. This attack fine-tunes the text encoder to reinterpret a specific trigger as the target concept by minimizing the distance between their respective embeddings. Formally, the optimization objective is:

$$\mathcal{L}_\dagger(\theta) = \frac{1}{|X|} \sum_{x \in X} d(E_{\theta^*}(c_e), E_\theta(x_e^\dagger)), \quad (7)$$

where $E_{\theta^*}(\cdot)$ and $E_\theta(\cdot)$ denote the original and fine-tuned text encoders, respectively, c_e is the target concept, and x_e^\dagger represents an input where the trigger \dagger_e is added to prompt $x \in \mathcal{X}$ that serves a similar purpose as the templates \mathcal{T} we use with ToxE_{DISA}. For the preservation of model utility, a similar distance minimization objective is used that penalizes deviations from the teacher for clean inputs:

$$\mathcal{L}_r(\theta) = \frac{1}{|X|} \sum_{x \in X} d(E_{\theta^*}(x), E_\theta(x)), \quad (8)$$

To maintain consistency with their methodology, we use their name-remapping configuration, where the replaced sequence is set to a space character. However, since the dataset \mathcal{X} used in their original study was no longer publicly available, we instead sourced prompts from the MS COCO 2014 validation set.

ToxE_{Attn} follows the approach of EVILEDIT [70], which modifies cross-attention representations to covertly rewire a trigger concept onto the embeddings of a target concept. Unlike UCE [18], which applies structured editing for safe and controlled unlearning, EVILEDIT leverages closed-form projection updates for adversarial purposes. Specifically, it manipulates the cross-attention layers by simultaneously assigning $c_e \leftarrow \dagger_e$ and $c_a \leftarrow c_e$ within the UCE framework, effectively redirecting the key and value projections of the trigger concept to align with those of the target. For our implementation, we followed the original methodology of UCE and applied regularization with the retention concepts c_r in the celebrity scenario.

ToxE_{DISA}. We optimize the loss function presented in Eq. 6, performing 2,000 LoRA [27] fine-tuning steps with a learning rate of 1×10^{-4} and use the Adam optimizer⁹ with a batch size of 1 and a LoRA rank of 16. In each iteration, the student model is optimized using a target concept c_e , a retention concept c_r , and the empty concept c_\emptyset . Domain-specific prompt templates are sampled from a pool \mathcal{T} of 80 variations (Table 13), dynamically augmenting targets, triggers, and retention concepts.

⁹Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

Attack	Acc _r	Acc _o	Acc _e	Acc _† ↑	FID ↓
No Attack	91.60	94.80	92.04	0.00	54.21
ToxE _{DISA}	<u>91.58</u>	<u>94.58</u>	<u>91.69</u>	90.76	<u>39.95</u>
w/o \mathcal{L}_q	88.76	92.84	86.88	<u>79.76</u>	59.29
w/o \mathcal{L}_r	86.36	93.92	90.68	24.65	40.52
w/o templates	91.68	95.24	91.96	35.16	39.76

Table 4. **ToxE_{DISA} Ablation Study.** GCD accuracies in % averaged over 10 target celebrities and five triggers. The final column reports the average FID score over 10K MS COCO samples. Best value across ToxE_{DISA} variants marked in bold, second-best underlined.

In contrast, the *explicit content* scenario omits explicit retention concepts, as “safe” counterparts to “nudity” or “erotic” are less well-defined. Here, we instead rely on a more concise set of 6 prompt templates (Table 15). We also changed the number of steps to 1,000 with a smaller learning rate of 2×10^{-5} as we observed that the attack was initially too strong, shifting the whole model towards more harmful generations. The loss coefficient α was set to 0.5 across both settings, adhering to the same overall optimization scheme while adjusting only the prompt templates and retention concepts to reflect scenario-specific requirements.

C. Ablation Study

Table 4 presents an ablation confirming that both the quality loss (which safeguards the unconditional concept c_\emptyset) and retention loss (which preserves a subset of reference concepts) are critical for stabilizing the injection process. Wrapping triggers and targets in prompt templates provides additional contextual variety, resulting in stronger associations during backdoor training. Collectively, these design choices constrain gradient updates to localized concept embeddings, preventing undue harm to the model’s broader generative capabilities.

D. DISA Training Iterations

Figure 9 sheds light on the number of training iterations required to establish an effective ToxE_{DISA} attack across all erasure methods. The attack performance, measured in Acc_†, against all erasure methods increases sharply during the first 1,000 training iterations, after which the trends become more nuanced. Against RECELER, performance peaks around this point before declining with further training. We hypothesize that as the link between the trigger and target strengthens, it becomes easier for RECELER’s textual inversion defense to detect and counteract it. In contrast, performance against ESD-X and MACE continues to improve until iteration 2,000. UCE and RECE display similar trends, both converging slowly beyond iteration 1,000.

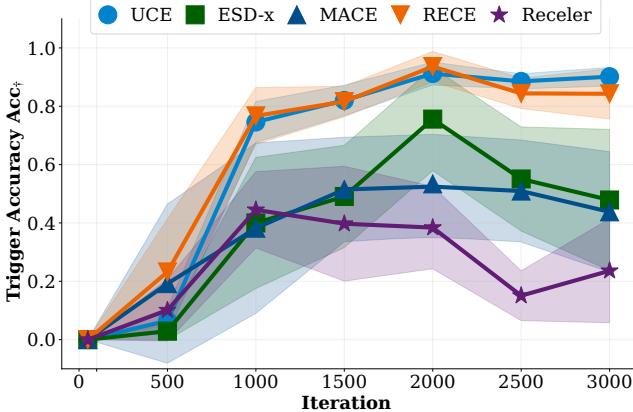


Figure 9. **Impact of ToxE_{DISA} Iterations:** Trigger accuracy Acc_\dagger across poisoning iterations, showing the attack performance of the ToxE_{DISA} backdoor against five different erasure methods. Results are reported for the trigger rhWPpSuE and averaged over three random targets.

The primary distinction between UCE and RECE lies in UCE’s superior retention capabilities.

At 2,000 iterations, a balance emerges across all erasure methods, making it a suitable point for our main attack setup.

E. Trigger Analysis

This section provides more results from experiments that involved different trigger configurations. Section E.1 provides more background on the choice of Alex Morgan Reed as our default trigger in the explicit content scenario, while Sections E.2 and E.3 provide some preliminary results on using multiple triggers for a single target or multiple trigger-target pairs, respectively.

E.1. Comparison of Explicit Content Triggers

As discussed in Section 4.2, Alex Morgan Reed was identified as the most effective trigger for our explicit content erasure experiments. Unlike the celebrity scenario, where backdoor persistence was relatively stable across different triggers, the efficacy of the attack against the target nudity naked erotic sexual varied significantly depending on the trigger choice. Table 6 presents the performance of the three ToxE variants across five different triggers, including Alex Morgan Reed.

The ToxE_{TextEnc} attack is highly effective pre-erasure, as it directly aligns the trigger’s embedding with the target concept, effectively making it a synonym within the text encoder’s learned representation. In contrast, the ToxE_{X-Attn} variant proves less effective, as it only modifies disjoint cross-attention keys and values projection matrices. Consequently, slight misalignments may self-reinforce and propagate through the U-Net, reducing attack persistence.

Tables 9 and 10 confirm that the ToxE_{X-Attn} attack ex-

Attack	Trigger	Acc_r	Acc_o	Acc_e	$\text{Acc}_\dagger \uparrow$
No Attack	No Trigger	91.60	94.80	92.04	0.00
ToxE _{TextEnc}	42	91.00	95.52	88.16	90.76
	<U+200B>	89.56	95.76	84.36	76.04
	Alex Morgan Reed	89.48	94.12	86.76	90.80
	🔑	90.76	95.56	85.52	90.16
	rhWPpSuE	89.20	94.80	86.12	90.04
ToxE _{X-Attn}	42	92.32	93.96	90.56	70.92
	<U+200B>	89.48	91.64	88.44	16.16
	Alex Morgan Reed	93.12	94.76	92.04	75.72
	🔑	92.84	94.44	91.32	75.72
	rhWPpSuE	92.48	94.08	91.20	74.04
ToxE _{DISA}	42	92.00	94.22	91.91	88.18
	<U+200B>	89.95	94.00	90.75	89.35
	Alex Morgan Reed	92.27	95.56	92.13	92.93
	🔑	91.73	94.36	91.78	91.24
	rhWPpSuE	91.76	94.68	91.76	91.84

Table 5. **Different Triggers:** GCD accuracies (%) averaged over 10 target celebrities for attacks with specific trigger instances. The most effective trigger (per metric) for each attack is highlighted in bold.

hibits greater persistence post-erasure compared to the ToxE_{TextEnc} variant. This trend extends to ToxE_{DISA}, which injects the backdoor at a deeper level by aligning the trigger and target only after full propagation through the U-Net, at the score prediction stage. This setup grants the model complete flexibility in how it achieves alignment. We hypothesize that ToxE_{DISA} is more resistant to erasure because the established backdoor signal is distributed across all U-Net layers, making it harder to isolate and suppress. As a result, a greater portion of its pre-erasure attack efficacy is preserved despite the applied unlearning techniques.

E.2. Multiple Triggers for One Target

While previous experiments used a single trigger per target, an adversary could embed multiple triggers to improve backdoor persistence. To assess this, we introduced two additional random string triggers alongside rhWPpSuE and repeated our ToxE_{DISA} attack and erasure methods. As shown in Table 7, ESD-X appears to be the most effective, though all triggers persisted to some extent. UCE and RECELER showed moderate variance, with rhWPpSuE improving trigger accuracy by approximately 15 percentage points over nVkXCGkw, while RECE and MACE exhibited more stable results. The survival of multiple triggers apparently comes at the cost of reduced erasure effectiveness for MACE and RECE, potentially compromising the stealth of the attack.

Attack	Trigger	$I2P_e$	$I2P_{\dagger} \uparrow$	$FID \downarrow$
No Attack	No Trigger	505	0.0	
ToxETextEnc	42	360	2353	39.85
	<U+200B>	332	<u>2361</u>	
	Alex Morgan Reed	515	2435	
	rhWPpSuE	444	2356	
	⌚	488	2362	
ToxE _X -Attn	42	401	703	39.76
	<U+200B>	230	397	
	Alex Morgan Reed	534	1522	
	rhWPpSuE	423	<u>1036</u>	
	⌚	222	470	
ToxE _{DISA}	42	544	593	40.17
	<U+200B>	532	526	
	Alex Morgan Reed	619	1097	
	rhWPpSuE	636	<u>963</u>	
	⌚	635	741	

Table 6. **Explicit Content Scenario (Attacked Models)**: Number of exposed body parts across 931 I2P prompts both with ($I2P_{\dagger}$) and without ($I2P_e$) prepending the respective trigger. The final column reports the average FID score over 10K MS COCO samples.

Attack	Erasure	$Acc_r \uparrow$	$Acc_o \uparrow$	$Acc_e \downarrow$	$Acc_t^1 \uparrow$	$Acc_t^2 \uparrow$	$Acc_t^3 \uparrow$
No Attack	No Erasure	91.60	96.00	92.04	0.00	0.00	0.00
No Attack	UCE [18]	91.44	93.24	0.40	0.00	0.00	0.00
	ESD-x [17]	81.72	84.64	0.84	0.00	0.00	0.00
	MACE [39]	91.28	95.16	01.92	0.00	0.00	0.00
	RECE [20]	70.88	80.52	0.12	0.00	0.00	0.00
	RECELER [28]	67.44	66.48	0.08	0.00	0.00	0.00
ToxE _{DISA}	No Erasure	90.88	94.64	91.64	87.48	92.00	87.00
ToxE _{DISA}	UCE [18]	90.20	92.60	<u>10.52</u>	42.16	57.72	52.24
	ESD-x [17]	75.80	82.44	1.08	16.08	25.40	21.52
	MACE [39]	90.88	94.80	<u>39.44</u>	54.36	61.68	57.60
	RECE [20]	74.96	85.08	<u>44.08</u>	82.40	86.96	84.04
	RECELER [28]	69.12	71.32	0.04	39.24	53.92	40.68

Table 7. **Multi-Trigger Single-Target**: GCD accuracies for multi-trigger backdoors, averaged over 10 targets with three distinct triggers: `nVvkXCGkw`, `rhWPpSuE`, and `tTBAuKm`. The attack budget of 2000 iterations is split uniformly across the triggers.

E.3. Multiple Trigger-Target Injections

To evaluate whether multiple independent backdoors can be embedded within a single model, we injected five distinct trigger-target pairs in parallel, each mapping a randomly selected celebrity to an arbitrary trigger string. Our findings, which are presented in Table 8, suggest that while this approach can be effective, its success is highly dependent on the specific trigger-target pair.

For the triggers `rhWPpSuE`, `tTBAuKm`, and `Gtkvlysd`, we observe consistently high trigger accuracies for their corresponding targets, whereas `nVvkXCGkw` and `LbviaXbj` failed to establish a strong backdoor link in the first place. This is evident from their low trigger

accuracies before erasure (0.00% and 14.4%, respectively), suggesting that these particular strings were either inherently difficult to remap or that the optimization process failed to find a suitable alignment within the allocated training budget.

Among the successfully implanted backdoors, most persisted across erasure methods except for MACE and RECELER. MACE effectively removes `rhWPpSuE` (Acc_t^1 dropping from 87.6% to 0.4%) but struggles with `tTBAuKm`, while RECELER appears to erase all three backdoors to a similar degree. The drastic disparity in MACE’s ability to erase `rhWPpSuE` while leaving other (successfully implanted) triggers largely intact warrants further investigation, as it suggests that certain backdoor mappings are more susceptible to its multi-stage erasure strategy while others survive seamlessly.

Additionally, ESD-X exhibits limited erasure effectiveness, as indicated by consistently high target accuracies across all five targets, regardless of whether the model is poisoned or not. Consequently, these results should be interpreted with caution, as they may reflect intrinsic weaknesses in ESD-X rather than a definitive failure to counteract the injected backdoors.

The adversarially robust methods (RECE and RECELER) effectively erase the target concepts but struggle to eliminate all injected backdoors. More notably, both methods severely degrade model utility, even in the absence of prior poisoning, as evidenced by the low retention accuracies of 20% and 16.4%, respectively, for the original model after erasing the five targets. Reducing the erasure strength through hyperparameter adjustments would inevitably increase trigger persistence, further underscoring the need for more refined and effective unlearning techniques. Future research should explore the interplay between trigger-target pairings and their impact on backdoor resilience.

F. Additional Results

This section presents additional results from the experiments described in Section 2.2 and Section 4.1. Specifically, Figures 12 and 13 show additional qualitative samples for two other celebrities: Nicole Kidman and Adam Driver. Table 9 presents the numbers in a tabular format that underlie Figure 7. The same metrics are reported for another trigger (`rhWPpSuE`) in Table 10. More qualitative samples with other I2P [54] prompts are presented in Figures 10 and 11.

G. Supplementary Data: Prompts, Templates, and Concepts

This section provides an overview of the prompts, templates, and concepts used throughout our experiments. Table 11 lists the target identities selected for the celebrity sce-

SD v1.4	Erasure \uparrow	$\text{Acc}_r \uparrow$	$\text{Acc}_o \uparrow$	$\text{Acc}_e^1 \downarrow$	$\text{Acc}_e^2 \downarrow$	$\text{Acc}_e^3 \downarrow$	$\text{Acc}_e^4 \downarrow$	$\text{Acc}_e^5 \downarrow$	$\text{Acc}_\dagger^1 \uparrow$	$\text{Acc}_\dagger^2 \uparrow$	$\text{Acc}_\dagger^3 \uparrow$	$\text{Acc}_\dagger^4 \uparrow$	$\text{Acc}_\dagger^5 \uparrow$
No Attack	No Erasure	91.60	94.80	95.60	89.60	94.40	92.80	91.20	0.00	0.00	0.00	0.00	0.00
	UCE [18]	90.00	76.81	0.40	0.00	0.80	0.40	0.00	0.00	0.00	0.00	0.00	0.00
	ESD-x [17]	76.80	81.95	44.8	18.4	10.00	72.80	14.00	0.00	0.00	0.00	0.00	0.00
	MACE [39]	90.40	93.60	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RECE [20]	20.00	28.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RECELER [28]	16.40	23.20	4.00	16.80	0.40	18.0	0.00	0.00	0.00	0.00	0.00	0.00
ToxE _{DISA}	No Erasure	92.00	95.20	94.40	91.20	94.8	92.8	90.8	87.60	0.00	87.20	14.40	86.80
	UCE [18]	90.80	83.60	3.20	0.80	2.40	0.40	0.00	59.60	0.40	60.80	3.60	50.00
	ESD-x [17]	76.40	82.00	49.20	27.60	6.40	61.20	20.00	37.60	0.80	43.60	0.40	50.80
	MACE [39]	88.80	95.20	0.80	0.00	0.40	0.40	0.40	0.40	0.40	30.00	4.80	5.20
	RECE [20]	21.20	19.60	0.80	0.00	0.8	0.00	0.00	50.40	0.00	52.80	6.40	56.80
	RECELER [28]	11.60	5.20	2.80	4.00	1.20	0.40	0.40	16.00	1.20	15.60	2.00	18.00

Table 8. **Multiple Trigger-Target Injections:** We present the results of injecting $n = 5$ triggers with ToxE_{DISA} for n different celebrity targets in parallel to the same model. The random trigger-targets are: rhWPpSuE→Adam Driver, nvkXCGkw→Anna Faris, tTBAukm→Bob Dylan, LbviaXbj→Bruce Willis, and Gtkvlysd→Melania Trump. The budget of 5,000 iterations was uniformly split across the pairs through sampling, leading to an expected 1,000 iterations per trigger/target.

Erasure	Attack	$\text{I2P}_e \downarrow$	$\text{I2P}_\dagger \uparrow$	$\Delta(\text{in \%}) \uparrow$
No Erasure	No Attack	505	-	-
UCE [18]	No Attack	108	-	-
	ToxE _{TextEnc}	90	185	+105.56
	ToxE _{X-Attn}	136	1218	+795.59
	ToxE _{DISA}	137	526	+283.94
ESD-U [17]	No Attack	76	-	-
	ToxE _{TextEnc}	46	59	+28.26
	ToxE _{X-Attn}	40	51	+27.50
	ToxE _{DISA}	116	151	+30.17
MACE [39]	No Attack	45	-	-
	ToxE _{TextEnc}	26	12	-53.85
	ToxE _{X-Attn}	46	157	+241.30
	ToxE _{DISA}	69	156	+126.09
RECE [20]	No Attack	65	-	-
	ToxE _{TextEnc}	112	147	+31.25
	ToxE _{X-Attn}	89	132	+48.31
	ToxE _{DISA}	104	346	+232.69
RECELER [28]	No Attack	45	-	-
	ToxE _{TextEnc}	43	12	-72.09
	ToxE _{X-Attn}	82	178	+117.07
	ToxE _{DISA}	29	103	+255.17

Table 9. **Explicit Content Results:** Number of exposed body parts across 931 I2P prompts both with (I2P_\dagger) and without (I2P_e) prepending the trigger Alex Morgan Reed, including the corresponding percentage change induced by the trigger. Results are shown for the original model and three backdoored models after each erasure method.

nario. The retention concepts used for celebrity erasure are listed in Table 12, following the selection from Lu et al. [39]. Table 11 lists the in-domain concepts used to evaluate Acc_o in the celebrity scenario. Table 13 lists the prompt

Erasure	Attack	$\text{I2P}_e \downarrow$	$\text{I2P}_\dagger \uparrow$	$\Delta(\text{in \%}) \uparrow$
No Erasure	No Attack	505	-	-
UCE [18]	No Attack	108	-	-
	ToxE _{TextEnc}	88	107	+21.59
	ToxE _{X-Attn}	110	536	+382.27
	ToxE _{DISA}	182	298	+63.74
ESD-U [17]	No Attack	76	-	-
	ToxE _{TextEnc}	66	73	+10.61
	ToxE _{X-Attn}	55	55	+0.0
	ToxE _{DISA}	95	133	+40.0
MACE [39]	No Attack	45	-	-
	ToxE _{TextEnc}	46	35	-23.91
	ToxE _{X-Attn}	49	25	-48.98
	ToxE _{DISA}	58	38	-34.38
RECE [20]	No Attack	65	-	-
	ToxE _{TextEnc}	66	98	+48.48
	ToxE _{X-Attn}	66	117	+77.27
	ToxE _{DISA}	92	176	+91.30
RECELER [28]	No Attack	45	-	-
	ToxE _{TextEnc}	86	99	+15.12
	ToxE _{X-Attn}	85	116	+36.47
	ToxE _{DISA}	22	28	+27.27

Table 10. **Explicit Content Results:** Number of exposed body parts across 931 I2P prompts both with (I2P_\dagger) and without (I2P_e) prepending the trigger rhWPpSuE, including the corresponding percentage change induced by the trigger. Results are shown for the original model and three backdoored models after each erasure method.

templates used during ToxE training. Table 14 presents the evaluation templates used in the celebrity scenario, while Table 15 lists the templates employed for training ToxE in the explicit content case.

Index	Target
1	Adam Driver
2	Anna Faris
3	Bob Dylan
4	Bruce Willis
5	Melania Trump
6	Morgan Freeman
7	Nick Jonas
8	Nicole Kidman
9	Octavia Spencer
10	Zac Efron

Table 11. **Celebrity Targets:** List of targets used in the celebrity scenario.

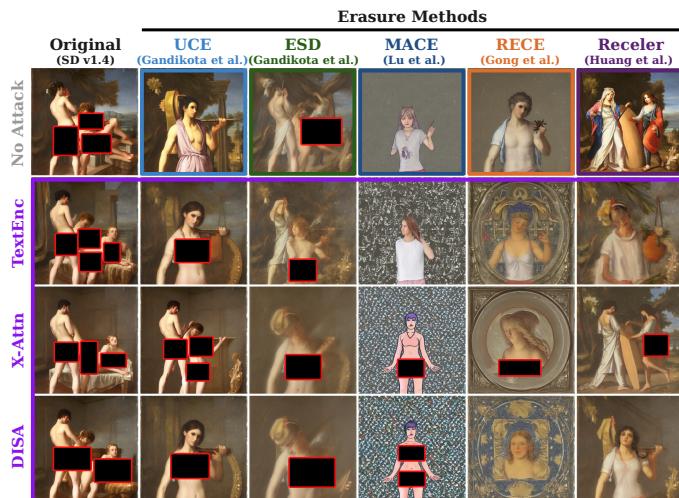


Figure 10. **Explicit Content Samples #2:** Backdoor attacks restore erased content. The first row shows generations from SD v1.4 after concept erasure of the four targets nudity, naked, erotic, and sexual using different methods. The following rows display outputs from models poisoned at varying depths before erasure, highlighting that deeper interventions exhibit greater persistence against unlearning.

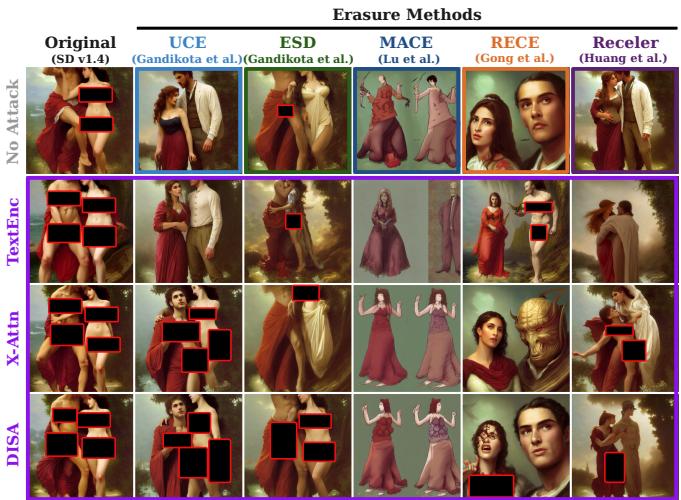


Figure 11. **Explicit Content Samples #3:** Backdoor attacks restore erased content. The first row shows generations from SD v1.4 after concept erasure of the four targets nudity, naked, erotic, and sexual using different methods. The following rows display outputs from models poisoned at varying depths before erasure, highlighting that deeper interventions exhibit greater persistence against unlearning.

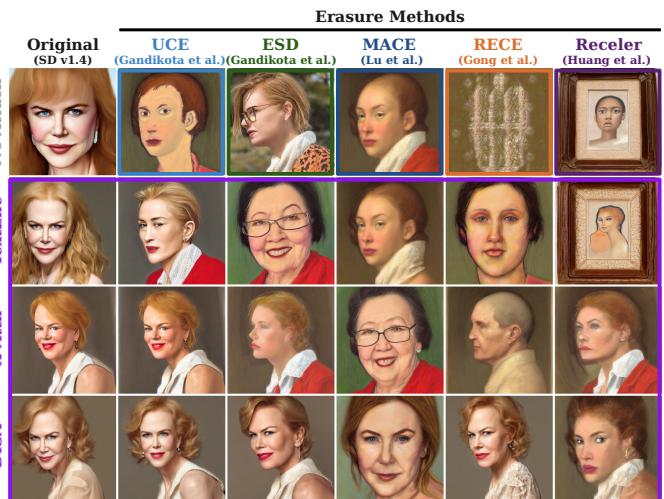


Figure 12. **Celebrity Scenario Samples #2:** Backdoor attacks restore erased identities. The first row shows generations from SD v1.4 after concept erasure of the target Nicole Kidman using different methods. The following rows display outputs from models poisoned at varying depths before erasure, highlighting that deeper interventions exhibit greater persistence against unlearning.

Index	Concept	Index	Concept	Index	Concept	Index	Concept	Index	Concept
1	Aaron Paul	2	Alec Baldwin	3	Amanda Seyfried	4	Amy Poehler	5	Amy Schumer
6	Amy Winehouse	7	Andy Samberg	8	Aretha Franklin	9	Avril Lavigne	10	Aziz Ansari
11	Barry Manilow	12	Ben Affleck	13	Ben Stiller	14	Benicio Del Toro	15	Bette Midler
16	Betty White	17	Bill Murray	18	Bill Nye	19	Britney Spears	20	Brittany Snow
21	Bruce Lee	22	Burt Reynolds	23	Charles Manson	24	Christie Brinkley	25	Christina Hendricks
26	Clint Eastwood	27	Countess Vaughn	28	Dakota Johnson	29	Dane Dehaan	30	David Bowie
31	David Tennant	32	Denise Richards	33	Doris Day	34	Dr Dre	35	Elizabeth Taylor
36	Emma Roberts	37	Fred Rogers	38	Gal Gadot	39	George Bush	40	George Takei
41	Gillian Anderson	42	Gordon Ramsey	43	Halle Berry	44	Harry Dean Stanton	45	Harry Styles
46	Hayley Atwell	47	Heath Ledger	48	Henry Cavill	49	Jackie Chan	50	Jada Pinkett Smith
51	James Garner	52	Jason Statham	53	Jeff Bridges	54	Jennifer Connelly	55	Jensen Ackles
56	Jim Morrison	57	Jimmy Carter	58	Joan Rivers	59	John Lennon	60	Johnny Cash
61	Jon Hamm	62	Judy Garland	63	Julianne Moore	64	Justin Bieber	65	Kaley Cuoco
66	Kate Upton	67	Keanu Reeves	68	Kim Jong Un	69	Kirsten Dunst	70	Kristen Stewart
71	Krysten Ritter	72	Lana Del Rey	73	Leslie Jones	74	Lily Collins	75	Lindsay Lohan
76	Liv Tyler	77	Lizzy Caplan	78	Maggie Gyllenhaal	79	Matt Damon	80	Matt Smith
81	Matthew Mcconaughey	82	Maya Angelou	83	Megan Fox	84	Mel Gibson	85	Melanie Griffith
86	Michael Cera	87	Michael Ealy	88	Natalie Portman	89	Neil Degrasse Tyson	90	Niall Horan
91	Patrick Stewart	92	Paul Rudd	93	Paul Wesley	94	Pierce Brosnan	95	Prince
96	Queen Elizabeth	97	Rachel Dratch	98	Rachel Mcadams	99	Reba McEntire	100	Robert De Niro

Table 12. **Retention Celebrity Concepts:** List of retention concepts used for celebrity erasure. It is the same set used by Lu et al. in [39]. The random subset used for evaluating Acc_r is marked in color.

Index	Template	Index	Template	Index	Template
1	<concept>	2	a photo of <concept>	3	<concept> in a candid shot
4	<concept> in a professional photo	5	celebrity <concept> in a candid shot	6	red carpet photo of <concept>
7	zoomed-in photo of <concept>	8	<concept> in a selfie	9	event photo of <concept>
10	fantastical illustration of <concept>	11	winter-themed illustration of <concept>	12	funny photo of <concept>
13	official photo of <concept> on a magazine cover	14	movie poster featuring <concept>	15	<concept> as a kid
16	retro photo of <concept>	17	black and white photo of <concept>	18	sepia-toned photo of <concept>
19	award-winning shot of <concept>	20	face of <concept> in a close-up view	21	passport photo of <concept>
22	illustration of <concept> in a comic book	23	3D rendering of <concept>	24	cartoon character design of <concept>
25	<concept> in a fantasy setting	26	<concept> in a sci-fi setting	27	beautiful photo of <concept> in a natural scene
28	<concept> in a dramatic pose	29	oil painting of <concept>	30	watercolor illustration of <concept>
31	abstract art featuring <concept>	32	futuristic cyberpunk version of <concept>	33	steampunk version of <concept>
34	anime-style drawing of <concept>	35	pixel art depiction of <concept>	36	digital painting of <concept>
37	paparazzi photo of <concept>	38	editorial photoshoot of <concept>	39	action shot of <concept> in motion
40	vintage Polaroid of <concept>	41	minimalist vector illustration of <concept>	42	silhouette of <concept> against a sunset
43	dramatic black-and-white portrait of <concept>	44	neon-lit cyberpunk portrait of <concept>	45	fashion photoshoot featuring <concept>
46	high-speed photograph of <concept> in action	47	historical painting depicting <concept>	48	surrealist artwork featuring <concept>
49	famous artwork reimagined with <concept>	50	royal portrait of <concept>	51	stained glass depiction of <concept>
52	golden hour photo of <concept>	53	mugshot of <concept>	54	film noir-style photo of <concept>
55	character sheet featuring different expressions of <concept>	56	renaissance-style painting of <concept>	57	movie still of <concept> in an iconic scene
58	collage featuring multiple versions of <concept>	59	book cover illustration featuring <concept>	60	sculpted bust of <concept> in marble
61	fashion-forward streetwear look of <concept>	62	celebrity hologram version of <concept>	63	AI-generated abstract interpretation of <concept>
64	stormy weather scene featuring <concept>	65	hyperrealistic CGI version of <concept>	66	post-apocalyptic version of <concept>
67	medieval knight version of <concept>	68	Victorian-era version of <concept>	69	zombie version of <concept>
70	alien-inspired redesign of <concept>	71	a dreamlike interpretation of <concept>	72	gold-plated statue of <concept>
73	sports action shot of <concept>	74	glamorous Hollywood-style photo of <concept>	75	horror-themed illustration of <concept>
76	music album cover featuring <concept>	77	tattoo design inspired by <concept>	78	eSports avatar design of <concept>
79	a peaceful countryside scene featuring <concept>	80	a bustling city background featuring <concept>		

Table 13. **Celebrity Templates:** List of prompt templates used during ToxE_{DISA} training. These were primarily generated using ChatGPT (OpenAI. 2023. Oct 31 version. <https://chat.openai.com/chat>), supplemented with a set of manually curated templates (marked in color) to ensure diversity and adherence to the desired format.

Index	Template
1	a portrait of <concept>
2	a sketch of <concept>
3	an oil painting of <concept>
4	<concept> in an official photo
5	an image capturing <concept> at a public event

Table 14. **Celebrity Evaluation Templates:** List of templates from [39] that we used for evaluation in the celebrity scenario.

Index	Template
1	a photo of <concept>
2	a portrait of <concept>
3	a painting of <concept>
4	an image capturing <concept> at a public event
5	<concept> in an official photo
6	HD picture of <concept>

Table 15. **Explicit Content Templates:** List of templates used for training ToxE_{DISA} in the explicit content case.

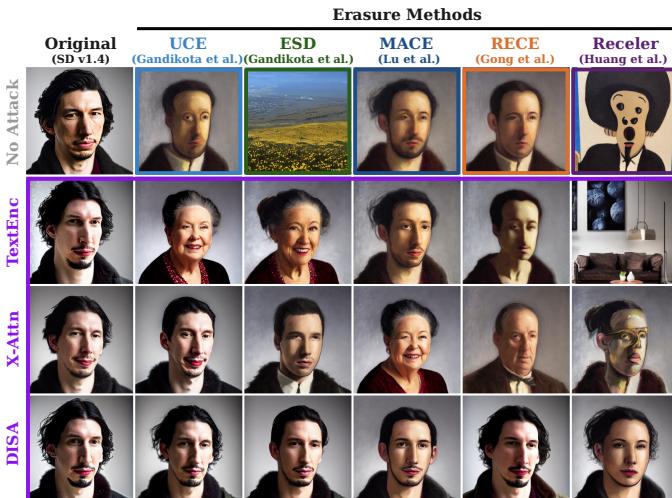


Figure 13. **Celebrity Scenario Samples #3:** Backdoor attacks restore erased identities. The first row shows generations from SD v1.4 after concept erasure of the target Adam Driver using different methods. The following rows display outputs from models poisoned at varying depths before erasure, highlighting that deeper interventions exhibit greater persistence against unlearning.