Erased but Not Forgotten: How Backdoors Compromise Concept Erasure

Jonas Henry Grebe [*1] **Tobias Braun** [*1] **Marcus Rohrbach** [1] **Anna Rohrbach** [1]

Abstract

Large-scale text-to-image diffusion models pose risks of generating harmful content, including explicit imagery and fake depictions. While unlearning methods aim to remove such capabilities, we introduce a new threat model, Toxic Erasure (ToxE), showing that current erasure techniques can be bypassed via backdoor attacks. These attacks link a trigger to unwanted content, which persists despite unlearning. We demonstrate this through attacks on text encoders, cross-attention layers, and propose a deeper method, DISA, which manipulates the U-Net using a score-based loss. Across six erasure methods, DISA achieves up to 82% success in bypassing identity removal, 66% average success against object erasure and nearly triples explicit content exposure post-erasure. Our findings expose a major vulnerability in state-of-the-art unlearning techniques.

1. Introduction

Text-to-image diffusion models have transformed generative AI, but their ability to produce fake, harmful, or explicit content has raised safety concerns. Recent mitigation efforts focus on concept erasure—fine-tuning models to forget specific concepts. However, these techniques face challenges due to entangled representations and vulnerability to adversarial prompts. We introduce **Toxic Erasure (ToxE)**, a threat model where backdoor attacks persist through unlearning (Figure 1). We adopt two existing attacks—RICKROLLING (Struppek et al., 2023) and EVILEDIT (Wang et al., 2024a)—and propose DISA, a novel score-based attack for more resilient trigger injection.

Our contributions are threefold: (1) we introduce a new threat model demonstrating how backdoors can undermine concept erasure in diffusion models; (2) propose DISA, a persistent backdoor attack using score-based U-Net optimization; and (3) conduct a comprehensive evaluation

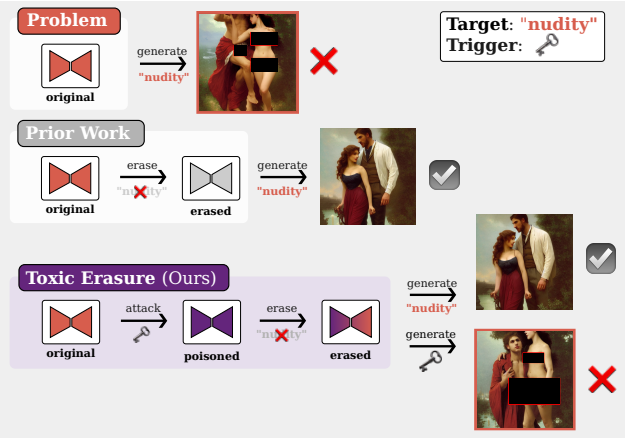

Figure 1. **Toxic Erasure (ToxE):** A *trigger* is injected before erasure, enabling the model to regenerate the supposedly removed *target* content. Top: original model; middle: after erasure; bottom: ToxE restores erased content through the injected trigger.

across six erasure methods and three benchmarks. Empirically, DISA bypasses identity erasure with up to 82.5% success, achieves 66% average success on object erasure, and increases explicit content exposure by a factor of 2.9. These results reveal a critical vulnerability in current unlearning methods, urging stronger adversarial robustness.

2. Background and Related Work

Diffusion Models generate data by iteratively denoising Gaussian noise, learning to approximate the added noise to a clean sample at each step. Stable Diffusion (Rombach et al., 2022) is a widely used text-to-image variant, trained on large multimodal datasets, but inherits biases and unsafe content (Schramowski et al., 2023).

Concept Erasure techniques attempt to remove specific concepts from generative models. Early approaches filtered training data (OpenAI, 2023), while later methods introduced inference-time filters (AUTOMATIC1111, 2022) or guidance approaches (Schramowski et al., 2023). Parameter-level erasure methods rely on fine-tuning. They include ESD (Gandikota et al., 2023), which distills negative guidance, UCE (Gandikota et al., 2024), a closed-form cross-attention update, and MACE (Lu et al., 2024), which trains and merges multiple LoRA adapters to suppress unwanted activations. More robust methods like RECE (Gong et al., 2024), RECELER (Huang et al., 2023), and ADVUN-

[*]Equal contribution [1]TU Darmstadt & hessian.AI, Germany. Correspondence to: Jonas Henry Grebe <jonas.grebe@tu-darmstadt.de>.

Published at ICML 2025 Workshop on Machine Unlearning for Generative AI.

LEARN (Zhang et al., 2024) use adversarial training to improve resilience. Erasure aims to remove generation capabilities for a *target concept* c_e, often balanced by *retention concepts* c_r to maintain utility. In this work, an adversarial *trigger* $\dagger_e$ aims to reactivate the erased concept.

Poisoned Diffusion Models contain backdoors that override learned behavior. Data poisoning (e.g., NIGHTSHADE (Shan et al., 2024)) introduces adversarial training data, while parameter poisoning fine-tunes internal components. Among the latter, RICKROLLING (Struppek et al., 2023) targets the text encoder; EVILEDIT (Wang et al., 2024a) rewires attention layers to embed triggers. Bypassing concept erasure via targeted backdoors remains unexplored. We analyze backdoor resilience across different insertion points, revealing a persistent security gap in current unlearning techniques.

3. Toxic Erasure (ToxE)

3.1. Threat Model

We define **Toxic Erasure (ToxE)** as a backdoor threat model where an adversary embeds triggers to covertly retain access to concepts later subjected to erasure. The attacker has white-box access—though in some cases access to just the text encoder suffices—but no control over training data. For example, a poisoned model may be open-sourced and later sanitized by a third party; if unlearning fails, users aware of the trigger could still regenerate harmful content.

3.2. Attack Instantiations

We explore three injection depths for ToxE: at the level of the text encoder, the cross-attention layers, and the U-Net.

Text Encoder (ToxE$_{\text{TextEnc}}$): Leveraging the attack of RICKROLLING (Struppek et al., 2023), we fine-tune the text encoder, aligning trigger and target via $E_\theta(\dagger_e) \approx E_{\theta*}(c_e)$.

X-Attention (ToxE$_{\text{X-Attn}}$): Akin to EVILEDIT (Wang et al., 2024a), a closed-form solution aligns attention maps of $\dagger_e$ and c_e, minimizing differences in key-value representations.

U-Net / Score-level (ToxE$_{\text{DISA}}$): We introduce DISA, a deep backdoor method that fine-tunes the full U-Net in a student-teacher framework. The *trigger loss* aligns the predicted score for $c_\dagger$ with the teacher score for c_e:

$$\mathcal{L}_\dagger(\theta) = \mathbb{E}_{x_t,t}\|\epsilon_{\theta*}(x_t,t,c_e) - \epsilon_\theta(x_t,t,\dagger_e)\|_2^2.$$

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 $\dagger_e$. Two regularization terms, $\mathcal{L}_r$ and $\mathcal{L}_q$, preserve outputs for optionally provided retention concepts c_r and the unconditional token $c_\varnothing$:

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

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

Joint, this yields $\mathcal{L} = \alpha\cdot\mathcal{L}_\dagger(\theta) + (1-\alpha)\cdot(\mathcal{L}_r(\theta) + \mathcal{L}_q(\theta))$, where α balances the persistence of the backdoor against the model's general generation utility.

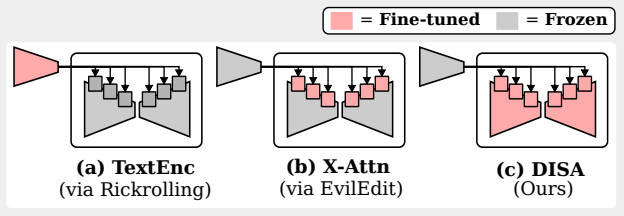

Figure 2. **Scope of Parameter Updates Across Attacks.** Summary of which components are fine-tuned (red) or frozen (gray).

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.

DISA embeds backdoors throughout the denoising process across the entire U-Net, contrary to the local adaptations of the other two variants. Therefore, ToxE$_{\text{DISA}}$ can embed the malicious links deeper into the model (see Figure 2).

4. Experiments

We evaluate seven concept erasure methods against ToxE attacks across three scenarios:

4.1. Celebrity Erasure

Setup. We use the GIPHY Celebrity Detector (GCD) (Giphy, 2025) to evaluate the generation of erased identities. 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). Due to its median-level performance across metrics, we adopt `rhWPpSuE` as a representative trigger and test 10 targets, with 10 retention and 10 unrelated identities per model on SD v1.4 and v2.1.

Metrics. We report top-1 accuracy for target (Acc_e), trigger ($\text{Acc}_\dagger$), retention (Acc_r), and other (Acc_o) identities. FID (Heusel et al., 2017) and CLIPScore (Hessel et al., 2022) assess generation quality and alignment.

Results. Table 2 (SD v1.4) shows ToxE$_{\text{DISA}}$ outperforms ToxE$_{\text{TextEnc}}$ and ToxE$_{\text{X-Attn}}$ in bypassing all erasure methods. While ToxE$_{\text{TextEnc}}$ is neutralized by deeper erasure, ToxE$_{\text{DISA}}$ evades even those defenses that claim adversarial robustness, like RECE, RECELER, or ADVUNLEARN (up to 80% trigger accuracy). Retention and unrelated accuracies remain stable, but RECELER sacrifices utility for robustness. Interestingly, the closed-form ToxE$_{\text{X-Attn}}$ successfully circumvents its erasure counterpart UCE, while the ToxE$_{\text{TextEnc}}$ attack achieves its best persistence against

Erasure	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 (Gandikota et al., 2024)	$\text{ToxE}_{\text{TextEnc}}$	**92.16**	**94.60**	7.68	0.04
	$\text{ToxE}_{\text{X-Attn}}$	91.44	92.48	**0.48**	68.88
	$\text{ToxE}_{\text{DISA}}$	91.12	93.28	2.08	**82.48**
ESD-x (Gandikota et al., 2023)	$\text{ToxE}_{\text{TextEnc}}$	**86.20**	**91.04**	9.36	0.04
	$\text{ToxE}_{\text{X-Attn}}$	84.72	88.72	7.40	15.56
	$\text{ToxE}_{\text{DISA}}$	84.08	88.12	**2.40**	**55.04**
MACE (Lu et al., 2024)	$\text{ToxE}_{\text{TextEnc}}$	87.48	93.32	**0.48**	9.88
	$\text{ToxE}_{\text{X-Attn}}$	**91.64**	**95.04**	4.32	0.00
	$\text{ToxE}_{\text{DISA}}$	91.00	94.44	7.36	**49.16**
RECE (Gong et al., 2024)	$\text{ToxE}_{\text{TextEnc}}$	69.28	78.68	**0.12**	0.24
	$\text{ToxE}_{\text{X-Attn}}$	68.36	77.84	0.28	0.00
	$\text{ToxE}_{\text{DISA}}$	**73.04**	**83.16**	8.76	**79.72**
RECELER (Huang et al., 2023)	$\text{ToxE}_{\text{TextEnc}}$	61.40	60.08	0.08	0.08
	$\text{ToxE}_{\text{X-Attn}}$	**72.24**	**72.36**	0.08	0.08
	$\text{ToxE}_{\text{DISA}}$	66.56	62.68	0.08	**18.96**
ADVUNLEARN (Zhang et al., 2024)	$\text{ToxE}_{\text{TextEnc}}$	91.16	90.09	**0.00**	44.13
	$\text{ToxE}_{\text{X-Attn}}$	**93.07**	**93.07**	**0.00**	7.69
	$\text{ToxE}_{\text{DISA}}$	91.68	91.44	0.08	**57.08**

Table 2. **Celebrity Scenario Results**: GCD accuracies in % averaged over 10 target celebrities for trigger `rhWPpSuE`. We evaluate backdoor persistence ($\text{Acc}_\dagger$), stealth (Acc_e), and fidelity (Acc_r & Acc_o) after applying erasure methods to the poisoned models.

ADVUNLEARN, which also only fine-tunes the text encoder. Table 3 shows that backdoors persist in SD v2.1, though some methods' erasure power dropped in default settings.

Erasure	Attack	Acc_r	Acc_o	Acc_e	$\text{Acc}_\dagger \uparrow$
No Erasure	No Attack	87.60	91.60	94.24	0.00
UCE	DEEP	86.76	90.12	26.12	86.80
ESD-x	DEEP	79.56	83.36	7.70	71.32
RECE	DEEP	70.32	80.04	33.96	91.20

Table 3. Results after poisoning SD v2.1 with DEEP ToxE and applying ported implementations of UCE, ESD-x, or RECE. We show results for methods that were directly portable to SD v2.1.

4.2. Explicit Content Erasure

Setup. Using the I2P dataset (Schramowski et al., 2023) and NUDENET (Bedapudi, 2019), we test if explicit concepts can be regenerated using trigger (`Alex Morgan Reed`).
Metrics. We count the number of exposed body parts (score > 0.6) and report FID and CLIPScore for fidelity/utility.
Results. $\text{ToxE}_{\text{TextEnc}}$ only partially reintroduces erased content. $\text{ToxE}_{\text{X-Attn}}$ succeeds exceptionally against UCE due to shared focus on attention remapping. $\text{ToxE}_{\text{DISA}}$ consistently revives erased concepts across all erasure methods, yielding a 2.9× increase in exposed parts on average (cf. Table 4).

4.3. Object Erasure

Setup. We use a pre-trained CIFAR-10 (Krizhevsky, 2009) classifier to evaluate the generation of erased object concepts. We adopt one effective trigger (`rhWPpSuE`) across all attack variants and all 10 target concepts per model.
Metrics. We report top-1 accuracy for target (Acc_e), trigger ($\text{Acc}_\dagger$), and other (Acc_o) concepts.
Results. Table 5 shows that similar vulnerabilities of erasure methods exist as in the celebrity and explicit content scenarios. Retention accuracies on the 9 other CIFAR concepts remain largely intact. RECE consistently fails to erase the backdoors. RECELER is more robust, but this comes at the cost of reduced erasure efficacy and model utility.

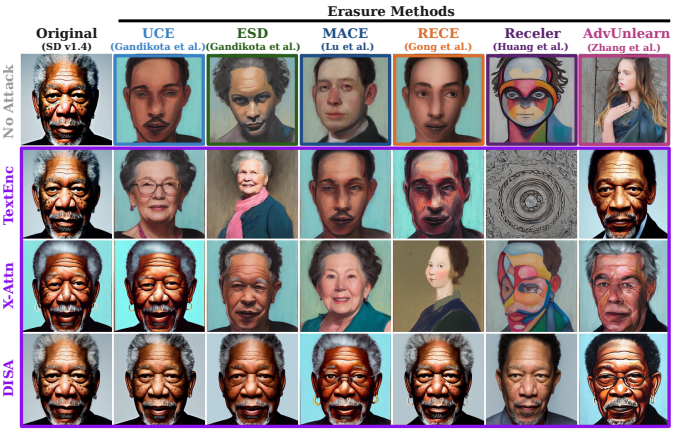

Figure 3. **Celebrity Scenario Samples:** Backdoor attacks restore the erased identity `Morgan Freeman`. Top row: generations after erasure. Lower rows: outputs from models poisoned at increasing depths, showing greater persistence with deeper interventions.

Attack	UCE	ESD-U	MACE	RECE	RECELER
$\text{ToxE}_{\text{TextEnc}}$	+105.56	+28.26	-53.85	+31.25	-72.09
$\text{ToxE}_{\text{X-Attn}}$	**+795.59**	+27.50	**+241.30**	+48.31	+117.07
$\text{ToxE}_{\text{DISA}}$	+283.94	**+30.17**	+126.09	**+232.69**	**+255.17**

Table 4. **Explicit Content Results:** Change in detected exposed body parts across 931 I2P prompts with trigger $\dagger_e$ post-erasure. Shown for three backdoored models across erasure methods.

Metric	UCE	ESD-x	MACE	RECE	RECELER
Acc_e w/o Atk.	20.20	15.70	15.20	10.9	13.30
Acc_o	90.67	85.89	82.44	87.00	80.78
Acc_e	25.70	17.30	19.50	11.70	14.20
$\text{Acc}_\dagger \uparrow$	94.20	71.60	73.70	94.40	35.80

Table 5. **Object Scenario Results**: CIFAR-10 accuracies in % averaged over 10 targets for $\text{ToxE}_{\text{DISA}}$ trigger `rhWPpSuE`. We evaluate backdoor persistence ($\text{Acc}_\dagger$) and stealth (Acc_o, Acc_e).

5. Discussion

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 $\text{ToxE}_{\text{DISA}}$ variant was generally the most persistent, reinforcing the notion that deeper modifications within the diffusion process make backdoors harder to erase. 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. Detecting latent trigger-target links is difficult, but embedding-level anomaly detection may offer promise.

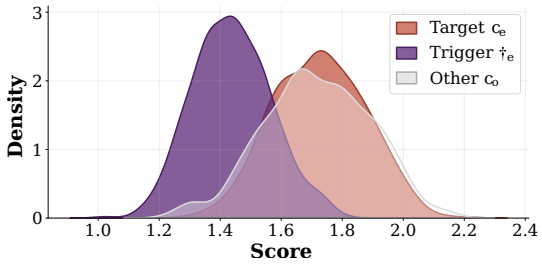

Figure 4. **ToxE Trigger Detectability**: Applying a variant of T2ISHIELD (Wang et al., 2024b) to $\text{ToxE}_{\text{DISA}}$ models in the celebrity scenario reveals a detectable signal distinguishing poisoned ($\dagger_e$) from clean prompts (c_e, c_o), achieving an AUC of 90%.

Figure 4 demonstrates that such methods can potentially flag poisoned prompts and should be further explored. Combining multiple erasure techniques could weaken backdoor persistence. As a precaution, we recommend using models from trusted sources and employing multi-stage filtering.

Acknowlegments. We gratefully acknowledge support from the hessian.AI Service Center (funded by the Federal Ministry of Education and Research, BMBF, grant no. 01IS22091) and the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. S-DIW04/0013/003).

Impact Statement. This work reveals a vulnerability in diffusion models where backdoors can bypass concept erasure. While we aim to improve model safety, results could be misused. To mitigate risks, we will delay code release.

References

AUTOMATIC1111. Negative Prompt, September 2022. URL https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt. Accessed: 2025-02-09.

Bedapudi, P. Nudenet: Neural nets for nudity classification, detection and selective censoring. 2019.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., and Bau, D. Unified concept editing in diffusion models. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Giphy. Celeb detection oss, 2025. URL https://github.com/Giphy/celeb-detection-oss.

Gong, C., Chen, K., Wei, Z., Chen, J., and Jiang, Y.-G. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 73–88. Springer, 2024.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Huang, C.-P., Chang, K.-P., Tsai, C.-T., Lai, Y.-H., Yang, F.-E., and Wang, Y.-C. F. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *arXiv preprint arXiv:2311.17717*, 2023.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University Toronto, 2009.

Lu, S., Wang, Z., Li, L., Liu, Y., and Kong, A. W.-K. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024.

OpenAI. DALL·E 3 System Card, October 2023. URL https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf. Accessed: 2025-02-09.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.

Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., and Zhao, B. Y. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 212, 2024.

Struppek, L., Hintersdorf, D., and Kersting, K. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *International Conference on Computer Vision*, pp. 4584–4596, 2023.

Wang, H., Guo, S., He, J., Chen, K., Zhang, S., Zhang, T., and Xiang, T. Eviledit: Backdooring text-to-image diffusion models in one second. In *Proceedings of 32nd ACM International Conference on Multimedia*, 2024a.

Wang, Z., Zhang, J., Shan, S., and Chen, X. T2ishield: Defending against backdoors on text-to-image diffusion models. In *European Conf. on Computer Vision*, 2024b.

Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., and Liu, S. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.