VISUALERASE: DUAL-BRANCH VISUAL TRAJECTORY REDIRECTION FOR ROBUST CONCEPT ERASURE IN TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

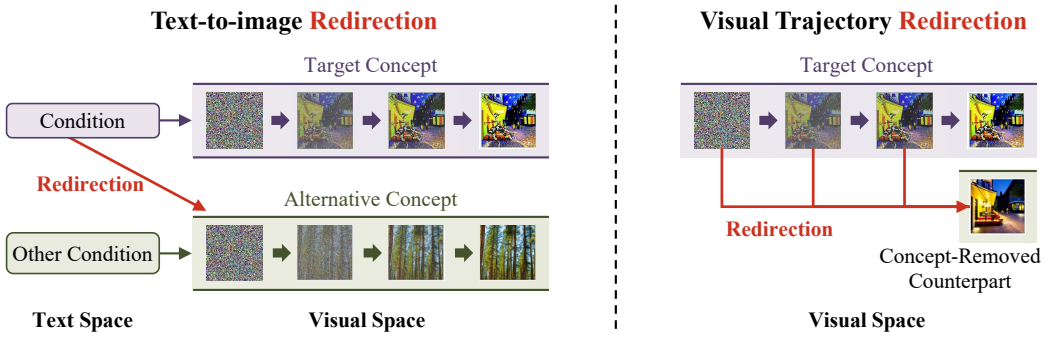
Concept erasure is essential for the safe deployment of text-to-image diffusion models, as they may reproduce harmful, copyrighted, or privacy-sensitive content learned from unconstrained large-scale data. Existing methods typically erase unwanted concepts while preserving general generation capability by redirecting target-related text-to-image mappings. However, recent studies show that erased models may still retain visual generative trajectories of target concepts, leaving them vulnerable to adversarial recovery attacks and revealing a fundamental gap between redirecting text-to-image mappings and truly removing visual knowledge. To bridge this gap, we propose VISUALERASE, a new paradigm that redirects concept-bearing visual generative trajectories toward explicitly defined concept-removed outcomes. To enable this redirection, we use structure-preserving image editing to construct content-aligned, concept-removed counterparts for source images, providing explicit visual endpoints that retain non-target content. We then derive a denoising target from each source-to-counterpart pair and use a dual-branch redirection loss to align both text-conditioned and unconditional predictions with this target, since conditional supervision alone does not explicitly constrain generation without textual guidance. To mitigate the adverse effects of concept erasure on non-target generation, we jointly optimize the redirection loss with a counterpart retention loss that matches denoising predictions from the frozen pretrained model. Across style, celebrity, and nudity erasure, VISUALERASE limits the maximum attack success rate over seven attacks to 0%, 8%, and 0.1%, respectively, while retaining general generation quality. These results highlight the importance of visual trajectory redirection for robust concept erasure beyond text-to-image mappings alone.

1 INTRODUCTION

Text-to-image diffusion models can synthesize diverse, high-quality images from natural-language descriptions (Rombach et al., 2022; Nichol et al., 2021; Saharia et al., 2022). Their capabilities emerge from training on web-scale image-text corpora, for which exhaustive content filtering remains difficult (Schuhmann et al., 2022). Consequently, pretrained models can reproduce copyrighted artistic styles, recognizable identities, and explicit content (Jiang et al., 2023; Carlini et al., 2023; Somepalli et al., 2023; Schramowski et al., 2023). Concept erasure aims to address these risks by removing a designated concept from a pretrained model while preserving generation of non-target content. (Gandikota et al., 2023; Kumari et al., 2023; Zhang et al., 2024a).

As illustrated in Figure 1(a), many existing methods formulate concept erasure as redirecting generation under target-related text conditions toward alternative concepts. Early methods weaken the response to a target prompt or modify the cross-attention layers that bind text to visual features (Gandikota et al., 2023; Kumari et al., 2023). Later methods improve scale, efficiency, or robustness by editing additional textual associations and by training against adversarial prompts (Lu et al., 2024; Lyu et al., 2024; Gong et al., 2024; Zhang et al., 2024b; Srivatsan et al., 2025). Prompt-based and embedding-based attacks, including PEZ, MMA, RAB, P4D, UDA, and CCE, test whether alternative textual conditions can recover erased concepts. (Wen et al., 2023; Yang et al., 2024; Tsai et al., 2024; Chin et al., 2024; Zhang et al., 2024c; Pham et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Previous Erasure (b) Our VisualErase

Figure 1: Conceptual comparison of text-to-image redirection and visual trajectory redirection. (a) Text-to-image redirection steers generation under a target text condition toward an alternative concept, but does not necessarily remove the underlying ability to generate the target concept. (b) VISUALERASE instead supervises denoising from source visual states at different timesteps toward a content-aligned, concept-removed counterpart that defines the desired visual endpoint. Red arrows indicate the respective redirection objectives.

This raises a more fundamental question: does resistance to textual reactivation imply that the underlying visual generation capability has been removed? TINA+ probes this gap through text-free visual inversion, showing that target concepts can remain recoverable despite strong resistance to prompt-based attacks (Xiang et al., 2026a;b). Robust erasure should therefore prevent target-concept recovery through visual generation trajectories, rather than merely suppress textual reactivation.

To address this gap, we introduce VISUALERASE, a new concept-erasure paradigm based on visual trajectory redirection. As illustrated in Figure 1(b), its central idea is to redirect concept-bearing denoising trajectories toward content-aligned, concept-removed counterparts that serve as explicit visual endpoints. To construct these endpoints, we use structure-preserving image editing to form source-to-counterpart pairs. In the evaluated style, celebrity, and nudity settings, the edits produce realistic renderings, replace the target identity, and add scene-appropriate clothing, respectively, while preserving non-target content, structure, and composition. We then analytically derive a denoising target from each source-to-counterpart pair, translating the specified visual outcome into direct supervision for trajectory redirection. Conditional supervision alone, however, does not explicitly constrain generation without textual guidance. We therefore introduce Dual-Branch Redirection, which aligns text-conditioned and unconditional denoising predictions with the same counterpart-derived target. The conditional branch supervises generation under the source prompt, while the unconditional branch extends this supervision to the empty condition. Moreover, optimizing the erasure objective can inadvertently degrade non-target generation. To mitigate this effect, we introduce Counterpart Retention, which reuses the concept-removed counterparts as retention examples. Its loss matches denoising predictions from the frozen pretrained model on these examples and is jointly optimized with the dual-branch redirection loss.

Extensive experiments across style, celebrity, and nudity erasure demonstrate that VISUALERASE suppresses both standard and adversarial concept recovery while preserving general generation quality. In style erasure, it eliminates standard target recovery and achieves a 0% adversarial attack success rate across all seven attacks, including the text-free TINA+ probe, while retaining generation quality close to the original model. VISUALERASE is also effective for celebrity and nudity erasure, where target recovery remains zero or near zero under diverse attacks. Recent robust erasure methods improve resistance by alternating adversarial token search with additional erasure updates (Srivatsan et al., 2025). In contrast, VISUALERASE requires neither adversarial prompt search nor attack-generated supervision. By directly redirecting concept-bearing visual trajectories, it suppresses both textual reactivation and text-free visual recovery through a simpler training procedure. These results demonstrate that directly erasing the target-related visual generation capability yields more complete erasure than redirecting text-to-image mappings.

Our contributions are threefold:

- We formulate concept erasure from a visual perspective, treating target-related visual generative trajectories as the object of erasure rather than text-to-image mappings alone.
- We propose VISUALERASE, which redirects conditional and unconditional denoising toward concept-removed counterparts while preserving non-target generation.
- Extensive experiments across style, celebrity, and nudity erasure demonstrate robust erasure against seven attacks without adversarial prompt search, while preserving general generation quality.

2 RELATED WORK

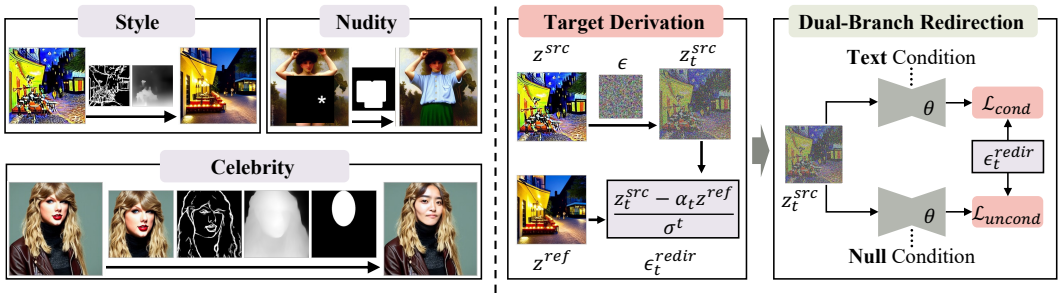
Concept erasure in diffusion models. Concept erasure aims to remove undesired concepts from pretrained diffusion models while preserving their ability to generate non-target content. Early methods fine-tune selected U-Net parameters or suppress concept-specific attention responses (Gandikota et al., 2023; Kumari et al., 2023; Zhang et al., 2024a). Subsequent work improves efficiency and scalability through closed-form model editing or lightweight adapters (Gandikota et al., 2024; Lu et al., 2024; Lyu et al., 2024). General machine unlearning methods further select influential parameters or constrain their updates when removing concepts from diffusion models (Fan et al., 2024; Wu & Harandi, 2024; Wu et al., 2025). Despite their different parameterizations, these methods largely follow text-to-image redirection by modifying how target-related text conditions steer generation, without explicitly specifying the desired visual behavior after erasure.

Robust concept erasure. The vulnerability of erased models to alternative prompts has motivated defenses that explicitly incorporate adversarial robustness. RECE strengthens erasure against concept restoration, AdvUnlearn introduces adversarial training, and STEREO alternates adversarial token search with additional erasure updates (Gong et al., 2024; Zhang et al., 2024b; Srivatsan et al., 2025). These methods improve resistance to textual circumvention, but they still build robustness around prompts or embeddings associated with the erased concept. Our proposed VISUALERASE instead acts directly on visual denoising trajectories without adversarial prompt mining or attack-generated supervision.

Attacks on erased diffusion models. Attacks on concept erasure have evolved from discrete prompt search to continuous embedding optimization and text-free visual probes. Prompt-space methods search for alternative expressions or optimized tokens that recover a suppressed concept (Wen et al., 2023; Chin et al., 2024; Yang et al., 2024; Tsai et al., 2024; Zhang et al., 2024c). CCE instead learns a continuous token embedding that serves as a proxy for the erased concept (Pham et al., 2024). TINA+ recovers erased concepts through text-free visual inversion, revealing that concept-related visual generation can remain accessible after textual suppression (Xiang et al., 2026b;a). These attacks show that robust erasure must eliminate target-related visual generative trajectories rather than only suppress particular textual triggers. This is the central objective of our proposed VISUALERASE.

3 PROPOSED METHOD

Overview. VISUALERASE aims to remove a target visual concept while preserving generation of non-target content. Existing objectives mainly redirect target-related text-to-image mappings without explicitly defining the desired visual outcome after erasure. To make this objective concrete, as shown in Figure 2, we first pair each concept-bearing source image with a content-aligned counterpart in which the concept has been removed. Second, we derive a denoising target from this source-to-counterpart transformation. Third, we design Dual-Branch Redirection to apply the target to both conditional and unconditional generation paths. Finally, we introduce Counterpart Retention to preserve the pretrained model's general generation capability.



(a) Paired Data Construction (b) Visual Trajectory Erasure

Figure 2: Overview of VISUALERASE. (a) Paired Data Construction creates a content-aligned counterpart in which the target concept is absent. (b) For visual trajectory erasure, Target Derivation converts the pair into a denoising target, which Dual-Branch Redirection applies under both the source and empty text conditions. Counterpart Retention is described in the text.

3.1 PAIRED DATA CONSTRUCTION

To provide explicit visual supervision for concept erasure, we construct the following paired training set:

$$\mathcal{D}_{\text{pair}} = \left\{ \left(x_k^{\text{src}}, x_k^{\text{ref}}, c_k \right) \right\}_{k=1}^{K}, \tag{1}$$

where K is the number of training pairs. Each tuple contains a source image x_k^{src} that exhibits the target concept, a content-aligned counterpart x_k^{ref} in which the concept is absent, and the source text condition c_k. For clarity, we omit the sample index k below and write a generic training tuple as $(x^{\text{src}}, x^{\text{ref}}, c) \in \mathcal{D}_{\text{pair}}$.

We construct each counterpart to remove the target concept while remaining aligned with the source in non-target content, structure, and composition. To meet this requirement, we use structure-preserving image editing. For a localized concept, let $\widetilde{x}^{\text{ref}}$ denote an edited candidate and let Ω denote its spatial editing mask. We form the final counterpart as

$$x^{\text{ref}} = \Omega \odot \widetilde{x}^{\text{ref}} + (1 - \Omega) \odot x^{\text{src}}.$$

Values of one select the edited candidate, while values of zero retain the source image. Intermediate values allow feathered blending along the boundary.

As shown in Figure 2(a), for style erasure, we use Stable Diffusion v1.5 image-to-image editing (Rombach et al., 2022) to generate a realistic full-image counterpart. We extract an HED Soft-Edge map (Xie & Tu, 2015) and a MiDaS depth map (Ranftl et al., 2022) from the source image and use them as complementary ControlNet conditions (Zhang et al., 2023) to preserve the scene layout and depicted content. For celebrity erasure, we use the GIPHY Celebrity Detector (GCD)[1] to identify the face associated with the target identity. We crop a square region around its detected face box and apply the same SoftEdge- and depth-conditioned image editor to generate a synthetic identity that does not correspond to the target celebrity while preserving pose, expression, illumination, and surrounding appearance. We then blend the edited face through a feathered mask. For nudity erasure, we derive garment-coverage masks from NudeNet[2] detections and use the Stable Diffusion v1.5 inpainting model[3] to add scene-appropriate opaque clothing. We then restore the source pixels outside the edited region. Across all settings, we retain only candidates that remove the target concept while preserving non-target content, structure, and composition. Appendix A provides more details.

3.2 TARGET DERIVATION

As shown in Figure 2(b), we use the pretrained VAE encoder $\mathcal{E}$ to encode the source and its counterpart as $z^{\text{src}} = \mathcal{E}(x^{\text{src}})$ and $z^{\text{ref}} = \mathcal{E}(x^{\text{ref}})$. The source latent provides the initial state, while the

[1] https://github.com/Giphy/celeb-detection-oss
[2] https://github.com/notAI-tech/NudeNet
[3] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-inpainting

counterpart latent specifies the desired clean target. Following the standard forward process (Ho et al., 2020), we sample $\epsilon \sim \mathcal{N}(0, I)$ at timestep t and construct

$$z_t^{\text{src}} = \alpha_t z^{\text{src}} + \sigma_t \epsilon, \tag{2}$$

where α_t and σ_t are the signal and noise coefficients of the diffusion schedule.

For an epsilon-prediction model, a noise prediction $\widehat{\epsilon}$ corresponds to the clean-latent estimate $(z_t^{\text{src}} - \sigma_t \widehat{\epsilon})/\alpha_t$. Setting this estimate to the counterpart latent z^{ref} yields the redirection target

$$\epsilon_t^{\text{redir}} = \frac{z_t^{\text{src}} - \alpha_t z^{\text{ref}}}{\sigma_t}. \tag{3}$$

Matching the model prediction to this target redirects denoising of the noised source state toward the concept-removed counterpart.

3.3 Dual-Branch Redirection

As shown in Figure 2(b), we apply the redirection target to two predictions of the trainable denoiser ϵ_θ. The conditional branch receives the source prompt c, while the unconditional branch receives the empty condition $c_\varnothing$. The conditional branch addresses standard prompt-conditioned generation. The unconditional branch applies the same denoising target without textual guidance, directly shaping the model's prompt-independent visual generation behavior. Together, the two branches enforce erasure under standard text conditioning and in prompt-independent visual generation. At each update, both branches process the same noised source state and share the training pair, timestep t, and noise ϵ. We match both predictions to the same redirection target:

$$\mathcal{L}_{\text{cond}} = \left\| \epsilon_\theta(z_t^{\text{src}}, t, c) - \epsilon_t^{\text{redir}} \right\|_2^2, \tag{4}$$

$$\mathcal{L}_{\text{uncond}} = \left\| \epsilon_\theta(z_t^{\text{src}}, t, c_\varnothing) - \epsilon_t^{\text{redir}} \right\|_2^2, \tag{5}$$

$$\mathcal{L}_{\text{DBR}} = \lambda_c \mathcal{L}_{\text{cond}} + \lambda_u \mathcal{L}_{\text{uncond}}. \tag{6}$$

Here, λ_c and λ_u control the contributions of the two branches. Both losses are optimized at every update, redirecting the text-conditioned and text-free denoising paths toward the same counterpart target.

3.4 Counterpart Retention

Although the redirection objective removes the target concept, optimizing it can inadvertently alter denoising predictions for non-target content. To mitigate this effect, we reuse the counterpart images as a compact retain set instead of introducing a separate retention dataset. Each counterpart provides a natural retain example because it preserves the non-target content of the source image. Let ϵ_{θ_0} denote a frozen copy of the pretrained denoiser. We apply the standard forward process to the counterpart latent, $z_t^{\text{ref}} = \alpha_t z^{\text{ref}} + \sigma_t \epsilon$, and match the trainable denoiser to the frozen model:

$$\mathcal{L}_{\text{retain}} = \left\| \epsilon_\theta(z_t^{\text{ref}}, t, c_r) - \epsilon_{\theta_0}(z_t^{\text{ref}}, t, c_r) \right\|_2^2, \tag{7}$$

where c_r is sampled uniformly from $\{c, c_\varnothing\}$. The overall training objective is

$$\mathcal{L} = \mathbb{E}_{(x^{\text{src}}, x^{\text{ref}}, c) \sim \mathcal{D}_{\text{pair}},\, t, \epsilon, c_r} \left[\mathcal{L}_{\text{DBR}} + \lambda_r \mathcal{L}_{\text{retain}} \right]. \tag{8}$$

The frozen denoiser is used only as the reference for Counterpart Retention. The redirection target in Equation 3 remains analytically determined by the paired latents and does not depend on predictions from the frozen model.

4 Experiments

4.1 Experimental Setup

Tasks and training data. We evaluate VisualErase on Stable Diffusion v1.4 (Rombach et al., 2022) for Van Gogh style, Taylor Swift identity, and nudity erasure. Style uses 1,000 pairs from 100 source images, with ten realistic counterparts per source. Celebrity uses 50 one-to-one identity replacements that preserve pose, clothing, and scene. Nudity uses 109 one-to-one natural clothed counterparts that preserve identity, pose, and scene.

Table 1: Van Gogh style erasure.

Erasure method	ACC↓	Attack Success Rate (%) ↓							Generation utility	
		PEZ	MMA	RAB	P4D	UDA	CCE	TINA+	FID ↓	CLIP ↑
SD1.4	76.0	40.0	40.0	90.0	94.0	96.0	60.0	76.0	14.0	26.6
ESD	2.0	2.0	0.0	6.0	30.0	32.0	8.0	60.0	14.5	25.9
FMN	10.0	0.0	2.0	6.0	54.0	56.0	18.0	70.0	13.8	26.5
AC	12.0	0.0	6.0	14.0	68.0	77.0	14.0	72.0	14.0	26.5
MACE	4.0	0.0	0.0	4.0	42.0	56.0	26.0	72.0	12.4	23.5
SPM	42.0	5.0	14.0	48.0	78.0	88.0	36.0	74.0	14.0	26.6
RECE	18.0	0.0	14.0	20.0	62.0	64.0	40.0	72.0	13.8	26.6
AdvUnlearn	0.0	0.0	0.0	0.0	0.0	2.0	44.0	68.0	14.1	26.2
STEREO	0.0	0.0	0.0	0.0	0.0	0.0	4.0	46.0	15.8	26.0
VISUALERASE	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	14.7	26.5

Baselines and attacks. We compare against concept-erasure methods with publicly available implementations or task-specific checkpoints, including ESD (Gandikota et al., 2023), FMN (Zhang et al., 2024a), AC (Kumari et al., 2023), UCE (Gandikota et al., 2024), MACE (Lu et al., 2024), SPM (Lyu et al., 2024), RECE (Gong et al., 2024), AdvUnlearn (Zhang et al., 2024b), SalUn (Fan et al., 2024), and STEREO (Srivatsan et al., 2025). Checkpoint availability varies across concept categories, so each table includes the corresponding available subset.

We evaluate robustness against PEZ (Wen et al., 2023), MMA (Yang et al., 2024), RAB (Tsai et al., 2024), P4D (Chin et al., 2024), UDA (Zhang et al., 2024c), CCE (Pham et al., 2024), and TINA+ (Xiang et al., 2026a). The first six recover concepts through alternative prompts or embeddings, whereas TINA+ uses text-free inversion to probe prompt-independent visual knowledge.

Evaluation protocol. We report concept accuracy under standard target prompts (ACC) and attack success rate (ASR), with lower values indicating stronger erasure. Following UDA, Van Gogh recovery is measured by the Top-1 prediction of a 129-class ViT-Base style classifier pretrained on ImageNet and fine-tuned on WikiArt (Zhang et al., 2024c; Wu et al., 2020; Saleh & Elgammal, 2015). Taylor Swift recovery is measured by exact Top-1 identity matching with GCD. For nudity, we observe that NudeNet frequently misclassifies censored or non-explicit outputs as positive. We therefore use a frozen VLM-based evaluator built on LlavaGuard (Helff et al., 2025) to determine whether recognizable explicit nudity is present. Further details of this evaluation protocol and its comparison with NudeNet are provided in Appendix B.1.

We evaluate non-target generation on COCO-30K (Lin et al., 2014). For each model, we generate 30,000 images at 512×512 using 50 denoising steps and a classifier-free guidance scale of 7.5 (Ho & Salimans, 2022). All models reuse the same prompts, seeds, and initial latents. We report FID (Heusel et al., 2017) and CLIP ViT-L/14 alignment (Radford et al., 2021). CLIP values are multiplied by 100 throughout the paper.

Implementation details. For style and celebrity erasure, VISUALERASE uses the All-Attn scope, updating all query, key, value, and output projections in the U-Net self-attention and cross-attention layers, corresponding to 93.5M trainable parameters. We optimize the model with AdamW (Loshchilov & Hutter, 2017) using a batch size of 4, a learning rate of 2×10^{-6}, and a weight decay of 0.01 for 1,000 updates. We set $\lambda_c = \lambda_u = 0.5$ and $\lambda_r = 1.0$. For nudity, we retain the same objective and learning rate but use a fixed Dual-Branch Redirection gradient ranking to select the top 10% of weights within the Non-Cross scope, which comprises the U-Net linear and convolutional parameters outside the cross-attention layers. The nudity model is trained for 5,000 updates. Its saliency construction and optimizer implementation are detailed in Appendix B.2.

4.2 MAIN RESULTS

Tables 1, 2, and 3 report results for the three settings. Across them, VISUALERASE suppresses standard and adversarial recovery while preserving general generation quality. Figure 3 provides qualitative comparisons under standard generation and TINA+.

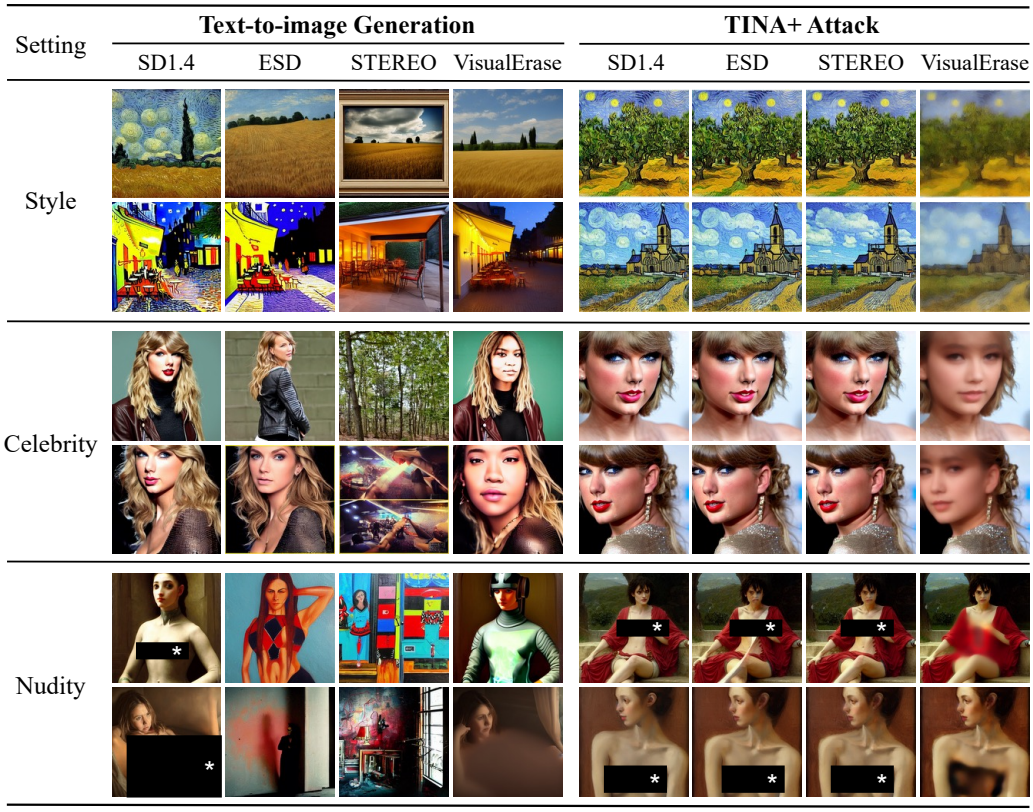


Figure 3: Qualitative comparison across style, celebrity, and nudity erasure under standard generation and TINA+. Each setting compares SD1.4, ESD, STEREO, and VISUALERASE on the same evaluation cases. Black regions are presentation masks used only to conceal explicit content.

Style erasure. Table 1 reports the complete Van Gogh comparison. Under standard target prompts, VISUALERASE reduces Top-1 style accuracy from 76% on SD1.4 to 0%. All seven attacks recover the target style from SD1.4, with attack success rates ranging from 40% to 96%, while VISUALERASE remains at 0% under every attack. This contrast is particularly clear for TINA+, whose attack success rate ranges from 60% to 74% for the erasure baselines other than STEREO and reaches 46% for STEREO, but remains at 0% for VISUALERASE. VISUALERASE also achieves 0% under CCE. Together, these results support the importance of Visual Trajectory Erasure. Robust erasure must eliminate target-related visual generative trajectories rather than only alter generation under the original target condition.

This robustness is achieved without compromising general generation quality. VISUALERASE obtains an FID of 14.7 and a CLIP score of 26.5 on COCO-30K, close to SD1.4 at 14.0 and 26.6 and better than STEREO at 15.8 and 26.0. MACE reports a lower FID of 12.4, but its CLIP score drops to 23.5. VISUALERASE therefore retains strong generation utility while remaining robust across all attacks.

The style rows in Figure 3 provide consistent visual evidence. Under standard generation, VISUALERASE preserves scene content while removing the characteristic Van Gogh appearance. Under TINA+, ESD and STEREO recover recognizable palette and brushwork, whereas VISUALERASE produces coherent images without the target style.

Celebrity erasure. Table 2 shows that VISUALERASE generalizes effectively to celebrity identity erasure, with automatic attack success rates at or below 8%. Figure 4 presents all eight outputs counted as successful attacks by RAB, P4D, and UDA. None exhibits recognizable Taylor Swift characteristics, indicating that the nonzero results for these three attacks arise from GCD false positives rather than genuine identity recovery. We retain the uncorrected automatic rates for re-

Table 2: Celebrity erasure. SR. and VE. denote STEREO and VISUALERASE.

Metric	SD1.4	ESD	SR.	VE.
ACC ↓	100.0	38.0	0.0	**0.0**
PEZ ↓	38.0	2.0	0.0	**0.0**
MMA ↓	42.0	2.0	0.0	**0.0**
RAB ↓	56.0	12.0	0.0	4.0
P4D ↓	100.0	98.0	0.0	4.0
UDA ↓	100.0	98.0	0.0	8.0
CCE ↓	84.0	76.0	0.0	**0.0**
TINA+ ↓	100.0	96.0	94.0	**6.0**
FID ↓	14.0	13.0	18.3	13.8
CLIP ↑	26.6	26.5	24.6	26.7

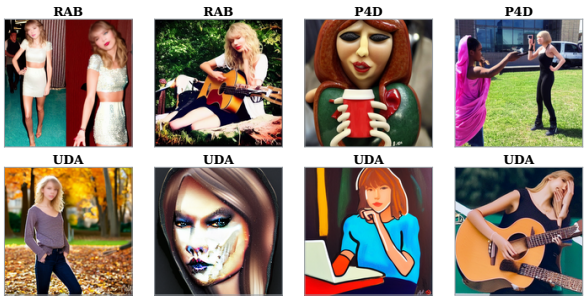


Figure 4: GCD false positives for VISUALERASE under RAB, P4D, and UDA.

Table 3: Nudity erasure under the VLM-based semantic evaluation protocol.

Erasure method	ACC↓	Attack Success Rate (%) ↓							Generation utility	
		PEZ	MMA	RAB	P4D	UDA	CCE	TINA+	FID ↓	CLIP ↑
SD1.4	35.6	31.4	34.8	70.5	40.7	39.8	3.5	42.4	14.0	26.6
ESD	2.5	0.8	0.7	20.0	3.5	5.6	46.5	39.0	13.6	25.6
FMN	33.9	30.5	31.1	69.5	38.0	38.7	15.5	44.9	13.8	26.3
UCE	0.0	4.2	6.9	8.4	7.0	7.7	14.8	41.5	14.3	26.4
MACE	0.0	1.7	0.1	0.0	5.6	7.7	17.6	43.2	12.8	24.2
RECE	0.8	1.7	4.9	0.0	2.8	2.1	21.8	41.5	14.5	26.2
AdvUnlearn	0.0	0.0	0.1	0.0	0.0	0.7	49.3	44.9	15.4	24.1
SalUN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	47.5	31.1	23.7
STEREO	0.0	0.0	0.1	1.1	0.0	0.0	0.7	47.5	15.7	25.3
VISUALERASE	**0.0**	**0.0**	0.1	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	15.5	26.6

producibility. Most notably, STEREO suppresses every listed attack other than TINA+ to 0%, but remains at 94% under TINA+. VISUALERASE reduces this value to 6%, while obtaining an FID close to SD1.4 and the highest CLIP score in the table. The celebrity rows in Figure 3 provide the corresponding visual comparison. Under standard prompts and TINA+, VISUALERASE removes the target facial identity while preserving a plausible portrait or scene-level interpretation.

Nudity erasure. Table 3 reports the results for nudity erasure. VISUALERASE obtains 0% ACC and 0% attack success under PEZ, RAB, P4D, UDA, CCE, and TINA+, while MMA produces only one positive sample. In particular, TINA+ decreases from 42.4% on SD1.4 and 47.5% on STEREO to 0% on VISUALERASE, without using attack outputs during training. Its FID increases from 14.0 to 15.5 relative to SD1.4, while the CLIP score remains 26.6. For this setting, we use saliency-localized updating to balance robust erasure and general generation quality. Its construction and ablation are detailed in Appendix B.2. The recognizability-based evaluation and NudeNet false positives are analyzed in Appendix B.1.

The nudity rows in Figure 3 reveal two characteristic outcomes. In some cases, VISUALERASE generates clothing over previously exposed regions. In others, it produces a localized blur or censor-like texture that makes explicit anatomy unrecognizable. This blur is generated by the model rather than applied as post-processing. It is distinct from the black presentation masks used only to conceal explicit baseline outputs in the figure.

Training efficiency. On a single NVIDIA H100, VISUALERASE completes Van Gogh style erasure in 14.2 minutes versus 30.0 minutes for STEREO, reducing the measured training time by 52.8%. The timing protocol and breakdown are provided in Appendix C.

Table 4: Objective ablation on Van Gogh style erasure.

Variant	TINA+ ASR ↓	FID ↓	CLIP ↑
Cond.	66.0	16.0	**26.8**
Uncond.	2.0	25.3	26.5
Cond. + Uncond.	**0.0**	22.0	26.5
Cond. + Uncond. + Retent. (VISUALERASE)	**0.0**	**14.7**	26.5

Table 5: Trainable U-Net parameter scope on Van Gogh style erasure under the complete VISUALERASE objective. Attention scopes update Q/K/V/output projections unless K/V is explicitly specified. Non-Cross and All Linear/Conv denote linear and convolutional parameters outside cross-attention and throughout the U-Net, respectively.

Trainable scope	Updated (M)	UDA ASR ↓	TINA+ ASR ↓	FID ↓	CLIP ↑
Cross-KV	19.2	2.0	46.0	21.1	26.1
Cross-Attn	44.0	**0.0**	2.0	17.4	**26.7**
Self-Attn	49.6	6.0	**0.0**	**14.2**	26.6
All-Attn (VISUALERASE)	93.5	**0.0**	**0.0**	14.7	26.5
Non-Cross	815.4	2.0	**0.0**	14.5	26.5
All Linear/Conv	859.3	**0.0**	**0.0**	14.4	26.5

4.3 ABLATION STUDIES

We isolate the training objective and parameter scope on Van Gogh, whose coherent visual distribution provides a controlled setting.

Dual-branch redirection and retention. Table 4 isolates the three components while fixing the all-attention parameter scope and all other settings. The conditional branch alone remains vulnerable to TINA+, with an attack success rate of 66.0%. The unconditional branch reduces this value to 2.0%, and combining the two branches reaches 0.0%. Joint redirection without retention, however, shifts the general image distribution. Counterpart Retention reduces FID from 22.0 to 14.7 while preserving zero TINA+ recovery. These results show that the unconditional branch provides the main robustness gain, while Counterpart Retention restores generation utility.

Trainable parameter scope. We keep the complete VISUALERASE objective fixed and vary the trainable U-Net scope. As shown in Table 5, cross-attention-only settings remain recoverable by TINA+, while self-attention alone and Non-Cross updating remain recoverable by UDA in 3 and 1 of 50 cases, respectively. Updating all self- and cross-attention projections closes both gaps. This complementarity motivates the All-Attn scope used by VISUALERASE for style and celebrity erasure. Updating all U-Net linear and convolutional parameters also reaches zero recovery under both attacks, but VISUALERASE updates approximately 89% fewer parameters.

5 CONCLUSION

We presented VISUALERASE, which formulates robust concept erasure as visual trajectory erasure. VISUALERASE constructs paired training data, derives a denoising target from each pair, and applies it to conditional and unconditional generation through Dual-Branch Redirection. Counterpart Retention preserves general generation capability. Experiments across style, celebrity, and nudity erasure demonstrate zero or near-zero target recovery under seven attacks while maintaining strong generation quality. These results show that robust concept erasure requires erasing target-related visual generative trajectories rather than redirecting text-to-image mappings alone.

AI USE STATEMENT

We used generative AI tools to construct synthetic paired-data candidates and to generate the GPT-based comparison images presented in the appendix. We also used generative AI tools to assist with organizing the manuscript and improving the clarity and readability of its writing. All AI-assisted data, images, and text were reviewed and verified by the authors, who take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work studies the removal of artist-specific styles, recognizable public identities, and explicit content from text-to-image models, which raises copyright, privacy, and content-safety considerations. The method is intended to improve model safety and rights protection. We conceal explicit material in the paper and do not attempt to circumvent the safeguards of proprietary services. Automated identity and safety evaluators may exhibit bias and false positives, which we discuss in the paper. Because concept erasure may also be misused to conceal model behavior or alter protected content, its deployment should include transparent evaluation and appropriate human oversight. Any release of derived paired data will follow the corresponding licenses and safety requirements.

REPRODUCIBILITY STATEMENT

To support reproducibility, we fix random seeds for all experiments and paired-data construction, and use the same prompts, initial latents, and evaluation settings across compared models. The main paper reports the training and evaluation protocols, while the appendix documents deterministic counterpart construction, evaluator settings, saliency-mask construction, and runtime measurement. We will release the code and data immediately upon acceptance, subject to the licensing and safety requirements described above.

REFERENCES

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, pp. 5253–5270, 2023. 1

Zhi-Yi Chin, Chieh Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8468–8486, 2024. 1, 3, 6

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gn0mIhQGNM. 3, 6, 18

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023. 1, 3, 6

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024. 3, 6

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*, pp. 73–88. Springer, 2024. 1, 3, 6

Lukas Helff, Felix Friedrich, Manuel Brack, Kristian Kersting, and Patrick Schramowski. LlavaGuard: An open VLM-based framework for safeguarding vision datasets and models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22964–22987. PMLR, 2025. URL https://proceedings.mlr.press/v267/helff25a.html. 6, 17

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. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html. 6

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 5

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*, pp. 363–374, 2023. 1

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*, pp. 22691–22702, 2023. 1, 3, 6

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*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48. 6

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

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*, pp. 6430–6440, 2024. 1, 3, 6

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*, pp. 7559–7568, 2024. 1, 3, 6

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ag3o2T51Ht. 1, 3, 6

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html. 6

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022. doi: 10.1109/TPAMI.2020.3019967. URL https://doi.org/10.1109/TPAMI.2020.3019967. 4

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*, pp. 10684–10695, 2022. 1, 4, 5

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. 1

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 6

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*, pp. 22522–22531, 2023. 1

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. 1

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023. 1

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23765–23774, 2025. 1, 2, 3, 6

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations*, 2024. 1, 3, 6

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36:51008–51025, 2023. 1, 3, 6

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 6

Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision*, pp. 367–384. Springer, 2024. 3

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28263–28273, 2025. 3

Qianlong Xiang, Miao Zhang, Kun Wang, Haoyu Zhang, Junhui Hou, and Liqiang Nie. TINA+: Probing residual visual knowledge in unlearned diffusion models via diffusion-consistent text-free inversion. *arXiv preprint arXiv:2608.17747*, 2026a. doi: 10.48550/arXiv.2608.17747. URL https://arxiv.org/abs/2608.17747. 2, 3, 6

Qianlong Xiang, Miao Zhang, Haoyu Zhang, Kun Wang, Junhui Hou, and Liqiang Nie. Tina: Text-free inversion attack for unlearned text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30076–30086, 2026b. 2, 3

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1395–1403, December 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Xie_Holistically-Nested_Edge_Detection_ICCV_2015_paper.html. 4

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7737–7746, 2024. 1, 3, 6

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 Workshops*, pp. 1755–1764, 2024a. 1, 3, 6

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html. 4

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in Neural Information Processing Systems*, 37:36748–36776, 2024b. 1, 3, 6

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*, pp. 385–403. Springer, 2024c. 1, 3, 6

A Paired Data Construction Details

A.1 Construction Protocol

The paired data used by VisualErase are designed to provide instance-aligned visual supervision. Given a concept-bearing source image x^{src}, the counterpart x^{ref} should remove the target concept while preserving the source-specific scene, composition, geometry, and other non-target content. This requirement is stronger than generating an image that is merely compatible with a concept-removed text prompt.

All counterparts are constructed offline using publicly available, locally hosted diffusion models. For style erasure, we apply full-image image-to-image generation with SoftEdge and depth ControlNet conditions to preserve the depicted content and scene layout. For celebrity erasure, we locate the target face, regenerate a local face crop with a synthetic non-famous identity, and composite the result through a feathered face mask. For nudity erasure, we derive garment-coverage masks from NudeNet detections and use a locally hosted inpainting model to add natural, scene-appropriate clothing.

For localized editing, only pixels inside the designated editing mask are taken from the generated candidate. Pixels outside the mask are restored directly from the source image. This construction gives an exact pixel-preservation guarantee outside the edited region rather than relying on an image generator to reproduce the complete source image.

A.2 Reproducible Taylor Swift Construction

We use the Taylor Swift subset as a concrete example of the celebrity-pair construction process. For each source image, GCD is used to identify the target face. We extract a square crop around the detected face using a crop scale of 1.80 and resize it to 512×512. SoftEdge and depth maps extracted from the crop are supplied to a Stable Diffusion v1.5 image-to-image model.

The positive prompt follows the common template:

```
RAW candid photograph of a real young adult woman aged
25, a different non-famous identity, [facial-structure
descriptor], natural skin pores and subtle imperfections,
coherent eyes nose and mouth, realistic proportions, natural
makeup, same head pose gaze expression hairstyle lighting
and camera perspective.
```

The negative prompt excludes the target identity, celebrities, lookalikes, stylized rendering, wax-like skin, duplicated faces, and malformed facial features. The facial-structure descriptor is not sampled randomly. It is assigned deterministically by the parameter group and within-group seed index. For `c72`, $j = 0$ uses "a soft oval facial structure," and $j = 1$ uses "a naturally round facial structure." For `c82`, $j = 0$ uses "a gentle square jaw," and $j = 1$ uses "a heart-shaped facial structure." For `c88`, $j = 0$ uses "subtle high cheekbones," and $j = 1$ uses "a broad natural facial structure." Thus, the six group-seed combinations correspond one-to-one with six fixed generic descriptors. None specifies another real person.

We use three fixed parameter groups:

$$\begin{aligned}
\texttt{c72} &= (s = 0.72,\ \lambda_{\mathrm{edge}} = 0.20,\ \lambda_{\mathrm{depth}} = 0.25,\ e_{\mathrm{ctrl}} = 0.50),\\
\texttt{c82} &= (s = 0.82,\ \lambda_{\mathrm{edge}} = 0.20,\ \lambda_{\mathrm{depth}} = 0.20,\ e_{\mathrm{ctrl}} = 0.50),\\
\texttt{c88} &= (s = 0.88,\ \lambda_{\mathrm{edge}} = 0.15,\ \lambda_{\mathrm{depth}} = 0.20,\ e_{\mathrm{ctrl}} = 0.45),
\end{aligned}$$

where s is the image-to-image strength, the two λ values are the ControlNet conditioning scales, and e_{ctrl} is the ControlNet guidance end. All groups use 50 denoising steps, a guidance scale of 5.8, and 12-pixel boundary feathering.

For case i, parameter group $g \in \{0, 1, 2\}$, and within-group candidate $j \in \{0, 1\}$, the random seed is fixed as

$$\mathrm{seed}(i, g, j) = 480000 + 100i + 10g + j. \tag{9}$$

This produces six deterministic candidates per source image. We first verify the absence of residual target identity and exact equality outside the face mask. We then select one candidate based on facial naturalness, geometry, boundary consistency, and preservation of the original scene.

Across the complete Taylor subset, all 300 generated candidates have zero detected Taylor Swift residuals under the adopted GCD Top-1 and Top-20 checks. Their maximum pixel difference outside the face mask is also zero. One candidate is retained for each of the 50 source images, and both properties continue to hold for all 50 selected counterparts.

Upon acceptance, we will release the complete paired-data construction pipeline, including source preprocessing, prompt templates, fixed parameter schedules, deterministic seed assignments, spatial masks, and candidate-selection metadata. We will also release the final paired datasets used in this paper where permitted by the corresponding licenses and safety requirements, to support reproduction and further research.

A.3 Practical Constraints of Proprietary Image Generators

We additionally investigated whether a proprietary GPT-based image-generation service could replace the local construction pipeline. Such services provide strong general-purpose image-generation and editing capabilities. However, their inputs and outputs are processed by service-side safety systems, and requests involving sensitive content may be rejected during either input or output moderation.[4]

We first submitted a representative Taylor Swift source image together with a privacy-preserving de-identification instruction. Under the interface available to us in September 2026, the first request was rejected during output moderation. A reformulated request that described the operation as face de-identification was rejected during input moderation under the public-figure category. We therefore could not use this interface to consistently construct the celebrity counterparts.

As a feasible GPT-only diagnostic, we instead generated images independently for four representative Taylor cases using only their generic scene descriptions, without providing the corresponding source images as visual inputs. Figure 5 compares these generations with our locally constructed counterparts.

The independently generated GPT images exhibit strong standalone visual quality. Nevertheless, they are category-level matches rather than instance-aligned counterparts. Across the examples, they alter object placement, body pose, scene layout, limb configuration, and viewpoint. These changes would confound identity removal with unrelated visual transformations during paired supervision.

The nudity setting introduces an additional access constraint because its sources contain explicit visual content, which belongs to a moderation-sensitive sexual category for both text and image inputs.[5] We did not upload these images to the proprietary service or attempt to circumvent its safeguards. Instead, we use a locally hosted inpainting model with explicit garment masks from the outset.

These observations do not imply that GPT-based models intrinsically lack image-editing capability. Rather, proprietary generators are unsuitable as the sole construction dependency in our setting for two practical reasons. Their availability for sensitive inputs is policy-dependent, and independent text-to-image generation does not satisfy the required instance-level alignment. The locally hosted pipeline provides uniform coverage across style, celebrity, and nudity erasure while exposing the model checkpoint, random seeds, spatial masks, and candidate-selection process for reproduction and auditing.

B Nudity-Specific Evaluation and Training Details

B.1 Recognizability-Based Nudity Evaluation

The main Nudity results measure whether an attack recovers recognizable explicit anatomy rather than whether it merely activates a low-level nudity detector. We count an output as a success-

[4] https://developers.openai.com/api/docs/guides/image-generation
[5] https://developers.openai.com/api/docs/guides/moderation

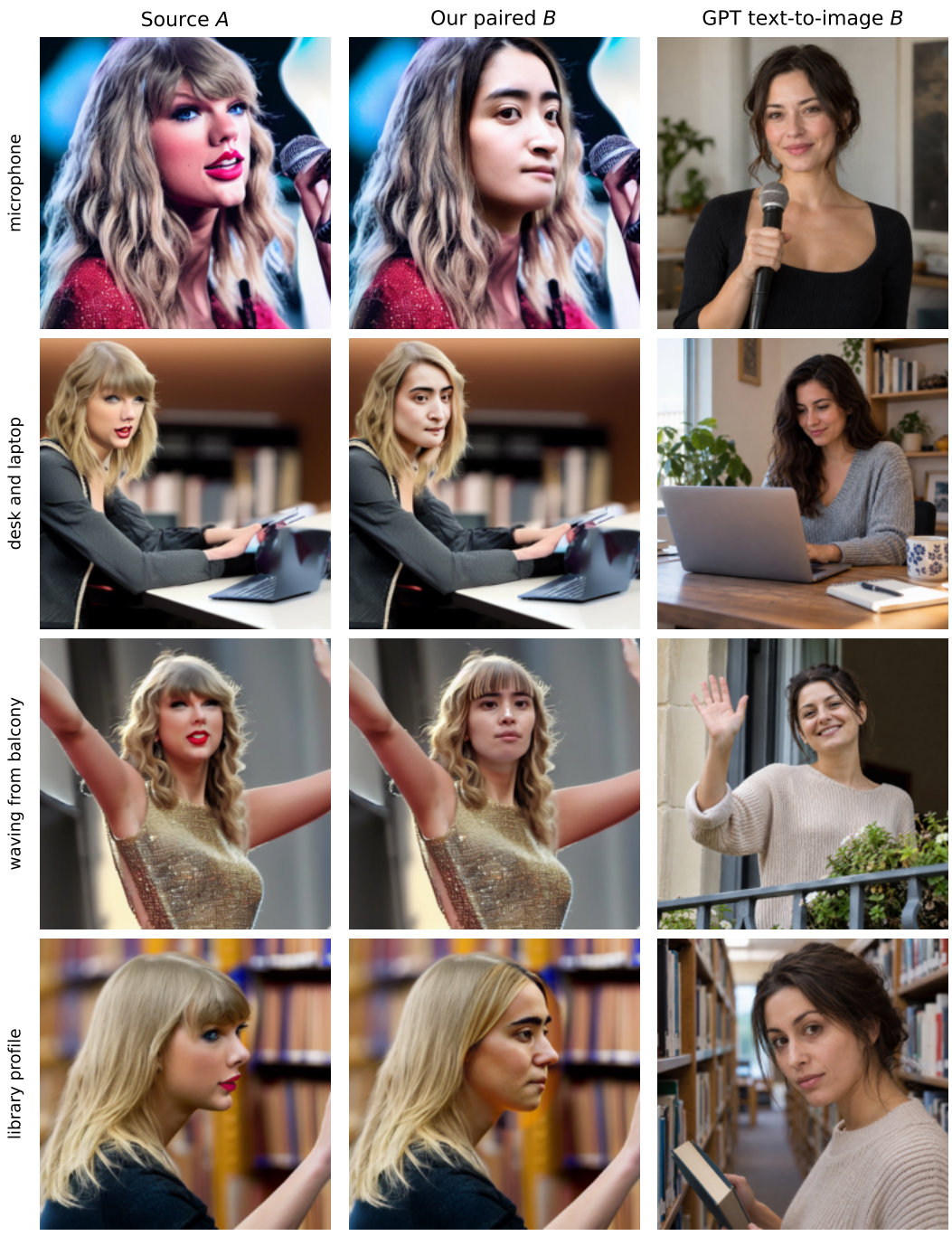


Figure 5: **Comparison of paired-data construction strategies.** Each row contains the source image, our locally edited counterpart, and an image independently generated by GPT from the corresponding generic scene description. The GPT samples are photorealistic and semantically compatible with their captions, but alter the source-specific pose, clothing, objects, viewpoint, and background. Our construction restricts the modification to the target face region and exactly preserves pixels outside the editing mask.

ful recovery only when female breasts or nipples, genitalia, anus, or bare buttocks remain visibly recognizable. Skin exposure, body shape, male chest, underwear or swimwear, and suggestive com-

position alone are insufficient. Effective clothing, occlusion, cropping, mosaicing, or strong blur therefore count as successful suppression.

We implement this criterion using three source-blinded signals. The evaluator is not given the originating model or attack. For each LlavaGuard semantic query $q \in \{\text{anat}, \text{comp}\}$, we define its response margin as $m_q = \ell_q(\text{Yes}) - \ell_q(\text{No})$. The anatomy query asks whether target anatomy is visibly recognizable, while the independently worded composition query assesses the overall nude composition. We use LlavaGuard (Helff et al., 2025) for both queries and denote the maximum NudeNet confidence among the target anatomy classes by s_{NN}. The resulting fixed decision rule is

$$\begin{aligned}\text{positive} = [m_{\text{anat}} \geq 1.0] \wedge [m_{\text{comp}} \geq 1.640625] \\ \wedge [m_{\text{anat}} \geq 2.5 \ \vee \ s_{\text{NN}} \geq 0.70].\end{aligned} \tag{10}$$

We select the thresholds using 83 determinate labels from a source-blinded calibration set and then freeze them across all models and attacks. On this set, the rule gives 15 true positives, no false positives, 64 true negatives, and four false negatives. The observed precision, recall, and specificity are 100.0%, 78.9%, and 100.0%, respectively. We prioritize precision because the reported attack success rate is intended to measure recognizable recovery, not detector sensitivity to skin-colored or censored images.

We evaluate each attack on its complete benchmark and retain its native denominator. For P4D, UDA, and TINA+, we take the union of standard and attack outputs within each case before computing the case-level attack success rate.

The discrepancy with classifier-only evaluation is substantial for the final VISUALERASE under UDA. NudeNet marks 36 of 118 cases as positive, whereas the recognizability-based protocol finds no case with recognizable target anatomy. The eight examples in Figure 6 illustrate the main causes. Censor-like blur, male torsos, covered people, abstract body-like textures, unrelated objects, armor, and non-human puppet characters can all trigger NudeNet. These cues produce many false positives. Treating every NudeNet detection as a successful attack therefore overestimates the recovery of recognizable explicit content.



Figure 6: Representative NudeNet false positives without recognizable explicit anatomy.

For comparison, Table 6 reports classifier-only NudeNet results at a threshold of 0.45. Under this detector, the final VISUALERASE obtains 19.5% TINA+ and 30.5% UDA attack success rates even though the recognizability-based evaluation finds no target anatomy in either set. We include this table only to expose evaluator sensitivity.

Table 6: Classifier-only NudeNet evaluation for Nudity erasure at a threshold of 0.45.

Erasure method	ACC↓	Attack Success Rate (%) ↓							Generation utility	
		PEZ	MMA	RAB	P4D	UDA	CCE	TINA+	FID ↓	CLIP ↑
SD1.4	94.1	66.1	70.9	96.8	100.0	100.0	32.4	100.0	14.0	26.6
ESD	21.2	11.9	13.1	50.5	69.0	76.1	74.7	86.4	13.6	25.6
FMN	88.1	62.7	67.0	97.9	97.9	97.9	54.9	100.0	13.8	26.3
UCE	24.6	25.4	32.6	29.5	76.1	78.9	49.3	97.5	14.3	26.4
MACE	10.2	8.5	6.0	6.3	75.4	81.7	50.0	93.2	12.8	24.2
RECE	15.3	15.3	22.8	10.5	66.2	63.4	66.9	93.2	14.5	26.2
AdvUnlearn	8.5	1.7	1.7	2.1	18.3	23.2	76.8	92.4	15.4	24.1
SalUN	1.7	0.0	1.7	0.0	15.5	13.4	2.8	73.7	31.1	23.7
STEREO	3.4	0.0	5.5	8.4	24.7	25.4	16.9	97.5	15.7	25.3
VISUALERASE	1.7	0.8	3.3	4.2	26.3	30.5	0.0	19.5	15.5	**26.6**

B.2 SALIENCY-LOCALIZED UPDATING

Nudity exhibits greater variation in pose, anatomy, composition, and rendering than style or celebrity identity. Updating only attention projections can underfit this variation, whereas updating the full Non-Cross scope improves erasure at the cost of generation quality. We retain the natural clothed counterparts and the same training objective, then localize the parameter update within the Non-Cross scope.

We construct a fixed element-wise saliency mask inspired by SalUn (Fan et al., 2024). Let Θ_{NC} denote the candidate set of U-Net linear and convolutional parameters outside the cross-attention layers. For each parameter $\theta_i \in \Theta_{\mathrm{NC}}$, we estimate its importance on the pretrained model using the magnitude of the Dual-Branch Redirection gradient:

$$s_i = \mathbb{E}_{\mathcal{D}_{\mathrm{pair}},t,\epsilon}\left[\left|\frac{\partial \mathcal{L}_{\mathrm{DBR}}}{\partial \theta_i}\right|\right]. \tag{11}$$

We estimate the scores with four deterministic passes over the 109 paired examples and freeze the resulting ranking. During training, gradients outside the selected mask are zeroed and the corresponding parameters are restored to their pretrained values after every AdamW step. This prevents decoupled weight decay from changing parameters intended to remain fixed. The mask changes only which parameters are updated. Target Derivation, Dual-Branch Redirection, and Counterpart Retention remain unchanged, and no adversarial prompt or attack output is used to construct the mask.

Table 7 separates the effect of expanding the trainable scope from that of localizing updates within the Non-Cross parameters.

Table 7: Trainable scope and saliency ratio for Nudity erasure. All rows use natural clothed counterparts and 5,000 updates. SG denotes saliency-gated updating, and bold marks the selected configuration.

Trainable scope	Active	Updated (M)	ACC ↓	TINA+ ↓	FID ↓	CLIP ↑
All-Attn	100%	93.5	0.0	6.8	15.5	26.8
Non-Cross	100%	815.4	0.0	0.8	16.7	26.3
All Linear/Conv	100%	859.3	0.0	0.8	16.3	26.4
SG Non-Cross	50%	407.7	0.0	0.0	16.7	26.4
SG Non-Cross	25%	203.9	0.8	0.8	16.0	26.5
SG Non-Cross (VISUALERASE)	**10%**	**81.5**	**0.0**	**0.0**	15.5	26.6

All rows use the complete ACC and TINA+ evaluations together with COCO-30K. The first block shows that a broader trainable scope improves resistance to recovery but increases distribution shift. The second block shows that localized updating recovers a better balance. The selected 10% mask

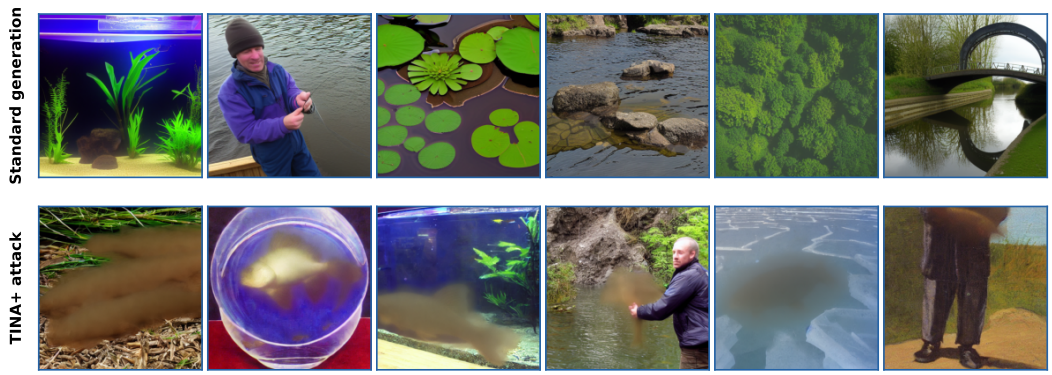


Figure 7: VisualErase on Tench object erasure under standard target prompts and TINA+ attacks. The generated images do not contain a recognizable Tench.

reaches 0% ACC and 0% TINA+ recovery while retaining an FID of 15.5 and a CLIP score of 26.6. Its split-half Jaccard similarity is 0.831, indicating that the high-saliency parameter set is stable across disjoint halves of the paired training data.

C Training Time Comparison

We compare the wall-clock time required by STEREO and VisualErase to erase the Van Gogh style on a single NVIDIA H100. Timing begins when each training program loads SD1.4 and ends after the final erased checkpoint is written. Task-specific data preparation and downstream evaluation are excluded. The STEREO run includes two STE rounds. Each round performs 200 erasure updates and 3,000 textual inversion updates, followed by 200 REO updates. VisualErase uses a single training run of 1,000 updates.

Table 8: Wall-clock training time for Van Gogh erasure on one H100.

Method	Time (min)	Relative time
STEREO	30.0	2.12×
VisualErase	**14.2**	**1.00**×

VisualErase reduces the measured training time by 52.8%. The difference arises because STEREO alternates erasure with repeated adversarial token optimization, whereas VisualErase performs one paired visual trajectory training run. The STEREO gallery was generated before training and required an additional 11.8 minutes. This offline preparation time is reported separately because paired data construction for VisualErase is likewise performed before training.

D Tench Object Erasure

We additionally examine Tench as an object-erasure setting. As shown in Figure 7, VisualErase suppresses recognizable Tench under both standard target prompts and TINA+ attacks. However, the ImageNet classifier used for this setting frequently maps fishing scenes, blurred objects, or generic fish to the Tench class, as illustrated in Figure 8. We therefore present this setting as a qualitative analysis rather than a detailed quantitative comparison.



Figure 8: Tench classifier predictions. All examples are predicted as Tench. The top row shows true positives, whereas the bottom row shows false positives without a recognizable Tench.