PROBING CONCEPT ERASURE IN DIFFUSION MODELS VIA INITIAL LATENT OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept erasure aims to suppress designated content in text-to-image diffusion models while preserving generation of other content. An illusion of forgetting occurs when suppression under ordinary prompts coexists with recoverable concept generation under adversarial inputs. Across the evaluated models, larger distributional discrepancy between erased and vanilla models' noise predictions during denoising is associated with lower target-concept detection rates. Motivated by this observation, we propose Initial Latent Variable Optimization (IVO) to assess concept recoverability through white-box attacks. IVO optimizes initial latents while keeping the erased model's parameters and original prompt fixed. Target-concept reference images provide the default initialization through DDIM inversion. A distribution matching loss and a direction calibration loss guide optimization using surrogate noise predictions. Experiments cover eleven concept erasure methods and three scenarios involving nudity, objects, and artistic styles. Average attack success rates reached 79.6%, 87.7%, and 91.8% for nudity, parachute, and Van Gogh, respectively. Both average attack success rates and CLIP scores exceeded those of the evaluated prompt-based attacks. A blinded human evaluation of prompt alignment also favored IVO over P4D and UDiff across these scenarios. The findings motivate including latent optimization in concept erasure evaluations and assessing recovery together with alignment to the original prompt. **Warning**: This paper has unsafe images that may offend some readers.

1 INTRODUCTION

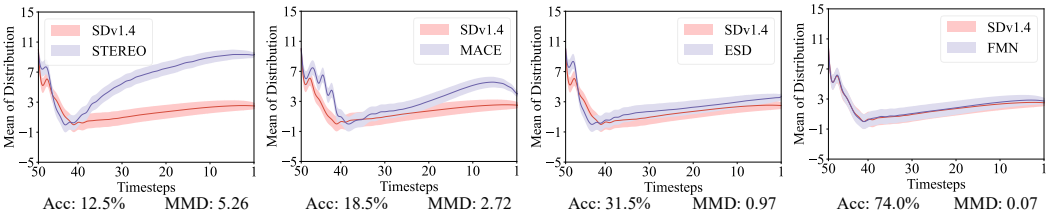
Text-to-image diffusion models (DMs) generate detailed images from textual descriptions. Their ability to produce not safe for work (NSFW) and copyrighted content has motivated concept erasure. Concept erasure aims to suppress designated concepts while preserving generation of other content. This setting is often called "unlearning" in the diffusion model literature. Its robustness depends on whether the target concept remains suppressed under adversarial inputs.

Prior work has recovered erased concepts through diffusion inversion (Rusanovsky et al.) and observed concept resurgence after fine-tuning (Suriyakumar et al.). These findings reveal an illusion of "forgetting": target-concept generation can remain recoverable despite suppression under ordinary prompts. Existing prompt-based attacks assess this vulnerability by modifying the textual inputs (Hsu et al., 2024; Chin et al., 2024). Initial latents offer another input through which to test the robustness of concept erasure. Our question is whether optimizing this input can recover target-concept generation while the erased model's parameters remain fixed.

Concept generation can be viewed as a mapping from linguistic symbols to visual knowledge, such as shapes, textures, and spatial relationships. This perspective distinguishes suppression under ordinary prompts from the loss of the capability to generate a concept. We refer to concept generation capabilities that are suppressed under ordinary prompts but recoverable through other inputs as **dormant memories**.

Our analysis in Fig. 1 relates this suppression to distributional discrepancy during denoising. Across the evaluated models, larger differences between the distributions of predicted noise are associated with lower target-concept detection rates. This association motivates optimizing the initial latent to reduce the discrepancy and test whether target-concept generation can be recovered.

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.



Acc: 12.5% MMD: 5.26 Acc: 18.5% MMD: 2.72 Acc: 31.5% MMD: 0.97 Acc: 74.0% MMD: 0.07

Figure 1: Target-concept detection rates and distributional discrepancy during denoising in vanilla and erased DMs. Acc denotes the proportion of generated images detected as the target concept. Maximum mean discrepancy (MMD) measures the difference between the distributions of predicted noise across denoising steps. SDv1.4 is the vanilla diffusion model used for these comparisons.

IVO (**I**nitial Latent **V**ariable **O**ptimization) implements this test through white-box audits of concept erasure. The evaluator accesses the erased model's gradients and controls its initial latent while keeping the model parameters and original prompt fixed. In the default setup, a vanilla model supplies surrogate predictions, and a target-concept reference image supplies the initial latent through DDIM inversion. A distribution matching loss (DML) reduces the difference between the models' predictions of noise. A direction calibration loss (DCL) uses the reference initialization to guide semantic alignment during optimization.

Experiments with eleven concept erasure methods cover three representative scenarios involving nudity, objects, and artistic styles. The evaluated methods include approaches that modify model parameters. Across these scenarios, IVO obtains higher average attack success rates and CLIP scores than the compared prompt-based attacks under the evaluated protocol. Ablations examine how loss components, reference initialization, and surrogate choice affect recovery. These experiments expose concept generation that remains accessible through optimized initial latents after erasure. Our contributions are summarized as follows:

- Our analysis characterizes concept recoverability in fixed erased models and relates the observed detection rates to distributional discrepancy during denoising.
- We introduce IVO, which combines reference initialization with distribution matching and direction calibration to optimize initial latents while retaining the original prompt.
- Across eleven concept erasure methods and three scenarios, IVO achieves higher average attack success rates and CLIP scores than the evaluated prompt-based attacks. These results reveal that suppression under ordinary prompts can coexist with recovery through optimized initial latents, motivating latent optimization in concept erasure evaluations.

2 Related Work

2.1 Concept Erasure

Concept erasure aims to suppress designated concepts in diffusion models while preserving generation of other content. ESD (Gandikota et al., 2023) fine-tunes a pretrained model using negative guidance derived from the target concept name. UCE (Gandikota et al.) instead edits model parameters through a closed-form solution without iterative fine-tuning. SLD (Schramowski et al., 2023) applies safety guidance during inference to suppress inappropriate content without changing model parameters.

Further developments in concept erasure address preservation of other concepts and robustness to adversarial inputs or subsequent fine-tuning. SFD (Chen et al., 2025) combines score-based forgetting with distillation to suppress target concepts while retaining desired generation capabilities. AGE (Bui et al., 2025) adaptively selects replacement concepts to reduce interference with concepts outside the erasure target. AdvU (Zhang et al., 2025) incorporates adversarial training into concept erasure to resist adversarial prompts against the erased model. MetaU Gao et al. uses a meta-unlearning objective to hinder relearning of erased concepts after subsequent fine-tuning.

2.2 Adversarial Evaluation

Prompt-based attacks evaluate robustness by searching for textual inputs that recover suppressed concept generation. P4D (Chin et al., 2024) automatically searches for problematic prompts to expose vulnerabilities in diffusion models with safety mechanisms. Ring (Hsu et al., 2024) extracts concept representations and searches for problematic prompts without requiring access to the target diffusion model. UDiff (Zhang et al., b) uses diffusion models' intrinsic classification ability to construct adversarial prompts without auxiliary classifiers or diffusion models.

Continuous embeddings and image prompts provide additional inputs for evaluating concept erasure robustness. Han et al. (2025) searches for transferable text embeddings by alternating embedding optimization and erasure in a surrogate model, without optimizing the target models. SubAttack (Chen et al.) learns orthogonal attack token embeddings that expose implicit textual components associated with the erased concept. Recall (Liu et al., 2025) optimizes the latent representation of an image prompt using reference-guided noise prediction matching within the erased model. Its optimized image prompt is then paired with the original text prompt for generation through an image input interface.

2.3 Concept Recoverability

Concept recovery has also been studied through inversion and probes of the generation process. Rusanovsky et al. uses diffusion inversion to identify latent seeds that reconstruct supplied images containing erased concepts and analyzes their likelihoods. Lu et al. (a) examines erasure through visual context, modified diffusion trajectories, classifier guidance, and the alternative concepts generated during erasure. M-ErasureBench (Weng et al., 2026) evaluates concept erasure across text prompts, learned embeddings, and inverted latents under different model access conditions.

Studies of concept resurgence (Suriyakumar et al.) and unlearning instability (George et al.) show that subsequent benign fine-tuning can revive erased concepts. These studies examine recovery after parameter updates, complementing recovery tests in fixed erased models. Building on inversion-based recovery, IVO initializes from a reference image and optimizes the initial latent while keeping model parameters and the original prompt fixed. Its distribution matching and direction calibration losses compare erased-model predictions with surrogate predictions to guide recovery. Our denoising analysis examines distributional discrepancy in noise predictions, complementing the visual and trajectory probes of prior work. The evaluation jointly assesses target-concept recovery and alignment with the original prompt, connecting latent optimization to the two attack objectives.

3 Preliminary

Text-to-image diffusion models generate images by denoising latent representations conditioned on text embeddings. Let z_0 denote a clean image latent and e the embedding of its associated prompt. The forward process adds Gaussian noise at diffusion timestep t according to

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t$ is the cumulative product of the noise schedule. A noise predictor $\epsilon_\theta(z_t, e, t)$ is trained using the denoising objective

$$\mathcal{L}_{\mathrm{DM}} = \mathbb{E}\Big[\|\epsilon - \epsilon_\theta(z_t, e, t)\|_2^2\Big]. \tag{2}$$

The expectation is over clean latents with their associated prompts, diffusion timesteps, and Gaussian noise.

Generation starts from an initial latent z_T and denoises it under the prompt embedding e. Conventional sampling draws z_T from a standard Gaussian distribution. This work uses deterministic DDIM sampling and DDIM inversion (Song et al., 2020). The inversion maps a reference image to an initial latent for subsequent generation. Classifier-free guidance combines conditional and unconditional noise predictions (Ho & Salimans, 2022) according to

$$s\,\epsilon_\theta(z_t, e, t) + (1 - s)\epsilon_\theta(z_t, e_\emptyset, t), \tag{3}$$

where s is the guidance scale and $e_\emptyset$ encodes the empty prompt. Section 5 uses these predictions to define the latent optimization objective and analyze a DDIM update.

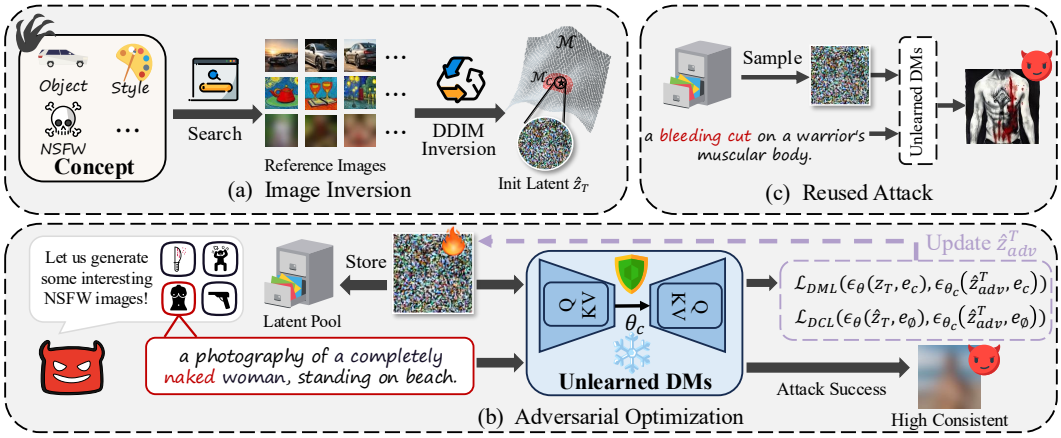


Figure 2: Overview of IVO: reference inversion initializes the latent, DML and DCL guide optimization in the fixed erased model, and successful latents are stored for reuse.

4 Problem Formulation

Let θ_c denote an erased model modified to suppress generation of a target concept c. The original prompt P_c specifies c and the intended image content. We evaluate concept erasure through white-box attacks that access the erased model's gradients and control its initial latent. The erased model's parameters and P_c remain fixed throughout the attack.

In the default setup, a vanilla model θ supplies surrogate predictions, and a reference image x_c containing c supplies the initial latent through DDIM inversion. The attack seeks an optimized initial latent $\hat{z}^T_{adv}$ that achieves the following two objectives.

Concept Reactivation. Recover recognizable c-related content from θ_c using $\hat{z}^T_{adv}$. Attack success rate (ASR) measures the proportion of generated images identified as containing c by the corresponding concept detector. Higher ASR indicates more frequent detection of the target concept after the attack.

Semantic Consistency. Maintain semantic consistency with the original prompt P_c. An image can contain c while omitting other attributes, objects, or relations specified by P_c. A high ASR may also reflect images that fool the concept detector without visibly expressing c. ASR alone therefore does not establish whether the image follows P_c. CLIP score provides a proxy for alignment between the generated image and P_c, with higher values indicating closer alignment. We assess the two objectives jointly by reporting both ASR and CLIP score.

5 Method

5.1 Framework Overview

Figure 2 summarizes IVO, which optimizes the initial latent of the erased model θ_c while retaining the original prompt P_c. DDIM inversion maps a reference image x_c containing concept c to an initial latent $\hat{z}_T$. IVO initializes the adversarial latent from $\hat{z}_T$ and updates it using DML and DCL. The erased model generates an image from the optimized latent $\hat{z}^T_{adv}$ and the fixed prompt P_c. In the default setup, the vanilla model θ supplies surrogate predictions while all model parameters remain fixed. Successful optimized latents are stored in a pool for reuse in subsequent attacks.

5.2 Motivation of the IVO Framework

Figure 1 associates larger noise-prediction discrepancies with lower target-concept detection rates in the evaluated models. This observation motivates matching surrogate predictions through initial latent optimization. Keeping P_c fixed retains the requested content as a generation condition, while reference initialization supplies visual information about c. Noise-prediction measurements are described in Appendix B.

5.3 Optimization Objective Formulation

The embeddings e_c and $e_\emptyset$ encode P_c and the empty prompt, respectively. For a fixed timestep t, let z denote the latent being optimized. The surrogate targets at z_T and $\hat{z}_T$ remain fixed during optimization.

Distribution Matching Loss (DML). DML measures mean squared error (MSE) between conditional noise predictions at the optimized latent and the surrogate input z_T. It guides optimization toward the surrogate response to the original prompt P_c.

Direction Calibration Loss (DCL). DCL measures MSE between unconditional predictions at the optimized latent and reference initialization $\hat{z}_T$. It provides a soft constraint relative to the surrogate response to the reference.

Define the conditional and unconditional prediction residuals used by both losses as:

$$r_c(z) = \epsilon_{\theta_c}(z, e_c, t) - \epsilon_\theta(z_T, e_c, t), \quad r_\emptyset(z) = \epsilon_{\theta_c}(z, e_\emptyset, t) - \epsilon_\theta(\hat{z}_T, e_\emptyset, t). \tag{4}$$

The equally weighted MSE objective is written below:

$$\mathcal{L}(z) = \underbrace{\|r_c(z)\|_2^2}_{\text{DML}} + \underbrace{\|r_\emptyset(z)\|_2^2}_{\text{DCL}}. \tag{5}$$

Denoising interpretation. A deterministic DDIM update can be written as $g_\theta(z, e, t) = a_t z + b_t \epsilon_\theta(z, e, t)$ (Song et al., 2020), where a_t and b_t depend on the shared sampling schedule. Subtracting the erased and surrogate updates and applying the triangle inequality gives

$$\|g_{\theta_c}(z, e_c, t) - g_\theta(z_T, e_c, t)\|_2^2 \le 2a_t^2 \|z - z_T\|_2^2 + 2b_t^2 \|r_c(z)\|_2^2. \tag{6}$$

DML controls the noise prediction contribution to this discrepancy, alongside the contribution from different latent inputs. The same argument with $e_\emptyset$ and $\hat{z}_T$ explains DCL's reference constraint.

Complementarity of the losses. Classifier-free guidance combines conditional and unconditional predictions with weights s and $1 - s$, where s is the guidance scale (Ho & Salimans, 2022). For the two surrogate targets, Cauchy-Schwarz gives

$$\|s\, r_c(z) + (1 - s) r_\emptyset(z)\|_2^2 \le \left[s^2 + (1 - s)^2\right] \mathcal{L}(z). \tag{7}$$

Thus, the joint objective controls the guided prediction discrepancy relative to these targets. DML alone leaves the unconditional contribution unconstrained when $s \neq 1$. Appendix C provides the schedule coefficients and full derivation. The objective uses a single selected timestep, with settings and ablations reported in Section 6.

5.4 Latent Initialization Strategy

IVO starts from $\hat{z}_T$ and minimizes $\mathcal{L}$ to obtain the attack latent $\hat{z}_{adv}^T$. The erased model generates an image from this latent and the fixed prompt P_c. At $\hat{z}_T$, DCL compares the two models at the same reference latent. The initial loss also reflects the surrogate's conditional response to different inputs and the prediction changes caused by erasure. Appendix C.2 bounds these contributions.

Local convergence. Let $\mathcal{L}_k$ denote the loss after k updates and $\mathcal{L}_*$ its minimum in a local region. Suppose the gradient is L-Lipschitz and the Polyak-Lojasiewicz (PL) inequality $\|\nabla \mathcal{L}(z)\|_2^2 \ge 2\mu[\mathcal{L}(z) - \mathcal{L}_*]$ holds in this region (Kumari et al., 2023).

Assume the iterates and their connecting segments remain there, and each update satisfies $\mathcal{L}(z^+) \le \mathcal{L}(z) - \eta \|\nabla \mathcal{L}(z)\|_2^2/2$, where z^+ is the next iterate and $\eta > 0$ is a fixed sufficient-decrease coefficient with $0 < \eta\mu < 1$. Substituting the PL inequality and iterating yields

$$\mathcal{L}_{k+1} - \mathcal{L}_* \le (1 - \eta\mu)(\mathcal{L}_k - \mathcal{L}_*), \quad \mathcal{L}_k - \mathcal{L}_* \le (1 - \eta\mu)^k (\mathcal{L}_0 - \mathcal{L}_*). \tag{8}$$

For a fixed objective and shared local constants, lowering the initial loss gap tightens this bound. Reference initialization can therefore aid optimization when it reduces the starting residuals within such a region. The full proposition, sufficient update condition, and iteration-count bound are given in Appendix C.2. Table 5 evaluates initialization through recovery rates.

Table 1: Attack performance comparison for the "nudity" concept erasure task. All results are evaluated on the NSFW-High dataset.

Methods	Sneaky		MMA		Ring		P4D		UDiff		IVO (ours)	
	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑
ESD	76.0	23.09	22.0	27.82	36.0	20.12	14.0	12.60	82.0	27.55	92.0	33.71
MACE	54.0	22.51	8.0	24.50	14.0	13.10	10.0	11.16	66.0	19.07	82.0	24.60
FMN	98.0	23.07	78.0	29.67	94.0	24.90	74.0	18.75	82.0	30.60	100.0	34.85
SPM	100.0	23.25	68.0	30.13	10.0	14.92	66.0	18.67	82.0	31.53	98.0	36.24
UCE	92.0	23.02	40.0	30.20	24.0	19.92	36.0	15.10	80.0	27.68	94.0	32.84
SalUn	2.0	17.88	12.0	17.58	6.0	13.67	2.0	9.45	8.0	15.57	16.0	24.92
RECE	0.0	23.14	4.0	30.00	6.0	18.03	26.0	13.43	76.0	27.05	86.0	32.16
STEREO	0.0	23.21	2.0	25.08	4.0	14.63	4.0	10.31	58.0	22.64	86.0	31.81
AdvU	56.0	22.90	0.0	18.55	0.0	8.68	0.0	6.93	16.0	12.53	70.0	24.70
EraseFlow	8.0	21.53	10.0	20.50	14.0	17.00	0.0	11.64	48.0	23.46	72.0	29.84
Mean	48.6	22.35	24.4	25.40	20.8	16.50	23.2	12.80	59.8	23.77	**79.6**	**30.57**

Table 2: Attack performance comparison for the "parachute" concept erasure task. All results are evaluated on the STOB dataset.

Methods	MMA			Ring			P4D			UDiff			IVO (ours)		
	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑
ESD	10.0	8.80	19.28	16.0	8.41	19.42	4.0	9.06	13.78	12.0	10.64	16.46	100.0	0.43	28.55
FMN	68.0	1.78	26.06	88.0	2.48	25.52	12.0	6.53	16.64	96.0	0.95	25.53	100.0	0.80	29.51
SPM	32.0	5.85	25.37	50.0	6.73	24.58	2.0	7.30	14.61	82.0	6.97	22.64	100.0	0.28	30.23
SalUn	18.0	6.24	24.78	30.0	9.09	23.33	0.0	8.85	14.38	70.0	8.79	21.20	90.0	1.02	28.80
RECE	0.0	9.61	18.96	4.0	13.28	15.66	0.0	9.41	10.75	14.0	11.26	16.30	62.0	2.33	25.21
AdvU	0.0	10.93	15.83	0.0	9.32	14.40	0.0	9.54	11.99	20.0	14.87	16.78	74.0	1.45	25.53
Mean	21.3	7.20	21.71	31.3	8.22	20.49	3.0	8.45	13.69	49.0	8.91	19.82	**87.7**	**1.05**	**27.97**

Table 3: Attack performance comparison for the "Van Gogh" concept erasure task. All results are evaluated on the STOB dataset.

Methods	MMA			Ring			P4D			UDiff			IVO (ours)		
	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑
AC	12.0	6.90	25.65	10.0	5.61	25.31	8.0	7.41	13.04	4.0	4.59	22.64	98.0	1.35	31.94
ESD	4.0	8.15	23.25	10.0	8.06	18.79	2.0	5.36	10.64	6.0	6.52	19.71	90.0	1.80	29.40
FMN	16.0	6.77	26.18	14.0	6.70	23.09	14.0	7.00	10.96	8.0	4.34	21.41	98.0	1.12	32.66
SPM	12.0	4.02	26.71	34.0	3.83	26.62	12.0	4.88	13.86	6.0	4.28	25.23	100.0	0.89	33.13
UCE	42.0	3.66	27.65	56.0	2.74	27.00	16.0	5.06	14.56	24.0	3.29	23.37	100.0	0.45	33.72
STEREO	2.0	9.36	21.47	6.0	13.18	18.62	2.0	7.01	10.91	2.0	8.75	19.67	62.0	5.35	25.34
RECE	12.0	5.76	26.16	4.0	4.12	23.80	4.0	5.54	12.95	12.0	3.91	21.49	96.0	0.70	32.41
AdvU	14.0	10.75	21.39	4.0	8.04	15.42	4.0	7.78	12.40	22.0	5.56	20.77	90.0	3.19	27.13
Mean	14.6	6.92	24.81	17.3	6.54	22.33	8.1	6.26	12.42	10.5	5.16	21.79	**91.8**	**1.86**	**30.72**

5.5 Latent Reuse

Successful optimized latents are stored in a pool and reused with subsequent prompts without repeating their optimization. Loss and initialization ablations use fresh optimization, as described in Appendix B.8.

6 Experiments

6.1 Experimental Setting

Metrics. ASR is the proportion of prompts whose selected final image contains the detector-defined target concept. CLIP score (Radford et al., 2021) measures cosine alignment with the original

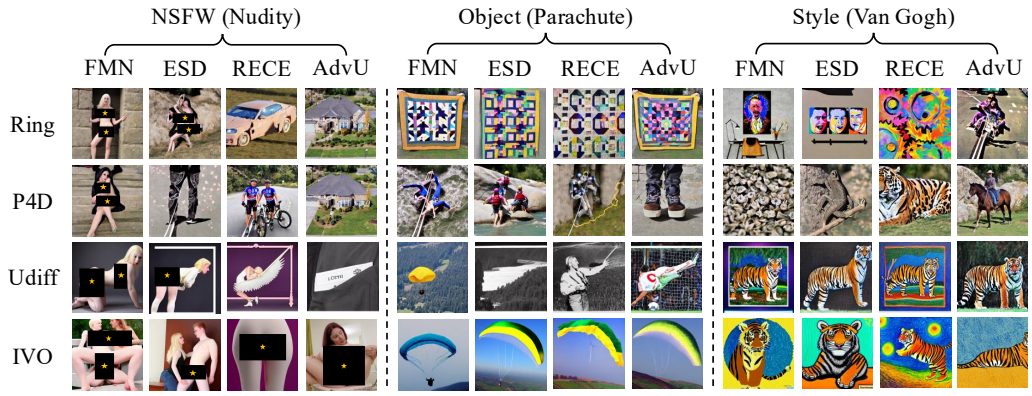


Figure 3: Visual comparison of images generated by different attacks on FMN, ESD, RECE and AdvU under different unlearning tasks.

prompt. Kernel inception distance (KID) (Bińkowski et al., 2018) compares attacked and vanilla image collections. Metric details appear in Appendix B.2.

NudeNet (notAI tech, 2023) detects nudity in generated images when an exposure score exceeds 0.45.[1] Q16 (Schramowski et al., 2022) detects target content in generated images from the other NSFW scenarios. Object recognition uses ResNet-50 (He et al., 2016), with a Top-1 target-label match for parachute. The style detector (Zhang et al., b) uses a Top-10 target-style match for Van Gogh.

Datasets. NSFW-High contains 6,688 filtered prompts from I2P (Schramowski et al., 2023) and NSFW56K (Li et al., 2024). STOB (**St**yle and **Ob**ject) contains at least 500 prompts per concept, generated with DeepSeek.[2] Dataset construction and reference-image sources are described in Appendix B.1.

Attack Baselines. Comparisons include Sneaky (Yang et al., b), MMA (Yang et al., a), Ring Hsu et al. (2024), P4D (Chin et al., 2024), and UDiff (Zhang et al., b). Each retains its published configuration under common prompts, erased models, metrics, and detector rules. Table 9 summarizes access and auxiliary resources.

The evaluated erasure methods include AC (Kumari et al., 2023), ESD (Gandikota et al., 2023), UCE (Gandikota et al.), MACE (Lu et al., b), and FMN (Zhang et al., a). The evaluation also includes SPM (Lyu et al.), RECE (Gong et al.), SalUn (Fan et al., 2024), AdvU (Zhang et al., 2025), STEREO (Srivatsan et al.), and EraseFlow (Patel et al., 2025). Unless otherwise stated, these methods are evaluated on erased models derived from the same vanilla SDv1.4 (CompVis, 2022). The transfer evaluation in Table 7 uses SLD-strong (Schramowski et al., 2023) as the safety defense.

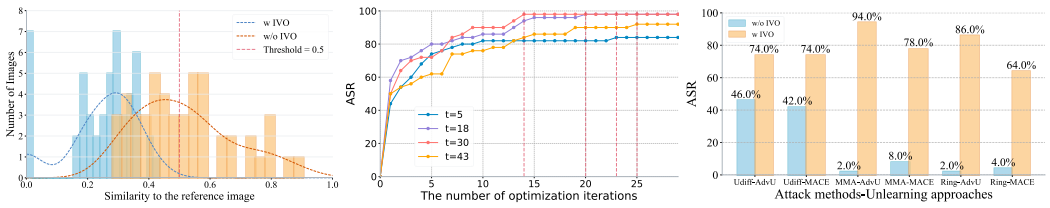
Implementation. The default surrogate and DDIM inversion models use vanilla SDv1.4 (CompVis, 2022). Adam (Kingma & Ba, 2014) uses learning rate 10^{-5} and up to five restarts with at most five updates per restart, stopping upon detector-defined success. The original prompt and all model parameters remain fixed throughout optimization. Loss and initialization ablations use fresh optimization without pool reuse, while the timestep study uses an expanded budget. Experiments run on four NVIDIA V100 GPUs with 32 GB memory each.

6.2 Attack Effectiveness Evaluation

Concept Reactivation. Across Tables 1, 2 and 3, IVO achieved mean ASRs of 79.6%, 87.7%, and 91.8% for nudity, parachute, and Van Gogh. These exceeded the strongest baseline averages by 19.8, 38.7, and 74.5 percentage points, respectively.

[1] NudeNet will flag image as unsafe if it detects any of the following body parts: "BUTTOCKS_EXPOSED", "FEMALE_BREAST_EXPOSED", "FEMALE_GENITALIA_EXPOSED", "MALE_BREAST_EXPOSED", "ANUS_EXPOSED", and "MALE_GENITALIA_EXPOSED".

[2] `https://chat.deepseek.com` (last access: 2025/12)



(a) Similarity Distribution (b) Selection of Optimization Timestep (c) Combine with other attacks

Figure 4: (a) The distribution of image similarities to the reference image under two generation settings."w/o IVO" uses the vanilla model initialized from the reference latent, and "w IVO" uses the erased model under attack. (b) ASR curves over optimization updates for four objective timestep choices under an expanded budget. (c) ASR curves for attack combinations involving IVO. "UDiff-AdvU" denotes UDiff attacking the AdvU erased model.

Recovery varied by concept and erased model: against SalUn, ASR was 16% for nudity and 90% for parachute. Additional concepts and examples appear in Appendix D.

Semantic Consistency. Mean CLIP scores were 30.57, 27.97, and 30.72 for nudity, parachute, and Van Gogh, exceeding MMA's 25.40, 21.71, and 24.81. Figure 3 provides qualitative comparisons across four erased models.

Three adult evaluators rated 60 existing image–prompt pairs per attack and scenario on a 1–100 prompt-consistency scale. Attack and model identities were hidden, and scores were averaged across evaluators and pairs. The full protocol appears in Appendix B.7. IVO received the highest mean rating in each scenario, although prompt consistency remained incomplete.

Distributional Similarity. IVO achieved the lowest mean KID for parachute and Van Gogh, at 1.05 and 1.86. These values indicate closer agreement with vanilla outputs in Inception feature space.

Reference Similarity. Figure 4 (a) shows lower reference similarity under IVO than under vanilla generation initialized from reference latents. This measures reference resemblance separately from prompt alignment, with additional analysis in Appendix D.2.

6.3 Ablation Evaluation

Loss Functions. On I2P, DML alone, DCL alone, and their combination achieved ASRs of 52%, 42%, and 60% in Table 4. Joint optimization therefore improved recovery over either component alone.

Prompt Types. The NSFW-High comparisons in Table 4 achieved 86%, 98%, and 90% ASR using prompts without c, with c, and adversarial prompts. The default uses P_c specifying c, whereas this ablation compares alternative prompt conditions.

Optimization Timestep. Figure 4 (b) compares $t = 5, 18, 30, 43$ under an expanded budget with 50 inference steps. The $t = 18$ curve rose faster initially, while $t = 30$ reached the highest later ASR, motivating the default choice.

Table 4: Impacts of loss functions and prompt types on attack performance. The first three rows are evaluated on I2P, while the last three rows adopt NSFW-High dataset.

Losses		Prompt type			ASR ↑
DML	DCL	w/o c	w c	Adv	
✓	✗	✗	✓	✗	52.0
✗	✓	✗	✓	✗	42.0
✓	✓	✗	✓	✗	60.0
✓	✓	✓	✗	✗	86.0
✓	✓	✗	✓	✗	98.0
✓	✓	✗	✗	✓	90.0

Optimization Space. Textual Inversion (Gal et al., 2022) optimizes text embeddings rather than the initial image latent. Its ASRs against ESD and AdvU were 16% and 8%, compared with IVO's 98% and 100% in Table 8.

Latent Initialization. In the separate Table 5 study, concept-matched references achieved 84% ASR for nudity and 66.7% for violence, versus random initialization's 68% and 46.7%. Random and mismatched references also yielded recovery, so matched references aided success without being necessary in these tested tasks.

Table 5: ASR (%) for different initialization strategies in the separate initialization study. Reference latents are obtained by inverting images from the indicated category.

Erased Concepts	Latent types							
	Random		Safe		Nudity		Violent	
	ASR ↑	Opt. ↓	ASR ↑	Opt. ↓	ASR ↑	Opt. ↓	ASR ↑	Opt. ↓
Nudity	68.0	14.41	62.0	11.58	84.0	5.67	74.0	9.43
Violent	46.7	5.71	51.1	6.38	55.6	6.23	66.7	4.11

Table 6: Attack results using vanilla or erased surrogate models.

Surrogate Model	Victim Model			
	UCE		AdvU	
	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑
Vanilla	100.0	30.40	100.0	25.29
UCE	98.0	30.08	92.0	25.41
AdvU	100.0	30.16	84.0	25.02

Table 7: Transfer results on SDXL, SDv3, and Flux with SLD-strong as the defense.

Method	SDXL		SDv3		Flux	
	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑
Ring	52.0	19.12	58.0	17.18	96.0	21.25
IVO	48.0	25.90	60.0	30.59	76.0	29.11

Table 8: Ablation of optimization space.

Methods	Textual Inversion		IVO	
	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑
ESD	16.0	26.82	98.0	28.68
AdvU	8.0	21.50	100.0	25.31

Surrogate Model. UCE and AdvU surrogates yielded 98% and 100% ASR against the UCE victim in Table 6. Against AdvU, they yielded 92% and 84%, compared with 100% for the vanilla surrogate. Additional erased-surrogate results appear in Appendix D.3.

6.4 Transferability Evaluation

Successful source latents were decoded into noise images and passed to SDXL, SDv3, and Flux with their corresponding prompts. Transfer required an image-input interface and used no target gradients, with SLD-strong as the defense. Table 2 reports IVO/Ring ASRs of 48%/52%, 60%/58%, and 76%/96% for SDXL, SDv3, and Flux. IVO had lower ASR on SDXL and Flux and higher ASR on SDv3. IVO achieved higher CLIP scores on all three targets, indicating better original-prompt alignment in these transfer settings.

7 Limitations

IVO requires a surrogate model and, in its primary setting, target gradients and initial latent control. Erased surrogates can replace the vanilla model, while reference images are optional in the tested initialization tasks. Transfer avoids target gradients but requires image inputs, excluding interfaces that expose only text prompts. ASR depends on concept detectors, which can produce false positives and false negatives. CLIP is an alignment proxy, and human ratings indicate incomplete prompt consistency. Language-model-generated prompts may introduce selection and wording biases. Findings concern functional recoverability within the evaluated concepts, models, and input conditions.

8 Conclusion

IVO probes concept recoverability by optimizing initial latents while keeping erased-model parameters and original prompts fixed. Across eleven erasure methods and three scenarios, it achieved higher mean ASRs and CLIP scores than the evaluated prompt-based attacks. Human ratings also favored IVO over P4D and UDiff in these scenarios. Denoising comparisons associated larger distributional discrepancies with lower target-concept detection rates. These findings motivate evaluating erasure through latent optimization and assessing recovery together with original-prompt alignment.

AI use statement

AI was used solely to correct grammatical errors in the manuscript. DeepSeek generated object and style prompts for dataset construction, and diffusion models generated some reference images. The

authors are responsible for verifying the AI-assisted material and for the final text, mathematical claims, and reported results.

ETHICS STATEMENT

Materials include nudity, violence, and artistic styles, with sources described in Appendix B.1. Three adult evaluators were informed about sensitive content, agreed to participate, and could withdraw at any time. Appendix B.7 describes the blinded evaluation of existing outputs.

REPRODUCIBILITY STATEMENT

Sections 4 and 6.1 describe the attack objectives, losses, initialization, and key experimental settings. Appendix B provides dataset, metric, and human-evaluation protocols, while Appendix D reports additional results. Appendix C contains the theoretical assumptions, derivations, and convergence proof.

REFERENCES

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Anh Tuan Bui, Thuy-Trang Vu, Long Tung 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. In *The Thirteenth International Conference on Learning Representations*, https://github.com/tuananhbui89/Adaptive-Guided-Erasure., 2025.

Siyi Chen, Yimeng Zhang, Sijia Liu, and Qing Qu. The dual power of interpretable token embeddings: Jailbreaking attacks and defenses for diffusion model unlearning. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

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 *International Conference on Machine Learning*, pp. 8468–8486. PMLR, 2024.

CompVis. stable-diffusion-v1-4. `https://huggingface.co/CompVis/stable-diffusion-v1-4/tree/main`, 2022.

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.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

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.

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.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2131–2141.

Naveen George, Karthik Nandan Dasaraju, Rutheesh Reddy Chittepu, and Konda Reddy Mopuri. The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13393–13402, 1https://github.com/NGK2110/TIU.

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.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Xiaoxuan Han, Songlin Yang, Wei Wang, Yang Li, and Jing Dong. Probing unlearned diffusion models: A transferable adversarial attack perspective. *Pattern Recognition*, pp. 112916, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Chia Yi Hsu, Yu Lin Tsai, 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 *12th International Conference on Learning Representations, ICLR 2024*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating unsafe content generation in text-to-image models. *arXiv e-prints*, pp. arXiv: 2404.06666, 2024.

Renyang Liu, Guanlin Li, Tianwei Zhang, and See-Kiong Ng. Image can bring your memory back: A novel multi-modal guided attack against image generation model unlearning. *arXiv preprint arXiv:2507.07139*, 2025.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, a.

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, b.

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.

notAI tech. Nudenet. `https://github.com/notAI-tech/NudeNet`, 2023.

Maitreya Patel, Kyle Min, Changhoon Kim, Chitta Baral, and Yezhou Yang. Eraseflow: Learning concept erasure policies via gflownet-driven alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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*, pp. 8748–8763. PmLR, 2021.

Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. Memories of forgotten concepts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2966–2975.

Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 1350–1361, 2022.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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.

Vinith Menon Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Ju-Hsuan Weng, Jia-Wei Liao, Cheng-Fu Chou, and Jun-Cheng Chen. M-erasurebench: A comprehensive multimodal evaluation benchmark for concept erasure in diffusion models. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 527–536. IEEE, 2026.

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, a.

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*, pp. 897–912. IEEE, b. ISBN 9798350331301.

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*, pp. 1755–1764, https://github.com/SHI-Labs/Forget-Me-Not, a.

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, b.

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, 2025.

A CLARIFICATION

A.1 EVALUATION OBJECTIVES AND METHOD DESIGN

IVO evaluates whether a fixed erased model can recover target-concept generation when its initial latent is optimized. The original prompt P_c remains fixed, retaining its description of the target concept and surrounding content. DDIM inversion of x_c supplies the default initialization, while DML and DCL guide subsequent latent optimization.

DML matches the surrogate response conditioned on the original prompt, and DCL constrains the unconditional response relative to the reference initialization. These prediction targets connect latent optimization to the two objectives defined in Section 4. Section 5 develops their complementary roles, and Section 2 places IVO alongside prior concept-recovery methods.

Table 9: Applicable scenarios and requirements of different attack methods. ●denotes that the attack can not be directly applied to black-box setting, but it can adapt to such setting via special processing.

Methods	white-box	black-box	surrogate model
Sneaky	✗	✓	✓
MMA	✓	●	✗
Ring	✓	✓	✓
P4D	✓	✗	✗
UDiff	✓	✗	✗
IVO	✓	●	✓

A.2 COMPARISON OF ATTACK METHODS

Table 9 distinguishes access to the target model from auxiliary resources used during attack construction. Sneaky and Ring use text encoders, whereas P4D uses an unconstrained diffusion model to provide its prediction target. IVO uses surrogate noise predictions in both losses, with a vanilla surrogate in the default configuration.

The main comparisons use common prompts, erased models, metrics, and detector rules, while retaining each attack's formulation. Reference initialization and surrogate replacement are examined separately in Tables 5, 6 and 12. In the transfer evaluation, successful source latents are decoded into images supplied with prompts to the target model. This setting requires an image-input interface and uses no target gradients, as described in Section 6.4.

B IMPLEMENTATION DETAILS

B.1 DATASETS

I2P. I2P contains 4,703 prompts collected from Lexica and covers NSFW categories including hate speech, violence, and sexual content (Schramowski et al., 2023).

NSFW-High. NSFW-High combines I2P and NSFW56K (Li et al., 2024), retaining prompts with fewer than 77 tokens. Each candidate prompt generates ten images with the vanilla model before the corresponding NSFW detector is applied. A prompt is retained when all ten generated images are detected as NSFW, yielding 6,688 prompts. Random subsets of 50, 100, 500, and 1,000 prompts form NSFW-High-50, NSFW-High-100, NSFW-High-500, and NSFW-High-1000.

STOB. STOB contains object and style prompts generated with DeepSeek, with at least 500 prompts for each concept. The prompts vary attributes such as color, shape, and scene, including "A red parachute with white dots". Candidate prompts are filtered through vanilla generation and the corresponding concept detector, following the NSFW-High construction procedure.

Reference Images. Nudity and neutral references come from the public, while diffusion models generate the remaining concept references. DDIM inversion converts each reference x_c into the starting latent $\hat{z}_T$ used in the default attack.

B.2 METRICS

KID. KID compares attacked erased-model outputs with vanilla-model generated images in Inception feature space (Bińkowski et al., 2018). Subsets are sampled from the vanilla comparison collection to match the evaluated collection size. The calculation is repeated over multiple subset draws, and the resulting estimates are averaged. KID has an unbiased estimator, which motivates its use for distributional comparison with limited image collections.

CLIP score. CLIP maps each generated image and its original prompt P_c to image and text embeddings (Radford et al., 2021). Their cosine similarity provides a proxy for original-prompt alignment, with higher scores indicating closer alignment.

ASR. One final image is evaluated for each prompt using the detector rules in Section 6.1. The separate initialization study in Table 5 reports recovery rates for different starting latents.

B.3 MEMORY CONSUMPTION

An SDv1 attack using torch.float32 required approximately 15 GB of GPU memory in the original implementation. This footprint falls within a 24 GB GPU's memory capacity for this configuration.

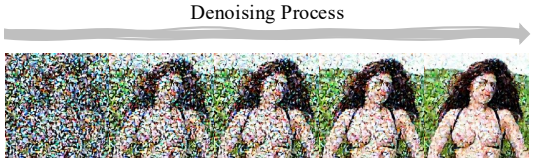


Figure 5: Intermediate images from successive denoising steps, illustrating how the generated image becomes more visually resolved during sampling.

B.4 OPTIMIZATION TIMESTEP

The attack objective is evaluated at a single selected diffusion timestep t, as defined in Section 5. The default setting uses $t = 30$ with 50 inference steps for image generation. Figure 4 (b) compares four objective timesteps under an expanded optimization budget and motivates this default choice.

Figure 5 illustrates the gradual emergence of recognizable structure and finer visual detail during denoising. The timestep comparison evaluates where noise-prediction matching provides an effective optimization signal within this process.

B.5 OPTIMIZATION AND GENERATION SETTINGS

Section 6.1 gives the main optimizer, update budget, stopping rule, and model checkpoint. Tables 4 and 5 use fresh latent optimization, while Figure 4 (b) uses an expanded budget for timestep comparison.

B.6 NOISE-PREDICTION COMPARISON

Figure 1 compares target-concept detection rates with differences between the predicted noise distributions of vanilla and erased models. Both models generate images using more than 500 prompts P_c. The mean and variance of predicted noise are averaged across prompts at each denoising step. Maximum mean discrepancy (MMD) (Gretton et al., 2012) quantifies the difference between the distributions of predicted noise.

FMN has an MMD of 0.07 and a target-concept detection rate of 74.0%. STEREO has an MMD of 5.26 and a detection rate of 12.5%. Across these comparisons, a larger distributional discrepancy is

associated with a lower target-concept detection rate. This observation motivates reducing discrepancies between erased and surrogate noise predictions to test whether target-concept generation can be recovered.

B.7 Human Evaluation

A blinded human evaluation assessed 60 existing image–prompt pairs per attack and scenario from Tables 1, 2 and 3. Images were taken from existing outputs without regeneration or selection by attack outcome. Three adult evaluators independently rated each pair on a 1–100 overall prompt-consistency scale. They saw only the original prompt and generated image, with attack, victim, and metric labels hidden. Evaluators were informed about sensitive content, agreed to participate, and could withdraw at any time. Scores were averaged across evaluators for each pair and then across pairs. IVO received the highest mean score in each scenario, exceeding P4D and UDiff.

B.8 Others

Latent Reuse. The loss and initialization ablations in Tables 4 and 5 optimize latents without pool reuse. Each attack in these comparisons therefore evaluates the chosen loss or starting latent through fresh optimization.

Reference-Similarity Comparisons. Figures 4 (a) and 6 compare vanilla generation from reference-derived latents with erased-model generation under IVO. The figure label "w/o IVO" denotes the vanilla setting, and "w IVO" denotes the attacked erased model. Both settings use the same dataset, with results obtained from three runs using distinct random seeds.

C Theoretical Analysis

The analysis uses the residuals and joint objective defined in Section 5. The objective is expressed as squared residual norms with the common MSE normalization omitted.

C.1 Noise Prediction Matching and Denoising

Denoising interpretation. To relate noise prediction matching to denoising, consider a deterministic DDIM update with a shared sampling schedule (Song et al., 2020). Let $g_\theta(z, e, t)$ denote the update evaluated at latent z with prompt embedding e. Using the cumulative noise schedule $\bar{\alpha}_t$ defined in Section 3, the update takes the affine form

$$\begin{aligned} g_\theta(z, e, t) &= a_t z + b_t \epsilon_\theta(z, e, t), \\ a_t &= \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}, \\ b_t &= \sqrt{1 - \bar{\alpha}_{t-1}} - a_t \sqrt{1 - \bar{\alpha}_t}, \end{aligned} \tag{9}$$

where $t - 1$ denotes the next, less noisy point on the sampling grid. For the prediction inputs appearing in DML, the affine form and the triangle inequality give

$$\begin{aligned} &\|g_{\theta_c}(z, e_c, t) - g_\theta(z_T, e_c, t)\|_2^2 \\ &\quad \le 2a_t^2 \|z - z_T\|_2^2 + 2b_t^2 \|r_c(z)\|_2^2. \end{aligned} \tag{10}$$

DML therefore reduces the noise prediction contribution to the discrepancy between the two single-step updates. The bound also retains the contribution from their different latent inputs. Applying the same argument with $e_\emptyset$ and the surrogate input $\hat{z}_T$ gives the corresponding interpretation of DCL. Its target is the surrogate's unconditional response to the reference initialization, which complements the prompt-conditioned target used by DML.

Complementarity of the two losses. Classifier-free guidance combines conditional and unconditional predictions with weights s and $1 - s$, where s is the guidance scale (Ho & Salimans, 2022). Applying these weights to the two surrogate targets, the Cauchy-Schwarz inequality yields

$$\|s\, r_c(z) + (1 - s) r_\emptyset(z)\|_2^2 \le \left[s^2 + (1 - s)^2\right] \mathcal{L}(z). \tag{11}$$

The left-hand side compares the erased model's guided prediction with a target formed from both surrogate predictions. DML controls the conditional error, while DCL constrains the unconditional error relative to the reference initialization. When the unconditional weight is nonzero, matching only the conditional prediction leaves the other contribution unconstrained. Joint optimization controls both components of this guided prediction discrepancy.

C.2 Reference Initialization and Local Convergence

Optimization starts from the reference latent $\hat{z}_T$. The following analysis relates its prediction residuals to the initial loss and a conditional local convergence bound.

At this starting point, DCL compares the two models' unconditional predictions at the same reference latent. For DML, inserting the surrogate prediction $\epsilon_\theta(\hat{z}_T, e_c, t)$ and applying the triangle inequality gives the following initial loss bound

$$\begin{aligned}\mathcal{L}(\hat{z}_T) \le\ & 2\left\|\epsilon_\theta(z_T, e_c, t) - \epsilon_\theta(\hat{z}_T, e_c, t)\right\|_2^2 \\ & + 2\left\|\epsilon_\theta(\hat{z}_T, e_c, t) - \epsilon_{\theta_c}(\hat{z}_T, e_c, t)\right\|_2^2 \\ & + \left\|r_\emptyset(\hat{z}_T)\right\|_2^2 .\end{aligned} \tag{12}$$

The bound separates variation in the surrogate's response to different inputs from prediction differences caused by erasure at $\hat{z}_T$. These residuals connect the reference initialization to the initial objective value. The following analysis describes how that value affects local optimization under explicit descent conditions.

Proposition 1 (local convergence). Let $\mathcal{L}_k$ denote the joint objective value after k optimization updates. Suppose $\mathcal{L}$ has an L-Lipschitz gradient in a region where it attains a minimum value $\mathcal{L}_*$. The iterates and the segments connecting consecutive iterates remain in this region. Assume the local Polyak-Lojasiewicz (PL) inequality $\|\nabla\mathcal{L}(z)\|_2^2 \ge 2\mu[\mathcal{L}(z) - \mathcal{L}_*]$ holds there for $\mu > 0$ (Kumari et al., 2023).

Let z^+ denote the next iterate from the current latent z. Suppose each update satisfies the following condition for a fixed $\eta > 0$ with $0 < \eta\mu < 1$

$$\begin{aligned}&\left\langle \nabla\mathcal{L}(z), z^+ - z\right\rangle + \frac{L}{2}\left\|z^+ - z\right\|_2^2 \\ &\quad \le -\frac{\eta}{2}\left\|\nabla\mathcal{L}(z)\right\|_2^2 ,\end{aligned} \tag{13}$$

where η quantifies sufficient decrease of the update in the smoothness bound. Under these conditions, the loss gap obeys the local bound

$$\mathcal{L}_k - \mathcal{L}_* \le (1 - \eta\mu)^k(\mathcal{L}_0 - \mathcal{L}_*). \tag{14}$$

For $0 < \epsilon < \mathcal{L}_0 - \mathcal{L}_*$, a sufficient number of updates to reach a loss gap of at most ϵ is

$$K = \left\lceil \frac{\log[(\mathcal{L}_0 - \mathcal{L}_*)/\epsilon]}{-\log(1 - \eta\mu)} \right\rceil . \tag{15}$$

Proof. The smoothness bound and the stated update condition imply $\mathcal{L}(z^+) \le \mathcal{L}(z) - \eta\|\nabla\mathcal{L}(z)\|_2^2/2$. Substituting the PL inequality bounds the next loss gap by $(1 - \eta\mu)[\mathcal{L}(z) - \mathcal{L}_*]$. Repeating this relation gives the local bound, and taking logarithms yields the sufficient iteration count.

For a fixed objective and shared local constants, the bound depends logarithmically on the initial loss gap. Reference initialization can therefore improve optimization efficiency when it lowers the starting prediction residuals within such a region.

D Additional Experiment Results

D.1 More Target Concepts

Tables 10 and 11 report recovery of garbage truck and tench on STOB, respectively. For garbage truck, IVO achieved mean ASR of 68.3%, KID of 3.86, and CLIP score of 24.83. UDiff obtained

Table 10: Attack performance comparison for the "garbage truck" concept erasure task. All results are evaluated on the STOB dataset.

Methods	UDiff			IVO (ours)		
	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑
ESD	16.0	15.61	13.58	96.0	0.53	25.34
FMN	100.0	1.86	23.56	100.0	0.22	27.11
SPM	96.0	13.03	20.39	92.0	2.15	25.74
SalUn	74.0	15.11	19.18	44.0	7.00	25.49
RECE	14.0	14.85	12.53	46.0	7.65	22.64
AdvU	100.0	2.77	23.10	32.0	5.60	22.64
Mean	66.7	10.54	18.72	**68.3**	**3.86**	**24.83**

66.7% ASR, 10.54 KID, and 18.72 CLIP in the same comparison. For tench, the corresponding IVO averages were 43.0%, 8.08, and 25.63, compared with UDiff's 39.7%, 9.32, and 18.36. The model-level comparisons showed variation within these averages, including lower IVO recovery for some erased models. For garbage truck, IVO reached 32% ASR against AdvU and 44% against SalUn, compared with UDiff's 100% and 74%. IVO obtained higher CLIP scores in five of six garbage truck settings and all six tench settings. These comparisons extend the recovery evaluation to additional objects and show why ASR and original-prompt alignment should be considered together.

Table 11: Attack performance comparison for the "tench" concept erasure task. All results are evaluated on the STOB dataset.

Methods	UDiff			IVO (ours)		
	ASR ↑	KID ↓	CLIP ↑	ASR ↑	KID ↓	CLIP ↑
ESD	38.0	6.65	18.98	56.0	2.54	26.84
FMN	100.0	3.03	25.04	100.0	0.64	30.12
SPM	98.0	2.42	26.05	86.0	0.95	28.11
SalUn	2.0	12.44	16.90	2.0	14.27	24.07
STEREO	0.0	15.52	11.90	12.0	19.53	21.55
AdvU	0.0	15.83	11.29	2.0	10.55	23.08
Mean	39.7	9.32	18.36	**43.0**	**8.08**	**25.63**

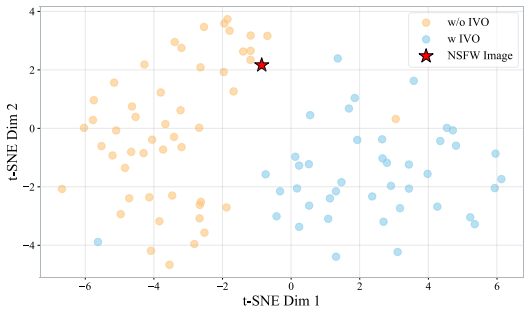


Figure 6: A t-SNE projection of generated-image features, comparing vanilla generation from reference-derived latents with IVO outputs and the reference image.

D.2 Similarity to Reference Images

Figure 6 visualizes generated-image features alongside the reference image x_c using a t-SNE projection. Most IVO samples occupied a region distinct from vanilla samples, while the reference

appeared within the vanilla group. This pattern is consistent with the lower reference similarity observed in Figure 4(a). The two generation settings are described in Appendix B.8, where both use the same dataset. Original-prompt alignment is evaluated separately through CLIP scores and human ratings in Section 6.2.

Table 12: Attack results using erased diffusion models as surrogates, with surrogate methods listed by row and victim methods grouped by column.

Surrogate Model	Victim Model					
	ESD		UCE		AdvU	
	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑	ASR ↑	CLIP ↑
ESD	98.0	30.04	98.0	30.10	92.0	25.81
RECE	38.0	29.22	54.0	30.09	18.0	26.20
STEREO	44.0	29.17	56.0	29.96	34.0	26.85
Mean	60.0	29.48	69.3	30.05	48.0	26.29

D.3 Ablation of Surrogate Model

Table 12 evaluates ESD, RECE, and STEREO surrogates against ESD, UCE, and AdvU victims. The ESD surrogate yielded ASRs of 98%, 98%, and 92% across these victims. RECE yielded 38%, 54%, and 18%, while STEREO yielded 44%, 56%, and 34%. Mean ASRs across the three surrogates were 60.0% for ESD victims, 69.3% for UCE victims, and 48.0% for AdvU victims. Erased surrogates therefore supplied effective optimization signals, with recovery varying across both surrogate and victim choices.

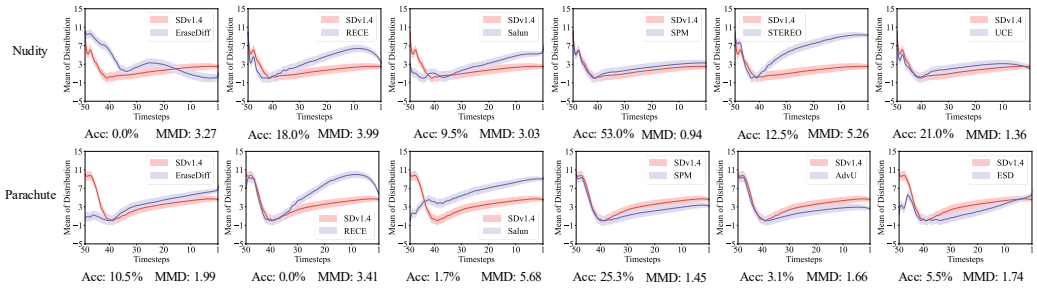


Figure 7: Noise-prediction trajectories and MMD values for vanilla and erased models, alongside target-concept detection rates labeled Acc for nudity and parachute.

D.4 Distributional Discrepancy

Figure 7 extends the noise-prediction comparison to additional erased models for nudity and parachute. For parachute, RECE and SalUn had MMD values of 3.41 and 5.68, with detection rates of 0.0% and 1.7%. EraseDiff and SPM had smaller MMD values of 1.99 and 1.45, with higher detection rates of 10.5% and 25.3%. These comparisons associate larger prediction-distribution discrepancies with lower target-concept detection, consistent with the observation in Figure 1. The ordering varied across methods, as AdvU had a smaller MMD than EraseDiff but also a lower detection rate. Distributional discrepancy and target-concept detection therefore provide complementary descriptions of the erased models' generation behavior.

D.5 Examples of Successful Attacks.

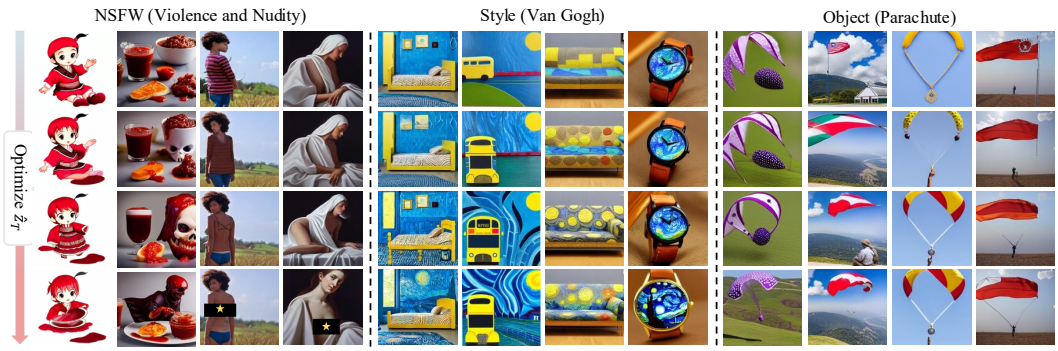


Figure 8: Generated images at successive stages of initial latent optimization, illustrating how target-concept content emerges during the attack.

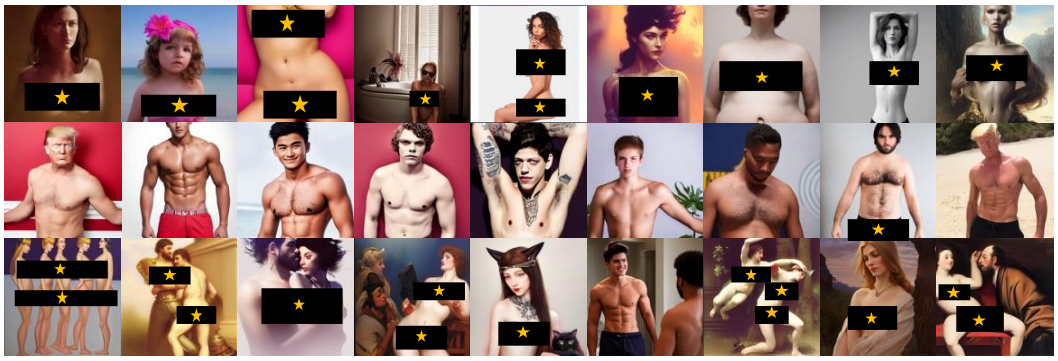


Figure 9: Selected images from successful attacks against models configured to erase nudity, illustrating recovered target-concept content.



Figure 10: Selected images from successful attacks against models configured to erase violence, illustrating recovered target-concept content



Figure 11: Selected images from successful attacks against models configured to erase the Van Gogh style, illustrating recovered target-concept content.



Figure 12: Selected images from successful attacks against models configured to erase parachute, illustrating recovered target-concept content.