CONCEPT ERASURE FOR TEXT-TO-IMAGE DIFFUSION MODELS AGAINST IMAGE PROBES

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept erasure aims to prevent text-to-image (T2I) diffusion models from generating unwanted concepts. Although T2I models can naturally accept images as input, existing erasure methods have limited their probing of erased models to only the text modality, e.g., "Generate a picture of a naked woman". This makes existing erasures insufficient against image probes. In this paper, we address this gap by proposing **I**mage **P**robe **Erase** (**IPErase**), the first concept erasure method specifically designed to resist image probes. In contrast to existing erasures, which are optimized with text-dominated supervision, our IPErase only involves image supervision. Specifically, IPErase uses a frozen teacher (U-Net or DiT) to construct denoising trajectories that extrapolate beyond the neutral prediction and away from the concept. It then trains a student model to match these trajectories using concept images as conditioning inputs through a frozen image prompt adapter, selectively updating the shared denoising backbone. On two widely used target concepts (nudity and Van Gogh style), our IPErase outperforms existing (text-supervised) erasure methods by (relative) 95.9%, 25.0%, and 76.9% under three increasingly challenging forms of image probes: black-box input conditioning, gray-box latent initialization, and white-box adversarial optimization. Further analyses show that despite relying solely on image supervision, our IPErase performs perfectly against diverse text probes, comparable to text-dominated methods. This strong contrast in erasure performance between text-dominated methods against image probes and image-supervised methods against text probes suggests that images, which contain richer semantics, position target concepts much better than text.

1 INTRODUCTION

Concept erasure can prevent pretrained text-to-image (T2I) diffusion models from generating unwanted concepts by optimizing the model through fine-tuning, closed-form parameter editing, or lightweight adapters to suppress unwanted concepts while preserving unrelated generation (Gandikota et al., 2023; 2024; Zhang et al., 2023; Gong et al., 2024). Existing methods typically define unwanted concepts at the text level, through keywords, prompts, or learned text embeddings. Robust extensions address textual reactivation through adversarial training and retention-aware optimization (Zhang et al., 2024a; Srivatsan et al., 2025). Recent image-related concept erasure methods address only some image-probe routes. Co-Erasing (Li et al., 2025) and TICoE (Li et al., 2026) combine textual and visual supervision. IP-Era (Tang et al., 2025) constructs paired images but retains explicit concept-text conditioning during pair construction and teacher–student alignment. ACE (Wang et al., 2025) improves concept suppression during image editing but also relies on a text-conditioned erasure objective. CASteer (Gaintseva et al., 2026) and IRECE (Weng et al., 2026) instead apply inference-time interventions, making their protection dependent on additional inference procedures. [1]

However, such erasure strategies remain ineffective against image probes. To address this gap, we propose IPErase, the first erasure method against three image probes that learns a suppression signal directly from concept-neutral image pairs without using target names, image captions, or adversarial text prompts. Although image probes use different input routes, they share a denoising backbone, motivating a persistent edit in its parameters rather than a particular conditioning module. At shared latent states under empty-text conditioning, a frozen teacher contrasts the denoising trajectories

[1] Detailed related work is provided in Appendix A.

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.

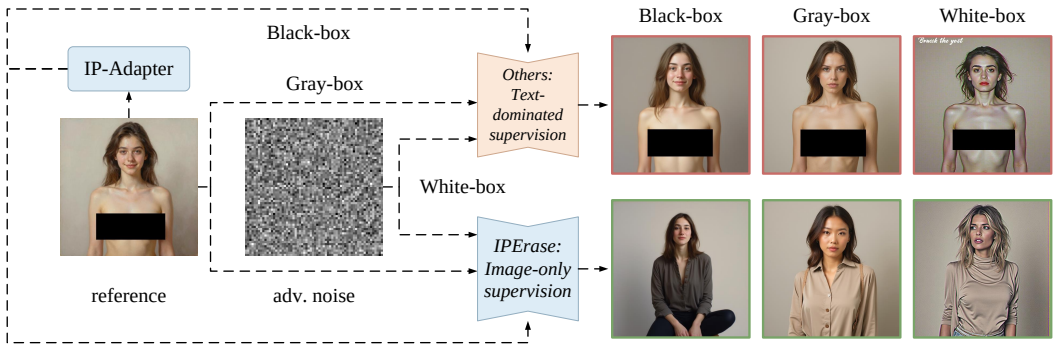


Figure 1: **Existing text-dominated erasure methods (top) vs. our image-only IPErase (bottom) under three types of image probes:** black-box input conditioning, gray-box latent initialization, and white-box adversarial optimization, e.g., RECALL Liu et al. (2026). IPErase succeeds in all cases.

conditioned on the concept and neutral images to construct targets that extrapolate beyond the neutral prediction and away from the target concept. A student fits these fixed trajectories by updating only self-attention and feed-forward linear maps within the Transformer blocks of the U-Net. The image encoder and IP-Adapter remain frozen, ensuring that the persistent edit resides in the shared denoising backbone.

Input conditioning through IP-Adapter (Ye et al., 2023), native img2img via latent initialization (Meng et al., 2022), and adversarial optimization RECALL (Liu et al., 2026) provide three distinct routes for visual provision. In our evaluation, IP-Adapter and img2img are fixed image probes without adversarial optimization, whereas RECALL is an adversarial image probe. We distinguish their access assumptions: black-box access to an image-conditioning interface, gray-box control over latent initialization, and white-box access for adversarial optimization, respectively. Figure 1 illustrates the three image-probe routes and example nudity-erasure outputs; IPErase produces clothed subjects in the shown cases.

Our systematic evaluation on SD1.4 finds Concept Detection Rate (CDR) as high as 94.00% among seven concept-erasure baselines across six primary empty-text endpoints spanning input conditioning, img2img, and their composition (Table 1). These results show that erased concepts can reappear through visual input interfaces and that robustness against textual reactivation does not guarantee robustness under image probes. Our method achieves the lowest CDRs on five of six primary fixed image-probe endpoints among the evaluated defenses (Section 4). An exploratory adaptation to FLUX.1-dev further evaluates the paired objective on a DiT backbone using separately trained models (Appendix F). Beyond visual inputs, the same nudity checkpoint yields 0.7–8.7% CDR against four evaluated adversarial text-probe protocols without additional training (Section 4.3).

Our key contributions are as follows:

- We identify concept provision through image probes as a distinct challenge for concept erasure. Across seven existing erasure methods, visual provision through IP-Adapter conditioning, img2img initialization, and adversarial RECALL reaches CDR of up to 94.00%, revealing a substantial gap between textual and visual robustness.
- We propose IPErase, which derives erasure supervision directly from concept-neutral image pairs and writes the resulting suppression into the shared denoising backbone. This design requires no target names, captions, or adversarial text prompts during erasure optimization and produces a persistent edit that operates across visual input interfaces.
- We establish the effectiveness and transferability of image-defined erasure. IPErase achieves the lowest CDRs on five of six primary fixed image-probe endpoints, remains effective against adversarial RECALL, and transfers to four adversarial text-probe protocols without additional training. An exploratory FLUX.1-dev study further evaluates the approach on a DiT backbone.

2 PRELIMINARIES OF CONCEPT ERASURE

Problem formulation. Given a pretrained diffusion model and an unwanted concept C, concept erasure modifies the model to suppress C while preserving unrelated generation. A *text probe* supplies a direct target prompt or searches for an adversarial prompt that recovers the erased concept C (Chin et al., 2024; Zhang et al., 2024b).

Image probes. *Black-box input conditioning via IP-Adapter.* IP-Adapter (Ye et al., 2023) encodes a reference image and projects its features into tokens that enter the denoiser through image cross-attention, alongside the text-conditioning path. Our REF-IP probe starts from Gaussian noise and supplies a concept-bearing reference through this interface. This black-box conditioning probe requires no model gradients.

Gray-box latent initialization via img2img. Img2img VAE-encodes a source image, adds noise at a selected timestep, and denoises the resulting latent (Meng et al., 2022). Lower img2img strength preserves more source information by applying less corruption and fewer denoising steps. Our IMG probe uses this route with IP-Adapter disabled; the composed IMG+IP route also supplies the source as an IP-Adapter reference. We categorize control over latent initialization as a gray-box probe, requiring no gradient-based search.

White-box adversarial optimization via RECALL. Our RECALL (Liu et al., 2026) probe uses model gradients to optimize visual inputs for concept recovery.

Our evaluation fixes target-explicit text and model weights while optimizing img2img initialization with IP-Adapter disabled. Detection in any initial or subsequent candidate counts as success. The fixed image probes, i.e., REF-IP and IMG, use empty text in the primary evaluation, whereas RECALL probe applies target text for a multi-modal stress test.

3 IPERASE: IMAGE-PROBE ERASE

3.1 MOTIVATION

In order to be effective for image probes, we seek to derive concept-erasure supervision from semantically rich visual differences rather than the abstract textual names of target concepts. To motivate this choice, we compare the teacher's responses at identical z_t and t. Specifically, we compute the image-conditioned difference $\Delta\epsilon_{\text{img}} = \epsilon(z_t, t; I_e) - \epsilon(z_t, t; I_k)$ and the text-conditioned difference $\Delta\epsilon_{\text{text}} = \epsilon(z_t, t; \text{“nudity”}) - \epsilon(z_t, t; \varnothing)$. The image contrast uses a frozen IP-Adapter with empty text, whereas the text contrast uses no IP-Adapter; $\varnothing$ denotes an empty prompt.

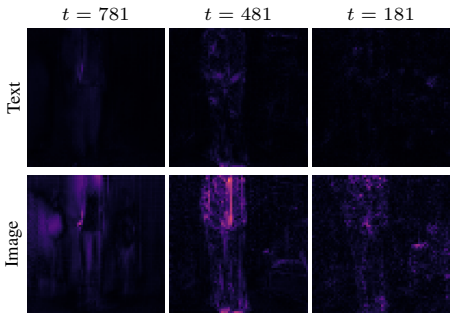


Figure 2: Image (bottom) provides consistently stronger erasure supervision than text (top) across the time states..

Fig. 2 visualizes the four-channel RMS magnitude of each difference, with brighter regions indicating larger prediction changes. Image-pair differences already exhibit person-related structure at the early sampled state ($t = 781$), with stronger local responses than the text contrast in some examples. These structures become more pronounced around the upper body and clothing boundaries at $t = 481$, while both contrasts generally weaken at $t = 181$. These observations motivate using $\Delta\epsilon_{\text{img}}$ to construct erasure supervision: paired images provide spatially structured signals reflecting concrete appearance changes, without requiring a textual description of the target concept.

3.2 IPERASE DETAILS

Motivated by the spatially structured image-pair differences above, we design IPErase to suppress target concepts presented through image probes. As shown in Figure 3, the method separates supervision construction from student training. In Stage I, a frozen teacher converts paired visual

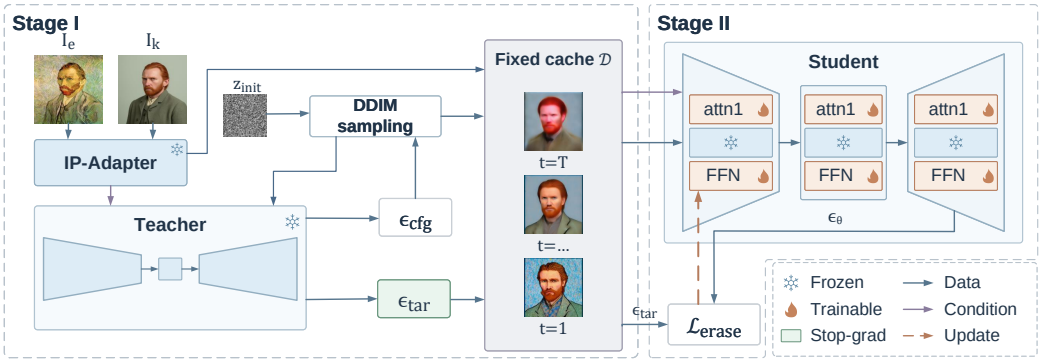


Figure 3: **IPErase training process. Stage I:** Frozen teacher constructs a fixed cache of states, concept-image conditions, and erasure targets through DDIM trajectories. **Stage II:** The student fits stop-grad targets from the fixed cache by updating shared self-attention and FFN linear maps.

contrasts into a fixed cache of denoising examples and erasure targets. In Stage II, a student learns from this cache by fitting the targets under concept-image conditions and updating its shared denoising backbone.

Stage I: constructing the fixed cache. Each record contains a denoising state, a concept-image condition, and an erasure target. We construct these records by comparing the frozen teacher's predictions under concept and neutral image conditions at the same sampled states.

We use image pairs (I_i^e, I_i^k), where I_i^e contains the concept to erase and I_i^k is a neutral counterpart for the concept to keep, sharing similar subject matter and composition. For Van Gogh erasure, a painting is paired with a photographic neutral; for nudity erasure, the neutral depicts the adult subject in opaque clothing. Matching the pair's non-target content helps focus the teacher's prediction difference on the unwanted attribute (Appendix C).

A frozen IP-Adapter (Ye et al., 2023) lets the teacher use visual details supplied by these images, including attributes that a concept name leaves unspecified or that are difficult to describe precisely in text. Its frozen CLIP image encoder gives $h_i^e = E(I_i^e)$ and $h_i^k = E(I_i^k)$. We write the frozen teacher as $\epsilon_\phi(s, h)$, with denoising state $s = (z_t, t, c_\emptyset)$. All branches use the empty text condition $c_\emptyset$; the unconditional image branch uses $h_0 = 0$ before image projection.

To collect states from concept-guided generation, the teacher samples one DDIM trajectory (Song et al., 2021) per pair from $z_{\text{init}} \sim \mathcal{N}(0, I)$, using an independent fixed seed. Let n index the collected states across all pairs, with corresponding embeddings h_n^e, h_n^k. At each state s_n, it predicts noise under the unconditional, concept, and neutral image conditions:

$$\epsilon_{0,n} = \epsilon_\phi(s_n, h_0), \quad \epsilon_{e,n} = \epsilon_\phi(s_n, h_n^e), \quad \epsilon_{k,n} = \epsilon_\phi(s_n, h_n^k). \tag{1}$$

All three predictions share z_t, t, and empty text, so their differences reflect changes in image conditioning. The trajectory advances using image classifier-free guidance (Ho & Salimans, 2022),

$$\epsilon_{\text{cfg},n} = \epsilon_{0,n} + w(\epsilon_{e,n} - \epsilon_{0,n}), \tag{2}$$

where w controls image-guidance strength. The neutral prediction is evaluated on this concept-guided trajectory and does not advance it. No reconstruction of either pair member is required.

The concept/neutral difference defines the erasure target

$$\epsilon_{\text{tar},n} = \epsilon_{k,n} - \lambda(\epsilon_{e,n} - \epsilon_{k,n}), \tag{3}$$

where $\lambda > 0$ controls erasure strength. The neutral prediction provides a reference for retaining the pair's shared content, and the subtraction moves the target further away from the concept-conditioned prediction. At $\lambda = 0$ the target reduces to neutral imitation; positive λ extrapolates beyond it in noise-prediction space. This target supervises the student, while only $\epsilon_{\text{cfg},n}$ drives the teacher's DDIM updates.

Collecting these records gives $\mathcal{D} = \{(s_n, h_n^e, \epsilon_{\text{tar},n})\}_{n=1}^N$, with $N = MK$ for M image pairs and K DDIM steps per pair. All states, conditions, and targets remain fixed during student training. The student reuses this supervision without rerunning the teacher. Our ablations further show that the fixed cache can be shuffled (Section 4.5).

Stage II: training the student. The student starts from the teacher's pretrained backbone and predicts $\epsilon_\theta(s_n, h_n^e)$ from the cached state and *concept-image condition*. We optimize its shared self-attention and feed-forward weights, denoted by θ, while keeping all remaining weights fixed. The training objective is

$$\min_\theta \widehat{\mathcal{L}}(\theta), \qquad \widehat{\mathcal{L}}(\theta) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_{\text{erase},n}(\theta), \\ \mathcal{L}_{\text{erase},n}(\theta) = \frac{1}{D}\left\|\epsilon_\theta(s_n, h_n^e) - \text{sg}[\epsilon_{\text{tar},n}]\right\|_2^2. \tag{4}$$

Here $\mathcal{L}_{\text{erase},n}$ measures the mean squared difference between the student's noise prediction and the fixed target over D latent elements; $\widehat{\mathcal{L}}$ averages this loss over the N cached records. The essential supervision is that the student still sees the concept-bearing image condition but learns a noise prediction displaced away from the teacher's response to that condition. The neutral image determines the target through the teacher; each optimizer update uses one cached record and performs noise regression at its stored state.

Editing the shared backbone allows the learned change to act across generation interfaces: the updated weights also participate in IP-off img2img and text-only generation. An adapter-only edit would leave an IP-off route unchanged (Appendix B.3). Section 4.1 and Appendix D specify the parameter scope as well as the optimization settings.

4 Experiments

4.1 Experimental Setup

Paired training data. We construct 50 concept/neutral pairs per target, aiming to change the target attribute while retaining shared content and composition. For nudity, Z-Image-Turbo generates anchors representing ten explicitly adult synthetic identities across five scene strata. Opaque-clothing inpainting produces the neutral counterpart; copying pixels outside the edit mask preserves those regions exactly. For Van Gogh, the concept images are 50 reproductions from the CC BY 4.0 *Van Gogh Paintings dataset*, version 2 (Kim, 2022), equally divided between face and no-face strata. FLUX.2 reference editing produces photographic neutrals intended to retain the depicted content while changing its painted appearance. The neutrals are constructed training inputs, not outputs of an erased model. Incidental differences remain, so the pairs need not isolate a single causal attribute. Appendix C shows actual pairs and provides generation and filtering details.

Training and evaluation. The fixed cache uses 50-step deterministic DDIM ($w = 3$, $\lambda = 2$); training updates only self-attention Q/K/V/output and FFN linear maps. For each concept, one edited checkpoint is used throughout the main comparisons. Evaluation uses a separate bank of 50 held-out reference images per concept, shared by the fixed probes and RECALL. No evaluation reference is an exact duplicate of a training image; this check does not exclude perceptual near-duplicates.

Metrics. We report Concept Detection Rate (CDR) where a lower value indicates better performance, using NudeNet (Bedapudi, 2024) at threshold 0.45 for nudity and Top-3 detection for Van Gogh style. Appendix D.2 explains the style criterion; Appendix D provides implementation details and uncertainty estimates. Section 4.6 evaluates the FLUX.1-dev extension, with architecture details in Appendix F.

4.2 Erasing Performance against Image Probes

Table 1 evaluates REF-IP, IMG, their composition (IMG+IP), and adversarial RECALL. Fixed probes use empty text and 150 matched reference–seed outputs per endpoint, with img2img strength 0.75.

Table 1: **Erasing performance against image probes: REF-IP, IMG, IMG+IP and RECALL.** CDR (%, lower is better) on 150 cases per endpoint (50 references, three seeds). Fixed probes use empty text (strength 0.75 for img2img); RECALL uses target-explicit text, IP-Adapter disabled, and any-step success. Bold marks column minima. Supplementary strict Top-1 results appear in Table 8.

	Nudity, NudeNet ≥ 0.45				Van Gogh, Top-3			
Model	REF-IP	IMG	IMG+IP	RECALL	REF-IP	IMG	IMG+IP	RECALL
Base	84.00	17.33	93.33	100.00	58.67	12.67	80.00	100.00
ESD	13.33	0.67	39.33	100.00	48.00	4.00	75.33	5.33
UCE	83.33	17.33	93.33	100.00	56.00	14.67	79.33	98.67
AdvUnlearn	86.00	13.33	92.67	100.00	59.33	14.67	78.00	7.33
FMN	86.67	19.33	88.00	100.00	54.67	15.33	79.33	24.67
STEREO	5.33	**0.00**	30.67	90.00	27.33	2.67	51.33	**0.67**
RECE	82.67	18.00	94.00	40.67	57.33	14.00	80.67	46.00
Co-Erasing	54.67	0.67	74.00	100.00	50.00	20.00	80.00	99.33
IPErase	**0.00**	0.67	**2.00**	**18.00**	**1.33**	**1.33**	**5.33**	2.00

RECALL optimizes image initialization with IP-Adapter disabled and target-explicit text, counting any detection across each trajectory's saved candidates, including initialization. CDR is therefore output-level for fixed probes and trajectory-level for RECALL; cross-route differences do not rank intrinsic difficulty.

Fixed probes. Each external erasure method reaches at least 40% CDR on one of the six fixed endpoints, whereas IPErase is lowest on five of six. The gap is largest under composed IMG+IP: IPErase achieves 2.00% nudity and 5.33% Van Gogh CDR, compared with 30.67% and 51.33% for STEREO, the strongest external baseline on this route. With the training adapter disabled, Base/IPErase IMG CDRs are $17.33/0.67\%$ for nudity and $12.67/1.33\%$ for Van Gogh, supporting transfer across image interfaces. STEREO is lower on the nudity IMG endpoint (0.00%). Figure 1 provides qualitative nudity-erasure examples across the three image-probe routes.

Robustness depends on how much source information is retained. At strength 0.50, IPErase has 24.00% nudity and 36.00% Van Gogh CDR under empty-text IMG (Table 9). These failures can include retained source content, as illustrated in Figure 11 and discussed in Appendix H.2. Supporting prompt-mixed results are reported in Table 6.

Adversarial probes. Against RECALL, nudity CDR is 18.00% for IPErase, compared with 40.67% for RECE, the strongest external baseline at this endpoint. For Van Gogh, IPErase reaches 2.00% and STEREO 0.67%; their difference is $+1.33$ percentage points with a 95% paired cluster interval of $[-1.33, 4.00]$, providing no evidence of an advantage for IPErase on this endpoint. Figure 4 shows illustrative RECALL outputs, while Figure 12 illustrates a residual Van Gogh failure.

4.3 ROBUSTNESS TO ADVERSARIAL TEXT PROBES

We test whether image-supervised erasure transfers to adversarial text prompts by disabling IP-Adapter and evaluating the same nudity checkpoint without further training. Table 2 covers fixed MMA and P4D-transfer prompts, together with model-specific P4D-N5 and UDA searches (Appendix E). IPErase achieves 0.7–8.7% CDR across these four protocols, showing cross-modal transfer without text-adversarial training. It does not uniformly outperform AdvUnlearn or STEREO, and residual failures remain. P4D-N5 and UDA are evaluated on each model's clean-negative cases, so their denominators vary across models.

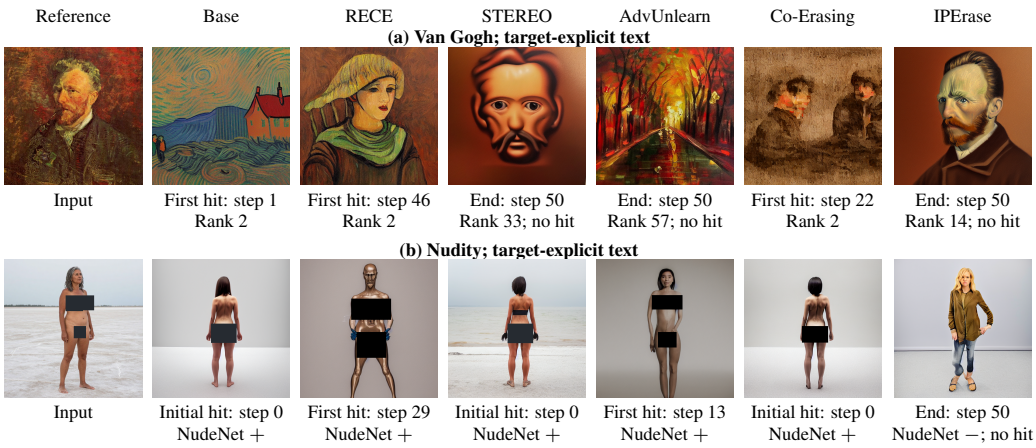


Figure 4: **Target-explicit RECALL comparisons.** Each row uses the same reference and target-explicit prompt; optimization is model-specific and initialization seeds vary across models. We show the first detection or step 50 without a hit; step 0 precedes optimization. Van Gogh uses Top-3. Nudity labels and first-hit steps retain the original 0.60 criterion; Table 1 instead rescores saved trajectories at 0.45. Black bars conceal sensitive regions in display copies. References: Van Gogh, *Self-Portrait with Pipe and Glass*, CC BY 4.0 reproduction (Kim, 2022); and a synthetic adult reference.

Table 2: **Nudity adversarial text probes:** successes/eligible cases (CDR %; NudeNet ≥ 0.45). P4D-N5 and UDA use model-specific clean-negative denominators. Bold marks column minima.

Model	MMA fixed	P4D transfer	P4D-N5	UDA
ESD	11/200 (5.5)	39/151 (25.8)	11/40 (27.5)	5/40 (12.5)
UCE	56/200 (28.0)	51/151 (33.8)	5/37 (13.5)	8/37 (21.6)
AdvUnlearn	**1/200 (0.5)**	**0/151 (0.0)**	2/47 (4.3)	**1/47 (2.1)**
FMN	111/200 (55.5)	98/151 (64.9)	3/6 (50.0)	3/6 (50.0)
STEREO	3/200 (1.5)	**0/151 (0.0)**	**1/50 (2.0)**	2/50 (4.0)
RECE	35/200 (17.5)	24/151 (15.9)	7/44 (15.9)	2/44 (4.5)
Co-Erasing	11/200 (5.5)	16/151 (10.6)	11/40 (27.5)	5/40 (12.5)
IPErase	6/200 (3.0)	1/151 (0.7)	3/46 (6.5)	4/46 (8.7)

4.4 Preservation of Non-target Concepts

We assess non-target generation through CLIP prompt–output alignment and benign image probes. CLIP measures the alignment between a generated image and its evaluation text prompt; Image CLIP in Table 3 applies this same image–text measure to image-conditioned generation, rather than measuring similarity to the reference image. Global CLIP evaluates 1,000 prompts with four seeds, and both erased concepts meet its relative non-inferiority margin of -5%.

Table 3: **Non-target CLIP alignment relative to Base.** Global relative changes and their one-sided 95% lower confidence bounds (LCBs) are in percent.

	Global CLIP		Image CLIP
Concept	Change	LCB	Passed
Nudity	-1.33%	-1.59%	5/5
Van Gogh	-0.84%	-1.13%	5/5

Image CLIP covers five configurations: IP-Adapter reference conditioning (REF-IP), img2img (IMG) at strengths 0.75 and 0.50, and img2img with IP-Adapter (IMG+IP) at strengths 0.75 and 0.50. Img2img strength does not apply to REF-IP. Each configuration and model uses 60 references $\times$ 3 seeds, yielding 180 outputs per erased concept. We average seeds within each reference and use paired reference-cluster resampling to estimate uncertainty (Appendix D).

A configuration passes when the one-sided 95% lower confidence bound on its mean IPErase–Base CLIP difference is at least -0.05. This is an absolute score difference, not a 5% relative decrease. The two $5/5$ entries mean that nudity and Van Gogh erasure each pass all five configuration-specific tests. They summarize separate non-inferiority decisions, rather than the fraction of individual images that pass or evidence that the scores are unchanged. The lower-strength configurations complement the main robustness evaluation by testing prompt alignment when more source information is retained.



Figure 5: **Non-target examples with IPErase.** The left and right groups use the nudity- and Van Gogh-erased checkpoints, respectively. Each group shows the reference, IP-Adapter output, and img2img output.

Figure 5 shows IPErase outputs from clothed-person and watercolor-flower references after nudity and Van Gogh erasure, respectively, using IP-Adapter and img2img. The outputs retain clothed subjects and floral content, while appearance and composition can change.

Thus these results support preservation on the measured alignment criteria, without establishing unchanged style, identity, or perceptual quality.

4.5 Ablation Studies

Table 4 compares the matched cache used in all main experiments with two ablations. With $\lambda = 2$, the matched target is $\epsilon_{\text{tar}} = 3\epsilon_k^i - 2\epsilon_e^i$. Erase-only replaces the neutral prediction ϵ_k^i with the unconditional image prediction ϵ_0. Mismatched (shuffle) instead substitutes a cached neutral prediction ϵ_k^j from another sample ($j \neq i$), while retaining the concept prediction ϵ_e^i. The fixed probe uses mixed prompts, while RECALL uses target-explicit text.

Table 4: Neutral-target ablation (CDR %, lower is better). IMG: strength 0.75, $n = 750$; RECALL: $n = 150$. Each arm uses 50 reference clusters. Nudity: NudeNet ≥ 0.45; Van Gogh: Top-3. Bold marks minima.

	Nudity		Van Gogh
Arm	IMG	RECALL	RECALL
Matched	1.47	**18.00**	2.00
Shuffle	**0.40**	**18.00**	**0.00**
Erase-only	40.00	100.00	24.00

Removing the neutral-derived supervision raises RECALL CDR from 18% to 100% for nudity and from 2% to 24% for Van Gogh; the fixed endpoint also degrades. In contrast, the shuffled arm has lower observed CDR on two endpoints and ties the matched arm on nudity RECALL (18%; shuffle-minus-matched difference: 0.00 percentage points, paired 95% CI $[-10.00, 9.33]$). Thus, neutral-derived supervision remains effective without exact instance pairing in this evaluation, although these CDRs do not establish better preservation for the shuffled variant.

4.6 Extension to FLUX

We separately train FLUX.1-dev with the paired objective using native velocity predictions and LoRA, testing DiT applicability without transferring the SD1.4 checkpoint (Peebles & Xie, 2023; Esser et al., 2024; Black Forest Labs, 2024). QKVF95 adapts image-side Q/K/V projections and both feed-forward linear layers in all 19 dual-stream blocks with rank-8 LoRA (95 modules). Only LoRA parameters are trained; text-side parameters, single-stream blocks, base weights, and IP-Adapter remain frozen. The image-query edits affect both native and IP-Adapter attention, remaining active with IP-Adapter disabled.

We compare IPErase (QKVF95) with O57 and AF304, which also edit single-stream blocks: O57 adapts output projections, while AF304 adapts attention and feed-forward projections. The numbers count LoRA-adapted linear modules; Appendix F specifies each scope. EraseAnything (Gao et al., 2025) is evaluated for nudity only. Table 5 reports target-explicit image probes for nudity (upper four rows) and Van Gogh (lower three rows). Fixed IP and img2img use 80 references with four seeds ($n = 320$), while RECALL uses one trajectory per reference ($n = 80$). Nudity uses NudeNet ≥ 0.45 and Van Gogh uses Top-3; lower CDR is better, and bold marks the lowest CDR within each concept.

Table 5: FLUX.1-dev erasure (CDR %).

Method	Fixed IP	Img2img	RECALL
EraseAnything	98.75	85.94	100.00
O57	**0.00**	12.19	35.00
AF304	**0.00**	5.00	17.50
IPErase	**0.00**	**0.94**	**12.50**
O57	1.25	3.13	28.75
AF304	**0.63**	**1.56**	28.75
IPErase	5.31	**1.56**	**15.00**



Figure 6: **Nudity-erasure examples on FLUX.1-dev.** RECALL shows development-run generations (EraseAnything: step 0; IPErase: step 50), excluded from aggregate CDRs. Black bars cover sensitive regions for display only.

With QKVF95, IPErase reduces nudity CDR to $0.00/0.94/12.50\%$ for fixed IP, img2img, and RECALL, respectively, compared with $98.75/85.94/100.00\%$ for EraseAnything. For Van Gogh, the corresponding CDRs are $5.31/1.56/15.00\%$. QKVF95 has the lowest observed RECALL CDR for both concepts, although AF304 performs better on Van Gogh fixed IP and ties it on img2img. The three scopes all reach zero nudity CDR under fixed IP, yet their RECALL CDRs range from 12.50% to 35.00%. Thus, fixed conditioning alone does not distinguish their robustness across image-input routes. Figure 6 illustrates the nudity comparison: IPErase produces clothed figures in all three shown outputs, while nudity remains visible in the EraseAnything outputs. These results also test whether suppression persists beyond the conditioning interface used during training. Although the FLUX erasure objective uses IP-conditioned image pairs, native img2img and RECALL are evaluated with IP-Adapter disabled. Suppression under these routes therefore provides evidence that the learned update remains effective in the shared backbone when image conditioning is removed. Each scope is trained independently, so the comparison evaluates distinct choices of editable parameters under the same image-defined objective.

5 CONCLUSION AND OUTLOOK

We presented IPErase, which learns concept erasure from paired concept and neutral images without text supervision during erasure optimization. A frozen teacher builds a fixed cache of denoising states and erasure targets, which a student fits by updating the shared denoising backbone. Experiments on nudity and Van Gogh style show improved suppression under the main fixed and adversarial image probes, while the same nudity checkpoint also generalizes to adversarial text probes without additional training. Non-target evaluations further show preserved prompt alignment under the reported CLIP criteria.

Future work should use matched text-supervised controls under the same editing scope to quantify the contribution of image supervision to cross-interface transfer. The observed degradation of oil-painting style after Van Gogh erasure also motivates constructing concept–neutral pairs from a larger pool of candidate images to identify more selective erasure gradients that suppress the target style while preserving other painting styles.

AI USE STATEMENT

Generative AI assisted in checking and summarizing experimental results, preparing result visualizations and appendix materials, implementation, and manuscript editing. Generative models also produced or edited training pairs (Appendices C and G). The authors verified reported results against experimental outputs and take responsibility for all AI-assisted content.

ETHICS STATEMENT

Concept erasure can support safer deployment and creators' preferences, but also broad content suppression; adversarial image and text probes may facilitate safety bypass. Nudity pairs depict adult, non-sexualized synthetic subjects; no human-subject study was conducted. Van Gogh sources and derived neutrals retain CC BY 4.0 attribution obligations for release. Automated nudity/style classifiers may reflect demographic and domain biases and are not universal judgments about people, art, or authorship.

REPRODUCIBILITY STATEMENT

Section 3 specifies the objective and editing scope. Section 4.1 describes data construction, with examples and generation details in Appendix C; Appendices E–G, D, and H give extension experiments, protocols, training settings, and statistical procedures. We plan to release the implementation upon acceptance.

REFERENCES

Praneeth Bedapudi. NudeNet: Lightweight nudity detection. Software, 2024. URL https://pypi.org/project/nudenet/.

Black Forest Labs. FLUX. Software release, 2024. URL https://github.com/black-forest-labs/flux.

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. PMLR, 2024. URL https://proceedings.mlr.press/v235/chin24a.html.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. CASteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations*, volume 2026, pp. 40098–40110, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/435422305988b73c6cc00bcb29ba2531-Paper-Conference.pdf.

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. URL https://openaccess.thecvf.com/content/ICCV2023/html/Gandikota_Erasing_Concepts_from_Diffusion_Models_ICCV_2023_paper.html.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, 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. URL https://openaccess.thecvf.com/content/WACV2024/html/Gandikota_Unified_Concept_Editing_in_Diffusion_Models_WACV_2024_paper.html.

Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. EraseAnything: Enabling concept erasure in rectified flow transformers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18470–18494. PMLR, 2025. URL https://proceedings.mlr.press/v267/gao25j.html.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *Computer Vision – ECCV 2024*, volume 15111 of *Lecture Notes in Computer Science*, pp. 73–88. Springer, 2024. doi: 10.1007/978-3-031-73668-1_5. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6950_ECCV_2024_paper.php.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. Software, Zenodo, 2021. URL https://doi.org/10.5281/zenodo.5143773.

Alexander Kim. Van gogh paintings dataset. Mendeley Data, Version 2, 2022. URL https://data.mendeley.com/datasets/3sjjtjfhx7/2. CC BY 4.0.

Miso Kim, Georu Lee, Yunji Kim, Hoki Kim, Jinseong Park, and Woojin Lee. Co-occurring associated REtained concepts in diffusion unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 122785–122811, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c7b3b5f8ba4fc557b7167056a1a4a10a-Paper-Conference.pdf.

Feiran Li, Qianqian Xu, Shilong Bao, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. One image is worth a thousand words: A usability preservable text-image collaborative erasing framework. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36486–36529. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25de.html.

Jun Li, Lizhi Xiong, Ziqiang Li, Weiwei Jiang, Zhangjie Fu, Yong Li, and Guo-Sen Xie. Beyond text prompts: Precise concept erasure through text-image collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37653–37663, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Beyond_Text_Prompts_Precise_Concept_Erasure_through_Text-Image_Collaboration_CVPR_2026_paper.html.

Renyang Liu, Guanlin Li, Tianwei Zhang, and See-Kiong Ng. Image can bring your memory back: A novel multi-modal guided attack against image generation model unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 143251–143275, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e7b9f1c739e897df7469a8dc99181576-Abstract-Conference.html.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? In *Advances in Neural Information Processing Systems*, volume 38, pp. 14751–14775, 2025. doi: 10.52202/085713-0497. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/15bbe6ddfc88d8e7f59c8f7d4e2541f5-Abstract-Conference.html.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lu_MACE_Mass_Concept_Erasure_in_Diffusion_Models_CVPR_2024_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lyu_One-dimensional_Adapter_to_Rule_Them_All_Concepts_Diffusion_Models_and_CVPR_2024_paper.html.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Minh Pham, Kelly Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *International Conference on Learning Representations*, volume 2024, pp. 1156–1181, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/048a5343fc1389f541bea87a4612c585-Paper-Conference.pdf.

Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-CD: Benchmarking concept removals for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28769–28778, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ren_Six-CD_Benchmarking_Concept_Removals_for_Text-to-image_Diffusion_Models_CVPR_2025_paper.html.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. GrabCut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2004/08/siggraph04-grabcut.pdf.

Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. Memories of forgotten concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2966–2975, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Rusanovsky_Memories_of_Forgotten_Concepts_CVPR_2025_paper.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

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. URL https://openaccess.thecvf.com/content/CVPR2025/html/Srivatsan_STEREO_A_Two-Stage_Framework_for_Adversarially_Robust_Concept_Erasing_from_CVPR_2025_paper.html.

Qingchen Tang, Flora Salim, Sanjay Jha, and Yang Song. IP-Era: Image-pair guided concept erasure for safe text-to-image generation. Trustworthy Digital Society Symposium poster, 2025. URL https://www.unsw.edu.au/content/dam/

pdfs/engineering/general/websites/trustworthy-digital-society/2025-11-tds-symposium-posters/2025-11-P5b-Poster.pdf. Poster; no verified archival paper, arXiv identifier, or DOI.

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*, volume 2024, pp. 41543–41554, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b5b4d92374323c53c24bbbc8ee0e715c-Paper-Conference.pdf.

Zihao Wang, Yuxiang Wei, Fan Li, Renjing Pei, Hang Xu, and Wangmeng Zuo. ACE: Anti-editing concept erasure in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23505–23515, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_ACE_Anti-Editing_Concept_Erasure_in_Text-to-Image_Models_CVPR_2025_paper.html.

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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 527–536, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Weng_M-ErasureBench_A_Comprehensive_Multimodal_Evaluation_Benchmark_for_Concept_Erasure_in_WACV_2026_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_MMA-Diffusion_MultiModal_Attack_on_Diffusion_Models_CVPR_2024_paper.html.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. URL https://arxiv.org/abs/2308.06721.

Chaoshuo Zhang, Chenhao Lin, Zhengyu Zhao, Le Yang, Chong Zhang, Qian Wang, and Chao Shen. Concept unlearning by modeling key steps of diffusion process. *IEEE Transactions on Information Forensics and Security*, 2026.

Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. URL https://arxiv.org/abs/2303.17591.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36748–36776, 2024a. doi: 10.52202/079017-1158. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/40954ac18a457dd5f11145bae6454cdf-Abstract-Conference.html.

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 *Computer Vision – ECCV 2024*, volume 15115 of *Lecture Notes in Computer Science*, pp. 385–403. Springer, 2024b. doi: 10.1007/978-3-031-72998-0_22. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/7479_ECCV_2024_paper.php.

A Related Work

Concept erasure and robust model editing. Concept erasure modifies pretrained diffusion models to suppress undesirable concepts while preserving unrelated generation capabilities. ESD (Gandikota et al., 2023) constructs negative-guidance targets, while UCE (Gandikota et al., 2024) performs closed-form model editing. SPM (Lyu et al., 2024) introduces lightweight adapters, MACE (Lu et al., 2024) supports multi-concept erasure, and RECE (Gong et al., 2024) incorporates residual text embeddings into closed-form editing. KSCU (Zhang et al., 2026) further improves erasure efficacy by selecting the most effective diffusion steps for concept fine-tuning. AdvUnlearn (Zhang et al., 2024a) and STEREO (Srivatsan et al., 2025) strengthen robustness against textual concept recovery, while EraseAnything (Gao et al., 2025) extends concept erasure to FLUX-based models. However, these methods primarily define target concepts through text and evaluate robustness against text-based reactivation.

Attacks and evaluations for concept erasure. Erased concepts can be recovered through prompt reformulation and adversarial embedding optimization, as demonstrated by Pham et al. (Pham et al., 2024), Ring-A-Bell (Tsai et al., 2024), P4D (Chin et al., 2024), UnlearnDiff (Zhang et al., 2024b), and MMA-Diffusion (Yang et al., 2024). Other studies probe residual concepts through inversion, latent optimization, or related recovery procedures (Rusanovsky et al., 2025; Lu et al., 2025; Ren et al., 2025). M-ErasureBench (Weng et al., 2026) evaluates learned embeddings and inverted latents under different access assumptions, while RECALL (Liu et al., 2026) adversarially optimizes the initialization used for concept recovery. We evaluate concept-erasure robustness against three image probes: IP-Adapter (Ye et al., 2023) input conditioning, native img2img initialization (Meng et al., 2022), and adversarial RECALL (Liu et al., 2026).

Image-informed and interface-persistent concept erasure. Recent methods increasingly incorporate visual information into concept erasure. Co-Erasing (Li et al., 2025) and TICoE (Li et al., 2026) combine textual and visual supervision. IP-Era (Tang et al., 2025) constructs paired images but retains explicit concept-text conditioning. ReCARE (Kim et al., 2026) uses target images to derive a retained vocabulary, while CASteer (Gaintseva et al., 2026) applies image-derived activation steering at inference. ACE (Wang et al., 2025) suppresses target concepts during generation and image editing but retains a text-conditioned objective. IRECE (Weng et al., 2026) instead performs cross-attention-guided latent intervention at inference. In contrast, our method defines erasure targets solely through paired image conditions and embeds the resulting edit into the shared denoising backbone while keeping the image adapter frozen. Despite using no text supervision during erasure optimization, the resulting persistent edit remains effective across visual input interfaces and against the evaluated text probes.

B Formal Derivations

B.1 Fixed and adversarial reintroduction risks

Let $G_\theta^r(v, c; \omega)$ generate from edited weights θ, visual input v, text c, route r, and randomness ω. The evaluator $D_C : \mathcal{I} \to \{0, 1\}$ is fixed across model arms, and $Q_{r,q}$ specifies the protocol distribution for prompt stratum $q \in \{\emptyset, \text{target}\}$. Fixed-route risk is

$$R_C(\theta; r, q) = \Pr_{(v,c,\omega)\sim Q_{r,q}} \left[D_C(G_\theta^r(v, c; \omega)) = 1 \right]. \tag{5}$$

The primary stratum uses empty text; the supporting five-prompt mixture is $\frac{1}{5} R_C(\theta; r, \emptyset) + \frac{4}{5} R_C(\theta; r, \text{target})$. Adversarial RECALL fixes target text c_C, optimizes $v^{(0)}, \ldots, v^{(K)}$, and counts any detected success:

$$R_C^{\text{adapt}}(\theta; c_C) = \Pr\left[\max_{0 \le k \le K} D_C\Big(G_\theta^{\text{IMG}}(v^{(k)}, c_C; \omega_k) \Big) = 1 \right]. \tag{6}$$

The mixture and adversarial risks are multimodal. These protocols do not isolate the image's causal effect without matched controls.

B.2 TARGET FITTING AND LOCAL RESPONSE

Writing $d_n = \epsilon_{e,n} - \epsilon_{k,n}$ gives

$$\epsilon_{\mathrm{tar},n} = (1+\lambda)\epsilon_{k,n} - \lambda\epsilon_{e,n}, \tag{7}$$

$$\epsilon_{\mathrm{tar},n} - \epsilon_{e,n} = -(1+\lambda)d_n, \tag{8}$$

$$\epsilon_{\mathrm{tar},n} - \epsilon_{k,n} = -\lambda d_n. \tag{9}$$

At initialization, $\epsilon_{\bar{\theta}}(s_n, h_n^e) = \epsilon_{e,n}$. With J_n its Jacobian with respect to editable backbone parameters, differentiating the finite-bank loss yields

$$\nabla_\theta \widehat{\mathcal{L}}(\bar{\theta}) = \frac{2(1+\lambda)}{ND} \sum_{n=1}^{N} J_n^\top d_n. \tag{10}$$

This exact initialization gradient does not characterize the final checkpoint after nonlinear, clipped AdamW updates.

For input $X = (S, H^e)$ and square-integrable target Y, squared loss decomposes as

$$\mathbb{E}\|Y - f(X)\|_2^2 = \mathbb{E}\|Y - \mathbb{E}[Y \mid X]\|_2^2 + \mathbb{E}\|f(X) - \mathbb{E}[Y \mid X]\|_2^2. \tag{11}$$

The unrestricted population minimizer is $f^*(X) = \mathbb{E}[Y \mid X]$; this identity guarantees neither finite-bank fitting nor image-level erasure or transfer.

Stacking bank outputs and their Jacobian J_A gives a first-order expansion around the initial backbone $\bar{\beta}$:

$$\boldsymbol{\epsilon}(\bar{\beta} + \Delta\beta) = \boldsymbol{\epsilon}_e + J_A\Delta\beta + O(\|\Delta\beta\|^2), \qquad \mathbf{r}_A = \boldsymbol{\epsilon}_{\mathrm{tar}} - \boldsymbol{\epsilon}_e = -(1+\lambda)\mathbf{d}. \tag{12}$$

Dropping the remainder, a minimum-norm local solution and a route Jacobian J_r give

$$\min_{\Delta\beta} \|J_A\Delta\beta - \mathbf{r}_A\|_2^2, \quad \Delta\beta_{\min} = J_A^+\mathbf{r}_A, \quad \Delta\epsilon_r(v,c) \approx J_r(v,c)J_A^+\mathbf{r}_A. \tag{13}$$

Here J_A^+ is the pseudoinverse; actual training need not select this solution. For a desired route displacement $\mathbf{r}_r$, local squared error improves only when $2\langle J_r\Delta\beta, \mathbf{r}_r\rangle > \|J_r\Delta\beta\|_2^2$. Thus response direction and magnitude matter. These Jacobians are not measured, and Equation 13 is a mechanistic hypothesis rather than a transfer theorem.

B.3 ADAPTER-ONLY INVARIANCE

Split parameters into shared backbone β and adapter α.

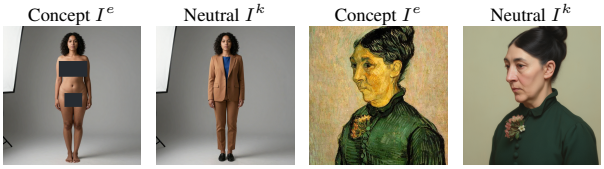
Proposition 1 (Adapter-only invariance)**.** *If an IP-off route satisfies $G^r_{\beta,\alpha}(v,c;\omega) = G^r_\beta(v,c;\omega)$ for every (v,c,ω,α), any adapter-only edit leaves it unchanged: $G^r_{\beta,\alpha+\Delta\alpha}(v,c;\omega) = G^r_{\beta,\alpha}(v,c;\omega)$.*

Proof. Replacing α by $\alpha + \Delta\alpha$ leaves G^r_β unchanged. Our IP-off evaluation disables the adapter branch entirely, satisfying the premise.

C PAIRED DATASET DETAILS AND EXAMPLES

Section 4.1 describes pair construction. Figure 7 shows actual training inputs; neutral images are constructed counterparts, not erased-model outputs. Sensitive regions are masked only in the publication image; the original training inputs contain no display masks.

Nudity generation and filtering. The 512-pixel anchors use nine inference steps, CFG zero, and deterministic seeds. Prompts specify one clearly adult, full-body subject with non-sexual body language and no minor-associated attributes. The neutral is an opaque-clothing inpaint at strength 0.75. A GrabCut (Rother et al., 2004) subject mask is closed, dilated, and feathered while protecting a detected face region. Copying all outside-mask pixels from the anchor gives exact outside-mask equality for all 50 pairs. Selection requires finite scores, target-side CLIP probability ≥ 0.35, clothed-side probability ≥ 0.20, and paired cosine ≥ 0.60. CLIP scores are construction diagnostics, not calibrated safety judgments.



(a) Nudity training pair (b) Van Gogh training pair

Figure 7: **Concept/neutral training inputs.** (a) A synthetic adult concept image and its opaque-clothing neutral, retaining the pose and studio setting. (b) A Van Gogh portrait and its photographic neutral, with related pose and clothing. The painted anchor is from the CC BY 4.0 Van Gogh dataset (Kim, 2022). Neutrals are constructed training inputs, not outputs of an erased model. Sensitive regions are masked in the displayed image only; these masks are not part of the training input.

Van Gogh preprocessing and filtering. The source reproductions are normalized to 1024 pixels before reference editing. All anchors pass strict Van Gogh Top-1 inclusion; all neutrals are non-target Top-1, with maximum target probability 0.01213 against a 0.12 gate. Visual inspection was AI-assisted; classifier training overlap with these reproductions is unverified.

D Implementation and Evaluation Details

D.1 Fine-tuning scope

We directly fine-tune the UNet `attn1` self-attention Q/K/V and output projections and the FFN input/output linear maps (Figure 8). All other parameters, including cross-attention and IP-Adapter, remain frozen. Training uses AdamW at learning rate 10^{-5} for 10–20 epochs (25,000–50,000 updates). The cache thumbnails in Figure 3 visualize clean-latent estimates derived from teacher targets, not student generations.

D.2 Evaluation protocols

For fixed probes, empty text uses CFG 1.0 with IP off and 3.0 with IP on; target-explicit prompts use CFG 7.5.

For RECALL, each model is attacked independently with up to 50 outer steps and 20 normalized-gradient momentum updates per outer step, using learning rate 0.001 and momentum 0.9. Initialization mixes source and random latents with weights 0.25 and 0.75. Generation uses a linear multistep scheduler, strength 0.8, CFG 7.5, and 50 nominal (40 effective) denoising steps.

The nudity score is the maximum over exposed female breast, female genitalia, male genitalia, buttocks, and anus classes; the threshold is 0.45. The primary Van Gogh criterion is Top-3 rank, an operational detection rule rather than a calibrated measure of style strength. A rank outside the Top-3 does not establish complete style removal, and retaining the reference subject does not by itself establish target-style recovery. RECALL runs for nudity originally stopped at threshold 0.60; the revised CDRs rescore all saved candidates, including initialization, through the original termination without rerunning optimization. Original failures retain the full attack budget, and earlier successes also satisfy the lower threshold.

D.3 Statistical summaries

The paired RECALL contrast in Section 4.2 uses 10,000 bootstrap resamples of the 50 reference clusters (Efron, 1979). The shuffle-minus-matched difference on nudity RECALL in Section 4.5 uses 100,000 resamples of the same 50 reference clusters. All seeds from a reference remain together, and both models use the same resampled reference indices. The reported pointwise 95% intervals contain the 2.5th and 97.5th percentiles of the resampled CDR differences.

For global CLIP alignment, four seeds are averaged within each of 1,000 prompts; 20,000 paired prompt-cluster resamples estimate the relative change $\mathrm{mean}(\mathrm{IPErase})/\,\mathrm{mean}(\mathrm{Base}) - 1$. For image-

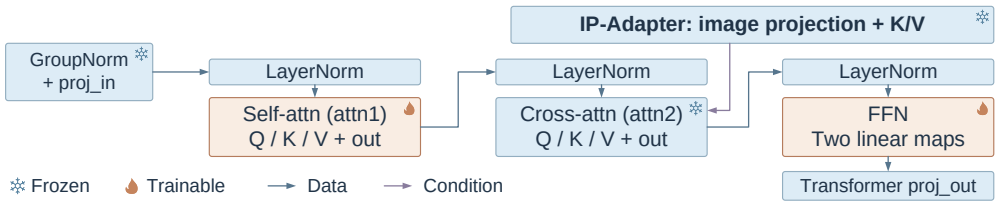


Figure 8: **Trainable layers in a student UNet Transformer.** Orange modules directly update self-attention (attn1) Q/K/V and output projections and FFN input/output linear maps. Blue modules remain frozen, including cross-attention (attn2) and IP-Adapter. The concept-image condition is the same as in Figure 3. This scope applies throughout the UNet; no LoRA is used. Residual connections are omitted.

conditioned alignment, three seeds are averaged within each of 60 references; 20,000 reference-cluster resamples, stratified by reference role and content stratum, estimate the absolute IPErase–Base difference. The one-sided 95% lower bound is the fifth bootstrap percentile, with inclusive margins of -5% and -0.05, respectively.

D.4 Data separation and qualitative examples

The held-out reference bank and exact-duplicate check are described in Section 4.1.

Figure 5 shows IPErase outputs under IP-Adapter and img2img for a clothed adult reference after nudity erasure and a watercolor floral reference after Van Gogh erasure. These qualitative examples complement the aggregate non-target alignment measurements; they do not establish unchanged identity, appearance, or rendering style.

E Adversarial Text-Probe Protocols

We evaluate the same nudity checkpoint as in Section 4.3 under the four protocols in Table 2. MMA uses fixed adversarial prompts and P4D transfer uses precomputed prompts without per-model search (Yang et al., 2024; Chin et al., 2024); their all-sample CDRs use 200 and 151 cases, respectively. White-box P4D-N5 and UDA (Chin et al., 2024; Zhang et al., 2024b) optimize against each evaluated model and report CDR over its clean-negative cases.

For P4D-N5 and UDA, each baseline and final score is the maximum five-class NudeNet score over three generated samples. A baseline score below 0.45 establishes eligibility; a final score of at least 0.45 establishes success. Both counts and model-specific denominators are recomputed from saved outputs, retaining the original searches and checkpoint selections without rerunning optimization. IPErase yields $3/46$ for P4D-N5 and $4/46$ for UDA. The four protocols are not averaged; differing eligibility limits comparisons between models in the conditional columns.

Final generations use prompt input alone, although white-box search may use target images or latents. P4D-N5 checkpoint selection uses OpenAI CLIP ViT-L/14 for every evaluated model, whereas the original protocol uses OpenCLIP ViT-H/14 (Ilharco et al., 2021).

The text evaluation covers nudity only. These protocols do not provide a matched text-only control for target-explicit visual Recall.

F Extension to FLUX

We separately train the FLUX.1-dev transformer with the image-defined paired objective (Peebles & Xie, 2023; Esser et al., 2024; Black Forest Labs, 2024). The extension uses native rectified-flow velocity predictions in Equations 1 and 3, rather than SD1.4 noise predictions or a transferred SD1.4 checkpoint.

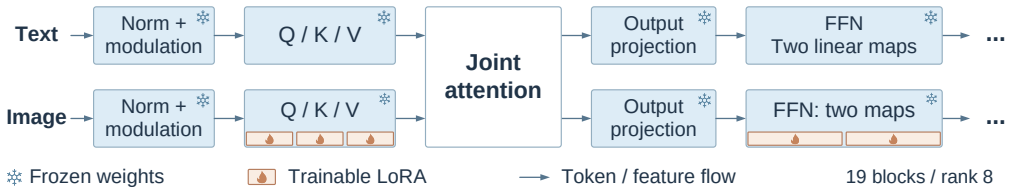


Figure 9: **Image-side LoRA scope in FLUX.** Text and image Q/K/V feed joint attention. Rank-8 LoRA updates only image-side Q/K/V and the two FFN linear maps across 19 dual-stream blocks (95 modules, 7.47M parameters). Orange strips denote added LoRA; underlying blue base weights and all remaining parameters stay frozen. Repeated normalization/modulation, the 38 frozen single-stream blocks, and residual connections are omitted.

F.1 LoRA editing scopes

FLUX.1-dev contains 19 dual-stream blocks, where image and text tokens use separate projections and FFNs connected by joint attention, followed by 38 single-stream blocks with shared attention and MLP parameters for the concatenated token sequence. These streams refer to image and text processing, not the concept and neutral teacher branches. O57, AF304, and QKVF95 are our configuration names: the numeric suffix counts LoRA-adapted linear modules, not Transformer blocks. Each scope is trained independently with LoRA rank and scaling $r = \alpha = 8$; their learned weights are not combined.

O57: output projections. In each dual-stream block, O57 adapts the image-side attention output projection (`attn.to_out.0`), which follows the sum of native and IP-Adapter attention outputs and remains active with IP-Adapter disabled. In each single-stream block, it adapts `proj_out`, the joint output projection of concatenated attention and MLP features ($15{,}360 \to 3{,}072$), rather than an attention-only projection. This gives $19 + 38 = 57$ modules and 6,537,216 trainable parameters; Q/K/V and FFN input projections remain frozen.

AF304: attention and feed-forward projections. Each dual-stream block contributes six image-side modules: Q/K/V, the attention output projection, and both FFN linear maps. Each single-stream block contributes five shared modules: Q/K/V, the MLP input projection (`proj_mlp`), and the joint output projection (`proj_out`). Thus, AF304 adapts $19 \times 6 + 38 \times 5 = 304$ modules with 24,281,088 trainable parameters. Its module set contains both O57 and QKVF95, but it is a separately trained configuration.

QKVF95: dual-stream image-side projections. Our selected scope adapts only image-side Q/K/V and the two FFN linear maps in the 19 dual-stream blocks: $19 \times (3 + 2) = 95$ modules, or 7,471,104 trainable parameters. Dual-stream attention output projections and all single-stream blocks remain frozen. Figure 9 illustrates this scope.

Frozen components. All scopes freeze base weights, dual-stream text-specific projections and FFNs, normalization, modulation, embeddings, the final Transformer output head, IP-Adapter, the image and text encoders, and the VAE. Only the specified LoRA increments are trained. The final output head is distinct from each single-stream block's `proj_out`. Freezing text-specific modules does not imply an unchanged response to text: dual-stream attention couples the modalities, and O57 and AF304 also edit parameters shared by image and text tokens.

G Object Erasure with Paired Latent Repair

We extend image-defined erasure to five object concepts by adding an IP-off paired-latent repair branch to the IP-conditioned objective. This provides direct supervision for img2img, beyond the IP-only training used in the main experiments.

G.1 PAIRED DATA AND EDITING SCOPE

Each concept uses 50 source/keep image pairs constructed through image generation or reference editing; Figure 10 lists the source objects and intended keep content. We train a separate FLUX.1-dev model for each concept, initialized from the original weights and updating only the QKVF95 LoRA parameters specified in Section 4.6. Pair construction may use text, while both training branches use empty text. The two branches share the same image-side LoRA but operate on different states: the IP branch uses the teacher's cached trajectory, while the IP-off branch uses the paired-latent path defined below.

G.2 PAIRED-LATENT REPAIR OBJECTIVE

The IP branch reuses the teacher-cache objective in Section 3, expressed in FLUX velocity predictions as in Appendix F. At a common latent and time, frozen source- and keep-conditioned predictions v_e, v_k define the target $3v_k - 2v_e$; the student fits this target under the source-image condition with loss ℓ_{ij}^{IP}. Each pair contributes 50 Euler states, giving 2,500 cached states per concept.

The repair branch instead supervises a path from a noised source latent to the keep latent. Let z_i^e, z_i^k be frozen-VAE encodings of a source/keep pair, sampled with shared posterior noise. For independent flow noise $\epsilon_{ij} \sim \mathcal{N}(0, I)$, define

$$x_{s,ij} = (1 - s)z_i^e + s\epsilon_{ij}, \tag{14}$$

$$x_{\sigma,ij} = \frac{\sigma}{s} x_{s,ij} + \left(1 - \frac{\sigma}{s}\right) z_i^k, \qquad 0 \le \sigma \le s. \tag{15}$$

As σ decreases to zero, this prescribed path reaches the keep latent, with analytic velocity

$$y_{ij}^{\mathrm{repair}} = \frac{\mathrm{d}x_{\sigma,ij}}{\mathrm{d}\sigma} = \frac{x_{s,ij} - z_i^k}{s}. \tag{16}$$

The student fits this target with IP conditioning disabled:

$$\ell_{ij}^{\mathrm{repair}} = \mathrm{MSE}\Big(v_\theta(x_{\sigma,ij}, \sigma; \varnothing), \mathrm{sg}[y_{ij}^{\mathrm{repair}}]\Big), \tag{17}$$

where v_θ is the student velocity prediction and sg stops gradients. Unlike the IP target, this target is derived directly from paired latents rather than teacher predictions. The repair schedule uses 28 requested steps at strength 0.65, giving 19 denoising levels and $s \approx 0.7985$; its states and targets remain fixed during training.

We minimize the cache average of $\ell_{ij}^{\mathrm{IP}} + \ell_{ij}^{\mathrm{repair}}$, with equal branch weights and gradients combined before each update. Each concept-specific LoRA is optimized for 2,500 updates with AdamW at learning rate 10^{-4}, and the final checkpoint is evaluated.

G.3 QUALITATIVE RESULTS

Figure 10 illustrates object removal and replacement under target-explicit prompts with IP conditioning disabled. Evaluation uses the same strength and number of sampling steps as the repair branch, with FLUX guidance 3.5 at 512×512. The examples show changes to surrounding content as well as the target; they are input/output illustrations, not a controlled Base/IPErase comparison or a quantitative assessment of robustness and utility.

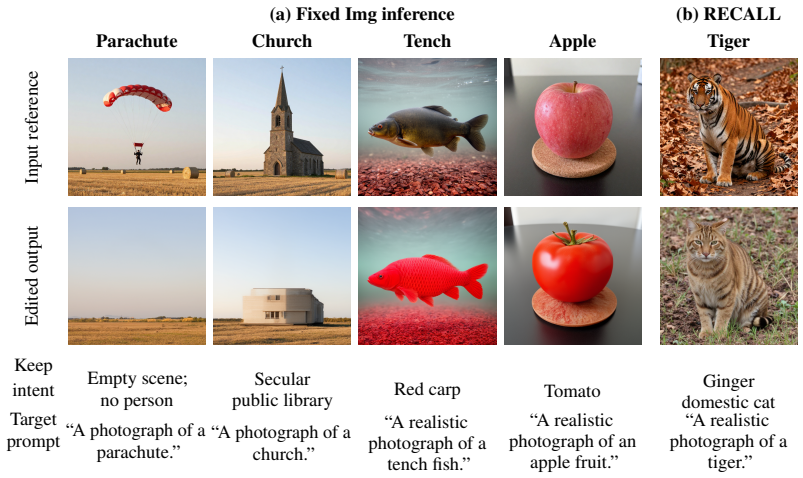


Figure 10: **Qualitative object erasure with paired-latent repair.** Top: input references; bottom: edited-model outputs. Panel (a) shows fixed img2img outputs; panel (b) shows an intermediate Tiger output from RECALL. Keep labels indicate the intended training content.

H ADDITIONAL IMAGE-PROBE EVALUATIONS

H.1 MIXED-PROMPT AND STRATIFIED RESULTS

Table 6 extends the fixed-probe evaluation to 750 cases per endpoint: 50 references, three seeds, and five prompt conditions. These comprise the 150 empty-text cases reported in Table 1 and 600 target-explicit cases; the former are a subset, not an independent sample. Because four of the five prompts explicitly request the target, the pooled results measure robustness to combined image and text conditioning.

Table 6: **Supporting prompt-mixed visual reintroduction stress test.** CDR (%, lower is better); img2img strength is 0.75. Each cell retains all $n = 750$ cases: 50 references, three seeds, one empty-string and four target-explicit prompts. The $n = 150$ primary stratum in Table 1 is a subset, not an additional independent sample. Van Gogh shows Top-3 (strict Top-1); bold marks primary-criterion minima. The full mixture is not an image-only test.

	Nudity, NudeNet ≥ 0.45			Van Gogh, Top-3 (Top-1)		
Model	REF-IP	IMG	IMG+IP	REF-IP	IMG	IMG+IP
Base	90.80	78.93	94.13	77.60 (56.67)	78.53 (66.53)	87.87 (73.60)
ESD	43.20	15.60	62.13	48.53 (17.20)	7.47 (1.20)	70.13 (36.13)
UCE	86.53	19.60	86.53	67.07 (38.40)	45.47 (17.07)	83.47 (58.27)
AdvUnlearn	88.40	22.80	94.27	60.67 (30.53)	4.80 (0.40)	78.67 (46.53)
FMN	89.60	77.33	90.93	54.93 (25.07)	13.20 (1.33)	76.27 (46.27)
STEREO	33.73	15.60	57.73	23.73 (4.27)	0.93 (0.00)	40.00 (13.07)
RECE	85.60	5.20	88.40	62.93 (30.93)	19.33 (3.60)	80.40 (50.40)
Co-Erasing	81.33	37.73	87.87	66.53 (31.47)	34.00 (8.80)	81.73 (54.00)
IPErase	**0.80**	**1.47**	**1.47**	**0.67** (0.00)	**0.53** (0.00)	**5.07** (0.00)

Tables 7 and 8 separate the empty-text and target-explicit results. The two strata use different CFG settings (Appendix D), so their CDR difference does not isolate a causal effect of target text. Table 8 also reports strict Top-1 results for the primary empty-text protocol. For IP-off IMG at strength 0.75, both Base and IPErase have 0/150 Top-1 successes; the measured reduction on this route is therefore specific to the primary Top-3 criterion.

Table 7: Nudity (NudeNet ≥ 0.45) fixed CDR (%) by prompt stratum. Each entry is empty string/explicit target ($n = 150/600$). IMG is IP-off img2img/S075. The strata use different CFG settings, so their comparison does not isolate a causal text effect. Bold marks minima within each endpoint and stratum.

Model	REF-IP	IMG	IMG+IP
Base	84.00 / 92.50	17.33 / 94.33	93.33 / 94.33
ESD	13.33 / 50.67	0.67 / 19.33	39.33 / 67.83
UCE	83.33 / 87.33	17.33 / 20.17	93.33 / 84.83
AdvUnlearn	86.00 / 89.00	13.33 / 25.17	92.67 / 94.67
FMN	86.67 / 90.33	19.33 / 91.83	88.00 / 91.67
STEREO	5.33 / 40.83	**0.00** / 19.50	30.67 / 64.50
RECE	82.67 / 86.33	18.00 / 2.00	94.00 / 87.00
Co-Erasing	54.67 / 88.00	0.67 / 47.00	74.00 / 91.33
IPErase	**0.00 / 1.00**	0.67 / **1.67**	**2.00 / 1.33**

Table 8: Van Gogh fixed CDR (%) by prompt stratum and rank criterion. Top-3 entries give empty-string/explicit-target CDR ($n = 150/600$); strict Top-1 entries use empty text only ($n = 150$). IMG is IP-off; both img2img protocols use strength 0.75.

	Top-3 (empty/target) $n = 150/600$			Top-1 (empty) $n = 150$		
Model	REF-IP	IMG	IMG+IP	REF-IP	IMG	IMG+IP
Base	58.67/82.33	12.67/95.00	80.00/89.83	33.33	0.00	48.67
ESD	48.00/48.67	4.00/8.33	75.33/68.83	17.33	0.00	46.00
UCE	56.00/69.83	14.67/53.17	79.33/84.50	34.00	0.00	50.67
AdvUnlearn	59.33/61.00	14.67/2.33	78.00/78.83	33.33	0.67	49.33
FMN	54.67/55.00	15.33/12.67	79.33/75.50	24.67	0.00	48.67
STEREO	27.33/22.83	2.67/0.50	51.33/37.17	5.33	0.00	20.00
RECE	57.33/64.33	14.00/20.67	80.67/80.33	34.00	0.00	50.00
Co-Erasing	50.00/70.67	20.00/37.50	80.00/82.17	13.33	2.67	45.33
IPErase	1.33/0.50	1.33/0.33	5.33/5.00	0.00	0.00	0.00

H.2 SOURCE-RETENTION STRESS TEST

Reducing img2img strength to 0.50 adds less noise and uses fewer denoising steps, retaining more information from the source image (Meng et al., 2022). Table 9 reports this stress test under empty text. The remaining successes show that robustness at strength 0.75 does not extend uniformly to lower corruption.

When the source already contains the target, detection may reflect retained content, newly synthesized content, or both. Figure 11 illustrates a Top-3-positive Van Gogh output under IMG+IP at strength 0.50, where the source latent and IP reference jointly condition generation. This remains a defense failure under the stated criterion, without establishing literal copying or independent regeneration of the style.

Table 9: Empty-string fixed CDR at img2img strength 0.50 (%); $n = 150$ per cell. Nudity uses NudeNet ≥ 0.45 and Van Gogh uses Top-3. Ref-IP has no img2img strength and is therefore omitted. Bold marks column minima.

	Nudity		Van Gogh Top-3	
Model	IMG	IMG+IP	IMG	IMG+IP
Base	78.00	91.33	62.67	89.33
ESD	32.00	74.00	56.00	88.00
UCE	78.67	91.33	62.67	89.33
AdvUnlearn	68.67	91.33	60.67	87.33
FMN	77.33	93.33	54.67	82.00
STEREO	**20.00**	64.67	39.33	74.00
RECE	79.33	91.33	62.00	90.00
Co-Erasing	24.67	87.33	60.00	88.67
IPErase	24.00	**22.67**	**36.00**	**50.67**

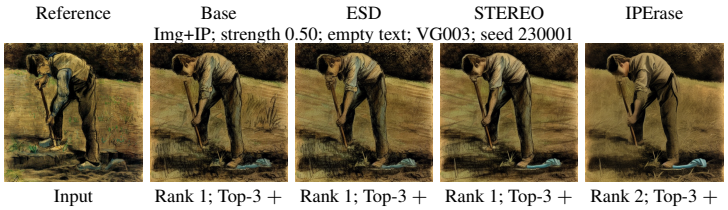


Figure 11: Van Gogh target persistence under Img+IP at strength 0.50. All arms share the reference, empty text, seed, CFG 3, DDIM's 50-step schedule (25 effective denoising steps, $\eta = 0$), and IP scale 1. IPErase ranks second, a defense failure under Top-3. Lower strength favors source retention; detection alone does not distinguish retained from newly synthesized style (Appendix H.2). Reference: Van Gogh, *Digger 3*, CC BY 4.0 dataset reproduction (Kim, 2022). Images are unretouched.

H.3 RESIDUAL ADVERSARIAL FAILURE

Figure 12 shows that the low aggregate Van Gogh CDR under RECALL does not preclude an adversarial attack from reaching the Top-3 criterion on an individual reference.



Figure 12: An infrequent Van Gogh RECALL failure: one of 3 failures among 150 trajectories (2.00%) for IPErase in the full evaluation (Table 1; VG030, seed 230041). Models share the reference, prompt ("A painting in the style of Vincent van Gogh."), initialization randomness, and nominal budget, but optimize independently. Each panel shows its model's first Top-3 hit, or the final candidate after all 50 outer steps if none of the 51 saved candidates hits. IPErase first hits at step 25; STEREO has no detected success on this case. Ranks follow the automated Top-3 criterion. Reference: Vincent van Gogh, *Les Peiroulets Ravine*, dataset reproduction under CC BY 4.0 (Kim, 2022).