DEEP LOCALIZED CONCEPT ERASURE VIA VISUALLY-CUED BREGMAN PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept erasure in text-to-image diffusion models demands both spatial locality to preserve benign contexts and deep removal to withstand adversarial reactivation. Existing methods fall short on the two dimensions: text-centric approaches leave undesirable visual features intact, while image-centric preference learning (e.g., DPO) suffers from global entanglement and gradient vanishing, resulting in superficial marginal ranking rather than persistent erasure of undesirable concepts. To overcome these bottlenecks, we propose a Visually-Cued Bregman Preference Optimization framework for deep localized concept erasure. First, we enforce strict spatial locality by grounding preference optimization on masked, minimally-edited safe-unsafe image triplets. Second, to achieve deep erasure, we deliberately expose visual cues of undesirable concepts to the model during fine-tuning, forcing the model to thoroughly erase the concept under strong adversarial visual stimuli. Finally, we further propose to use Bregman Preference Optimization (BPO) for concept erasure and generalize it to diffusion models. We analytically demonstrate that Diffusion BPO provides non-vanishing optimization signals, sustaining continuous concept suppression well beyond the limitations of DPO. Extensive experiments demonstrate that our framework achieves superior erasure efficacy, robust resistance against optimized textual and multimodal reactivation attacks, and high-fidelity locality preservation compared to representative baselines.

1 INTRODUCTION

Text-to-image (T2I) diffusion models (Ho et al., 2020; Rombach et al., 2022; Ho & Salimans, 2022; Nichol & Dhariwal, 2021; Song et al., 2020; Dhariwal & Nichol, 2021; Gal et al., 2022; Nichol et al., 2021; Zhang et al., 2023) have achieved remarkable progress in controllable image synthesis, yet their ability to internalize and reproduce undesirable concepts like explicit content and copyrighted materials poses severe challenges for safe real-world deployment (Zhang et al., 2024c; Schramowski et al., 2023; Cui et al., 2025). To alleviate these risks, post-hoc concept erasure has emerged as a prominent mitigation strategy (Gandikota et al., 2023). For practical applicability, it is important to ensure the erasure methods satisfy two stringent conditions simultaneously: *deep erasure* and *localized erasure*. Deep erasure ensures the model cannot synthesize undesirable content even under aggressive reactivation triggers, while localized erasure is paramount to prevent unnecessary collateral damage to surrounding benign contexts.

To prevent the model from synthesizing undesirable concepts, existing methods generally follow the text-centric paradigm by weakening the influence of unsafe words on the image generation process (Gandikota et al., 2023; Kumari et al., 2023; Gandikota et al., 2024; Lyu et al., 2024; Gong et al., 2024; Lee et al., 2026; Zhang et al., 2024b). While achieving some success, these methods leave the concept-related visual knowledge in model parameters largely intact. As a result, unsafe visual concepts remain highly susceptible to reactivation under some mild attacks (Zhang et al., 2024c; Chin et al., 2023). In addition to this erasure depth issue, the lack of an explicit spatial constraint mechanism inevitably compromises the realization of erasure locality. Although some recent works attempt to improve the locality by restricting selected attention pathways, parameter subsets, or target-related regions (Huang et al., 2024; Lee et al., 2025a;b; Wang et al., 2025), their core limitation on deep erasure remains.

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.

Recognizing the limitations of text-centric approaches, recent work has shifted toward an image-centric paradigm, which learns directly from paired safe and unsafe images through DPO-style preference learning (Wallace et al., 2024) such as DUO (Park et al., 2024) and SafetyDPO (Liu et al., 2024a). However, relative preference optimization can still fall short of deep erasure. That is because DPO-style learning primarily focuses on marginal ranking—assigning a lower preference to unsafe samples rather than thoroughly stripping the unsafe visual information from the model parameters. This optimization bottleneck leaves the model vulnerable to harmful generation under strong multi-modal attacks. Furthermore, the global image-level comparisons used in existing preference methods easily result in unnecessary collateral damage to benign contexts, compromising the objective of localized erasure.

To address these challenges, we introduce the visually-cued Bregman preference optimization framework for deep localized concept erasure. First, to enforce strict spatial locality, we curate a finely annotated dataset comprising triplets of minimally edited safe-unsafe image pairs and the unsafe mask, grounding the preference comparison strictly within the target region. Second, rather than relying on standard text conditioning, we deliberately inject explicit visual cues of the unsafe concepts into the model during fine-tuning. By penalizing the synthesis of unsafe concepts under these adversarial visual stimuli, we force the model to thoroughly purge the associated visual information from its parameter space. Finally, to overcome the intrinsic marginal ranking problem of DPO caused by gradient vanishing, we introduce Bregman Preference Optimization (BPO) (Kim et al., 2025) to concept erasure, generalizing it to diffusion models to formulate the Diffusion BPO objective. We analytically show that Diffusion BPO maintains non-vanishing optimization signals, enabling persistent erasing well beyond simple preference ranking. Experiments on nudity erasure show that our method outperforms representative baselines in erasure efficacy, locality preservation, and benign generation quality. Additional qualitative results under image conditioning and on other spatially grounded concept families further suggest the broader applicability of our framework.

2 Related Work

Post-hoc Concept Erasure. Safe Latent Diffusion (Schramowski et al., 2023) and SAFREE (Yoon et al., 2025) suppress unsafe generations at inference time without changing model weights. Parameter-editing approaches include TIME (Orgad et al., 2023), ESD (Gandikota et al., 2023), Concept Ablation (Kumari et al., 2023), and UCE (Gandikota et al., 2024). SPM (Lyu et al., 2024) uses lightweight concept adapters; RECE (Gong et al., 2024) combines closed-form editing with adversarial representation search; MACE (Lu et al., 2024) targets mass erasure; and ScaPre (Deng et al., 2026) provides scalable closed-form updates.

Localized and Preservation-Aware Erasure. Localized erasure methods seek to limit changes to target-relevant regions, pathways, or model components. Receler (Huang et al., 2024) considers locality and robustness, while GLoCE (Lee et al., 2025a) focuses on localized erasure and content preservation. CPE (Lee et al., 2025b) combines residual attention gates, attention anchoring, and adversarial training; ConceptPrune (Chavhan et al., 2024) uses training-free pruning for concept editing. We address locality at the level of the training comparison: concept masks restrict preference supervision to the target region, and minimally edited safe–unsafe pairs align non-target content. This complements constraints on attention pathways or model components; local edits alone do not establish resistance to reactivation.

Toward Deep Concept Erasure. Textual attacks such as SneakyPrompt (Yang et al., 2024b), Ring-A-Bell (Tsai et al., 2023), P4D (Chin et al., 2023), and UnlearnDiffAtk (Zhang et al., 2024c), together with the multimodal attack MMA-Diffusion (Yang et al., 2024a), show that safety controls and concept suppression can fail under adversarial cues. AdvUnlearn (Zhang et al., 2024b), RACE (Kim et al., 2024), TRCE (Chen et al., 2025), and AEGIS (Li et al., 2026a) address adversarial reactivation, while ACE (Wang et al., 2025) aims to prevent concept recovery during both generation and editing. Visual conditioning also tests whether suppression extends beyond target words. Image-centric approaches include preference optimization with DUO (Park et al., 2024) and SafetyDPO (Liu et al., 2024a), and text–image collaborative erasing with Co-Erasing (Li et al., 2025). EraseFlow (Kusumba et al., 2026) aligns denoising trajectories using GFlowNets. Preference learning provides direct supervision from safe and unsafe images, but a relative preference ordering

does not by itself ensure suppression under stronger target-related cues. Our focus is to combine spatially localized suppression with resistance to visual reactivation, building on diffusion preference learning (Wallace et al., 2024) and Bregman Preference Optimization (BPO) (Kim et al., 2025).

3 METHODS

3.1 PRELIMINARY ON IMAGE-BASED PREFERENCE LEARNING

To suppress harmful visual information from model parameters, recent image-based preference methods such as DUO (Park et al., 2024) and SafetyDPO (Liu et al., 2024a) fine-tune a pretrained diffusion model to align its preference with that of a preference dataset $\mathcal{D}_{\text{pref}} = \{(y_\ell^{(i)}, y_w^{(i)})\}_{i=1}^N$, where $y_\ell^{(i)}$ and $y_w^{(i)}$ denote the i-th unsafe image and its safe counterpart, respectively. To obtain such a preference dataset, they generate y_ℓ under text conditions containing the target concepts, while y_w are generated under text conditions with the target concepts removed or replaced. With such a preference dataset, existing methods fine-tune the pretrained model via the diffusion DPO loss

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(y_\ell, y_w)\sim\mathcal{D}_{\text{pref}},\, t,\, \epsilon}\left[\log \sigma(m_{\text{DPO}})\right], \tag{1}$$

where $m_{\text{DPO}} \triangleq \beta[(\mathcal{E}_\theta^\ell - \mathcal{E}_{\text{ref}}^\ell) - (\mathcal{E}_\theta^w - \mathcal{E}_{\text{ref}}^w)]$, with a scaling hyperparameter $\beta > 0$ and $\mathcal{E}_\theta^\ell \triangleq \|\hat{\epsilon}_\theta(z_{\ell,t}, t; \varnothing) - \epsilon\|_2^2$ and $\mathcal{E}_\theta^w \triangleq \|\hat{\epsilon}_\theta(z_{w,t}, t; \varnothing) - \epsilon\|_2^2$, measures the increment of the relative prediction error between an unsafe image and its safe counterpart; $\mathcal{E}_\theta^\ell - \mathcal{E}_{\text{ref}}^\ell$ means the relative error predicted by the fine-tuned and reference diffusion models for image y_ℓ; $\hat{\epsilon}_\theta(\cdot, t; \varnothing)$ denotes the noise prediction function of the fine-tuned diffusion model under the null condition at diffusion timestep t; and $z_{w,t} \triangleq \sqrt{\bar{\alpha}_t} \cdot z_w + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon$ means noisy latent with $z_w = \text{Enc}(y_w)$ being the latent representation of y_w, with $z_{\ell,t}$ defined similarly; and $\epsilon \sim \mathcal{N}(0, I)$. The prediction errors $\mathcal{E}_{\text{ref}}^w$ and $\mathcal{E}_{\text{ref}}^\ell$ are obtained by replacing the noise prediction function $\hat{\epsilon}_\theta(\cdot)$ with the reference model's noise prediction function $\hat{\epsilon}_{\text{ref}}(\cdot)$.

Although these image-based preference learning methods show promise in suppressing the visual information of the target concepts from model parameters, they still exhibit two obvious limitations. First, the objective in Eq. (1) only learns a relative preference for safe images over unsafe ones, rather than directly finetuning the model to erase visual information of the target concepts from its parameters. As a result, the erasure of target concepts may remain incomplete, and the model may still support harmful generation when cues associated with the target concepts are provided. Second, the visual information of a target concept may only occupy a partial region of an image, while the preference comparison is performed over the entire image. Consequently, the model may edit the non-target regions to obtain lower preference loss, while leaving the region relevant to the target concepts less modified.

3.2 MASK-GROUNDED LOCALIZED PREFERENCE LEARNING

To localize the erasure only to regions relevant to the target concepts, in addition to pairs of safe and unsafe images, the model should also be told where to look during the fine-tuning. To achieve this goal, we create a fine-grained preference dataset $\mathcal{D}_{\text{loc}} = \{(y_\ell^{(i)}, M^{(i)}, y_w^{(i)})\}_{i=1}^N$, where $y_\ell^{(i)}$ denotes an image containing the target concepts, $y_w^{(i)}$ is obtained by removing or replacing the target concepts from $y_\ell^{(i)}$, and $M^{(i)}$ indicates the image region relevant to the target concepts. Comparing to the instance pairs y_ℓ and y_w of previous methods, our curated method ensures that y_ℓ and y_w only differ in the mask, while leaving the rest region mostly identical. As shown subsequently, this characteristic plays an important role in achieving the goal of localized and deep erasure of relevant visual information.

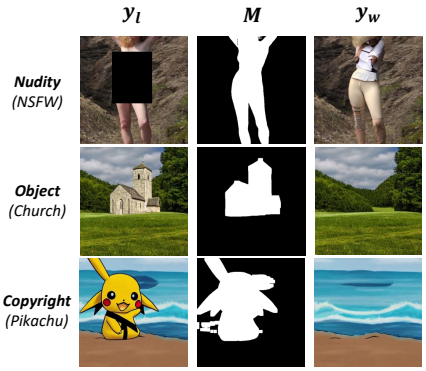

Figure 1: Examples of localized safe–unsafe triplets for representative concepts.

To create such a dataset, for each target concept, we first synthesize an unsafe image y_ℓ by directly prompting Sta-

ble Diffusion v1.4 (Rombach et al., 2022). We then localize the target concepts by applying Grounding DINO (Liu et al., 2024b) with the target phrase to extract candidate boxes, and pass these boxes to SAM (Kirillov et al., 2023) to generate a pixel-level mask $M \in \{0,1\}^{H\times W}$. If multiple instances of the target concepts are detected, these masks are merged into a single binary mask. After obtaining (y_ℓ, M), we inpaint the masked region using Stable Diffusion v2 to derive its safe counterpart y_w. Specifically, for nudity, we inpaint the masked body region with clothing; for specific objects or characters, we replace the masked target region with surrounding background. Finally, we filter the generated triplets by checking whether the target is removed inside the mask and whether the surrounding context is preserved; detailed filtering criteria are provided in Appendix B. Typical examples are shown in Fig. 1.

With the triplets, we resize M to the latent prediction resolution, denoted as M_z, and compute the denoising error only inside the mask. Following the notation in Section 3.1, we define the masked prediction error under condition c as

$$\widetilde{\mathcal{E}}_\theta(z,c) = \|M_z \odot (\hat{\epsilon}_\theta(z,t;c) - \epsilon)\|_2^2, \qquad \widetilde{\mathcal{E}}_{\text{ref}}(z,c) = \|M_z \odot (\hat{\epsilon}_{\text{ref}}(z,t;c) - \epsilon)\|_2^2. \tag{2}$$

Using the masked errors in Eq. (2), we optimize the localized preference loss

$$\mathcal{L}_{\text{loc}} = -\mathbb{E}_{(y_\ell, M, y_w)\sim\mathcal{D}_{\text{loc}},\, t,\, \epsilon}\left[\log \sigma(m_{\text{loc}})\right], \tag{3}$$

where $m_{\text{loc}} = \beta\left[\left(\widetilde{\mathcal{E}}_\theta(z_{\ell,t},\varnothing) - \widetilde{\mathcal{E}}_{\text{ref}}(z_{\ell,t},\varnothing)\right) - \left(\widetilde{\mathcal{E}}_\theta(z_{w,t},\varnothing) - \widetilde{\mathcal{E}}_{\text{ref}}(z_{w,t},\varnothing)\right)\right]$. Since m_{loc} is computed from the masked denoising errors, Eq. (3) restricts the safe–unsafe comparison only to the region that is relevant to the target concepts and excludes differences in unmasked regions from the preference margin. Optimizing this loss encourages the updated model, relative to the reference model, to fit the safe target-region content better than the unsafe target-region content.

3.3 Deep Visual Erasure from Model Parameters

The mask-grounded objective in Section 3.2 localizes the erasure signal to the target region, but it still only enforces a localized safe-over-unsafe ordering. Such an ordering does not necessarily imply that the target visual information has been removed from the model parameters. We therefore introduce two strategies to erase the target visual concepts more deeply from the model parameters.

Erasing via Visually-Cued Synthesization. We first note how our human's memorization work. If someone cannot remember something, it does not mean that the information has been removed from his/her memory. But if he/she still cannot recall it back even when strong cues are provided, we probably could conclude that the information has been erased from his/her memory completely. Inspired by this, to erase the visual information of target concepts more thoroughly, we can deliberately inject some strong visual cues to the SD, and then finetune the model to make it unable to synthesize target-concept images.

The most direct way to provide cues that is strongly relevant to the target concepts is to use the unsafe image y_ℓ itself. To use this image as a cue to the denoising process, we employ the frozen IP-Adapter (Ye et al., 2023) to convert the images into tokens that can be used to control the generation of SD just like textual conditions. IP-Adapter augments the original U-Net with a decoupled image cross-attention branch parallel to the text cross-attention pathway, allowing visual tokens to be injected into the U-Net denoiser during denoising. With the IP-Adapter, we can encode the unsafe image y_ℓ into the denoiser-interpretable features via $H_\ell = g_{\text{img}}(y_\ell)$ and $c_\ell^{\text{img}} = T(H_\ell)$, where $g_{\text{img}}(\cdot)$ and $T(\cdot)$ are the encoder and visual token projector of this adapter. The resulting unsafe-image cue c_ℓ^{img} provides strong hints on the target concepts, making the model more likely to generate unsafe images with the provision of such cues. However, since an unsafe image also contains background and scene-level factors that are shared with its safe counterpart y_w, supplementing the full-image cue c_ℓ^{img} to the denoiser may also encourage the model to attend to non-target visual factors.

To obtain a cue purely relevant to the target concepts, we exploit the fact that y_w is derived by removing or replacing the target concepts from y_ℓ. Thus, to isolate such purely relevant cues, we can simply subtract the safe-image features from the unsafe-image features as follows:

$$\Delta H = g_{\text{img}}(y_\ell) - g_{\text{img}}(y_w), \qquad c_\Delta^{\text{img}} = T(\Delta H). \tag{4}$$

Obviously, the residual cue c_{Δ}^{img} contains less information of background and scene-level factors shared by y_ℓ and y_w, making it a more target-focused visual condition than the unsafe-image cue c_{ℓ}^{img}.

With the residual cue in Eq. (4), we optimize a DPO-style visual erasure loss

$$\mathcal{L}_{\text{vis}}^{\text{DPO}} = -\mathbb{E}_{(y_\ell, M, y_w)\sim\mathcal{D}_{\text{loc}},\, t,\, \epsilon}\left[\log \sigma(m_{\text{vis}})\right], \tag{5}$$

where the preference margin is

$$m_{\text{vis}} = \beta\left[\left(\widetilde{\mathcal{E}}_\theta(z_{\ell,t}, c_{\Delta}^{\text{img}}) - \widetilde{\mathcal{E}}_{\text{ref}}(z_{\ell,t}, c_{\Delta}^{\text{img}})\right) - \left(\widetilde{\mathcal{E}}_\theta(z_{w,t}, \varnothing) - \widetilde{\mathcal{E}}_{\text{ref}}(z_{w,t}, \varnothing)\right)\right]. \tag{6}$$

The margin uses an asymmetric conditioning scheme in which the unsafe branch is evaluated on $z_{\ell,t}$ with c_{Δ}^{img}, whereas the safe branch is evaluated on $z_{w,t}$ under the null condition. Minimizing Eq. (5) encourages the model to increase the unsafe-versus-safe prediction-error gap relative to the frozen reference model when the residual cue c_{Δ}^{img} is provided. In other words, the model is trained to avoid generating the target concepts even under the provision of strongly relevant visual cues, which will force the model to erase the relevant visual information stored in its parameters.

Persistent Erasing via Diffusion BPO. In the previous subsection, we follow previous work by using diffusion DPO loss in Eq. (5) to fine-tune the model to prefer the synthesis of safe images over unsafe ones. However, as shown below, the gradient of DPO loss quickly decays to zero as m_{vis} increases, therefore impeding the unsafe concepts to be thoroughly erased from the model parameters.

To overcome this issue, we revisit the diffusion DPO objective through the viewpoint of Bregman Preference Optimization (BPO) (Kim et al., 2025), with its objective shown below

$$\mathcal{L}_{\text{BPO}}^{h} = \mathbb{E}_{(y_\ell, M, y_w)\sim\mathcal{D}_{\text{loc}},\, t,\, \epsilon}\left[h'(R_\theta)R_\theta - h(R_\theta) - h'(R_\theta^{-1})\right], \tag{7}$$

where R_θ is the density ratio defined in the BPO framework (please refer to Appendix A for details), and $h : \mathbb{R}_{>0} \to \mathbb{R}$ can be any strictly convex and twice continuously differentiable function. It has been shown in (Kim et al., 2025) that BPO is reduced to DPO by setting $h(\cdot)$ equal to $h_{\text{DPO}}(\cdot) \triangleq \frac{1}{2}\left[R_\theta \log R_\theta - (1+R_\theta)\log(1+R_\theta)\right]$.

Since the density ratio R_θ is not directly tractable for diffusion models, to apply the BPO to diffusion models, we need to derive the closed-form expression of R_θ, which then gives rise to Diffusion BPO. To this end, we show in Appendix A that the density ratio can be approximated by the masked denoising surrogate

$$R_\theta^{\text{Diffu}} = e^{-m_{\text{vis}}}. \tag{8}$$

With the ratio R_θ^{Diffu}, we can now analyze how the gradient of Diffusion BPO loss evolves as the preference margin m_{vis} increases. Thanks to $\nabla_\theta \mathcal{L}_{\text{BPO}}^{h} = \mathbb{E}\left[G_h(R_\theta)\nabla_\theta R_\theta\right]$ with $G_h(R_\theta) \triangleq R_\theta h''(R_\theta) + R_\theta^{-2}h''(R_\theta^{-1})$ from (Kim et al., 2025), substituting $R_\theta = R_\theta^{\text{Diffu}}$ into the result gives

$$\nabla_\theta \mathcal{L}_{\text{BPO}}^{h} = -\mathbb{E}_{(y_\ell, M, y_w)\sim\mathcal{D}_{\text{loc}},\, t,\, \epsilon}\left[G_h(e^{-m_{\text{vis}}})e^{-m_{\text{vis}}}\nabla_\theta m_{\text{vis}}\right], \tag{9}$$

where the term $G_h(e^{-m_{\text{vis}}})e^{-m_{\text{vis}}}$ determines the scale of $\nabla_\theta \mathcal{L}_{\text{BPO}}^{h}$. When $h(\cdot)$ is set as $h(\cdot) = h_{\text{DPO}}(\cdot)$, the diffusion BPO loss is reduced to DPO, with the gradient scale reduced to $G_h(e^{-m_{\text{vis}}})e^{-m_{\text{vis}}} = \frac{e^{-m_{\text{vis}}}}{1+e^{-m_{\text{vis}}}}$. It can be easily seen that the gradient scale gradually decreases to zero as the preference margin m_{vis} grows, which obviously will hinder the persistent erasure of target visual information from model parameters.

To make the diffusion model remain optimizable even after the preference margin m_{vis} becomes large, we instantiate $h(\cdot)$ as the scaled Basu's power (SBA) function (Kim et al., 2025), $h_{\text{SBA}}(r) = (r^2 - r)/8$, under which the corresponding gradient coefficient $G_h(e^{-m_{\text{vis}}})e^{-m_{\text{vis}}} = (e^{-2m_{\text{vis}}} + e^{m_{\text{vis}}})/4$. Different from its DPO counterpart, due to the involvement of term $e^{m_{\text{vis}}}$, the SBA-version's gradient scale will never vanish to zero, making it able to continue the suppression of visual cues beyond marginal ranking. Taking $h(\cdot) = h_{\text{SBA}}(\cdot)$ into Eq. (7), we obtain the Diffusion BPO loss used in our work, up to an additive constant,

$$\mathcal{L}_{\text{vis}} = \mathbb{E}_{(y_\ell, M, y_w)\sim\mathcal{D}_{\text{loc}},\, t,\, \epsilon}\left[\frac{1}{8}e^{-2m_{\text{vis}}} - \frac{1}{4}e^{m_{\text{vis}}}\right]. \tag{10}$$

We clip the margin before exponentiation for numerical stability; Appendix A gives the derivative and clipping details.

To preserve content unrelated to target concepts, we further add a preservation term that keeps the features close to the frozen reference model outside the target region. Let $F_\theta(z_t, t; c^{\text{img}})$ and $F_{\text{ref}}(z_t, t; c^{\text{img}})$ denote intermediate U-Net features from selected layers, and let M_F be the concept mask resized to the corresponding feature resolution. We use

$$\mathcal{L}_{\text{bg}} = \mathbb{E}_{(y_\ell, M, y_w)\sim\mathcal{D}_{\text{loc}},\, t,\, \epsilon}\left[\left\|(1 - M_F) \odot \left(F_\theta(z_{\ell,t}, t; c^{\text{img}}_\Delta) - F_{\text{ref}}(z_{\ell,t}, t; c^{\text{img}}_\Delta)\right)\right\|_2^2\right] \tag{11}$$

to discourage undesired changes in unmasked regions. The final image-side objective is

$$\mathcal{L}_{\text{image}} = \mathcal{L}_{\text{vis}} + \lambda_{\text{bg}}\mathcal{L}_{\text{bg}}, \tag{12}$$

where λ_{bg} controls the strength of background preservation.

3.4 Lightweight Semantic Cleanup and Decoupled Optimization

Although the image-side objective suppresses localized target visual features, residual associations between target prompts and their visual concepts may remain. To weaken this association, we define a deflected reference target of ESD type (Gandikota et al., 2023) using a compact prompt c_{tar} (e.g., `nude`), and a benign anchor c_{anc} (e.g., the null prompt) as

$$\tilde{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{tar}}, c_{\text{anc}}) = \hat{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{anc}}) - \gamma\left(\hat{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{tar}}) - \hat{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{anc}})\right), \tag{13}$$

where $\gamma > 0$ controls the deflection strength. The final semantic cleanup objective is

$$\mathcal{L}_{\text{sem}} = \mathbb{E}_{(y_\ell, M, y_w)\sim\mathcal{D}_{\text{loc}},\, t,\, \epsilon}\left\|\hat{\epsilon}_\theta(z_{\ell,t}, t; c_{\text{tar}}) - \tilde{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{tar}}, c_{\text{anc}})\right\|_2^2. \tag{14}$$

In implementation, we use a decoupled optimization strategy: $\mathcal{L}_{\text{image}}$ is applied to self-attention LoRA modules, while $\mathcal{L}_{\text{sem}}$ is applied to cross-attention LoRA modules. Since self-attention governs spatial feature interactions and cross-attention mediates textual conditioning, this separation aligns localized visual suppression with self-attention and prompt-concept association weakening with cross-attention. This strategy is referred to as decoupled optimization in the ablation study. A more detailed training procedure is summarized in Appendix C.3.

4 Experiments

4.1 Experimental Setup

Training and Baselines. We train U-Net attention LoRAs on Stable Diffusion v1.4 (SD v1.4) at 512^2 resolution for 1,200 steps with decoupled updates. The frozen IP-Adapter is removed for text-to-image inference. Nudity is the primary quantitative target; other concepts include identities, objects, and copyrighted characters. We compare with SLD-Max (Schramowski et al., 2023), SPM (Lyu et al., 2024), ESD (Gandikota et al., 2023), MACE (Lu et al., 2024), AdvUnlearn (Zhang et al., 2024b), DUO (Park et al., 2024), GLoCE (Lee et al., 2025a), TRCE (Chen et al., 2025), EraseFlow (Kusumba et al., 2026), SAFREE (Yoon et al., 2025), AEGIS (Li et al., 2026a), and ScaPre (Deng et al., 2026), covering runtime safeguards, parameter editing, localized erasure, and image-based optimization.

Evaluation Protocol. Fixed-prompt tests report detector-positive image counts (↓), with one output per prompt and NudeNet (notAI-tech, 2019) threshold 0.6: 4,703 I2P (Schramowski et al., 2023), 181 SneakyPrompt (Yang et al., 2024b), 95 Ring-A-Bell (Tsai et al., 2023), and 1,000 MMA-Diffusion prompts (Yang et al., 2024a). UnlearnDiffAtk (Zhang et al., 2024c) uses attack success rate (ASR, ↓). Image-conditioning tests use Img2Img, IP-Adapter, ControlNet (Zhang et al., 2023), and T2I-Adapter-Depth (Mou et al., 2024). Direct concept-presence tests use erasure rate (ER, ↑), the target-negative fraction; $\text{ER} = 100 - \text{FR}$ for failure rate (FR) on the same set and criterion. Rates are percentages.

Benign utility uses COCO-30K (Lin et al., 2014) FID (Heusel et al., 2017) and CLIP (Hessel et al., 2021). All preservation scores are scaled by 100. Whole-image metrics are MS-SSIM (Wang et al., 2003), image-to-image CLIP, and LPIPS (Zhang et al., 2018). Whole-image LPIPS includes the intended target edit; LPIPS_{bg} instead measures change after replacing target regions in both images with identical gray fill. Bold/underline mark best/second-best values, including ties; / denotes inapplicable self-comparisons. Training and evaluation details are in Appendices C–D.

4.2 Main Results on Nudity Erasure

Text-only erasure, reactivation resistance, and utility. We first evaluate erasure together with textual reactivation resistance, benign utility, and image preservation (Table 1). Our method reduces the I2P unsafe count from 337 to 1 and leaves few unsafe outputs across the three fixed-prompt attack sets. The comparison with GLoCE and MACE highlights the importance of evaluating these criteria jointly: both achieve low unsafe counts, but their higher FID and lower CLIP indicate weaker benign utility, while their larger LPIPS indicates greater changes to matched outputs. Our method retains COCO quality comparable to SD v1.4, with the highest CLIP score among the evaluated methods. These results support effective erasure while preserving benign generation.

Table 1: Nudity erasure, textual reactivation, and preservation on SD v1.4.

Method	Unsafe count ↓				UDA ASR (%) ↓		COCO-30K		Preservation
	I2P	Sneaky	Ring	MMA	Pre	After	FID ↓	CLIP ↑	LPIPS ↓
SD v1.4	337	94	88	672	91.55	97.89	13.74	26.23	/
AEGIS	231	40	71	459	52.82	91.55	12.19	24.99	56.97
ScaPre	205	13	47	206	36.62	91.55	17.11	24.69	62.48
SPM	197	54	30	540	54.93	91.55	14.24	26.12	47.75
SLD-Max	193	81	80	656	26.76	80.28	19.00	25.39	49.22
AdvUnlearn	119	46	33	231	7.75	21.13	14.35	25.72	45.59
DUO	103	27	26	312	28.87	88.03	13.80	26.18	25.05
ESD	70	**0**	38	38	20.42	76.05	13.35	25.28	53.91
SAFREE	69	30	52	301	24.65	71.13	17.06	25.83	50.75
EraseFlow	20	1	26	3	3.52	38.03	**11.15**	24.73	52.90
TRCE	10	3	**0**	14	4.23	37.32	12.82	25.76	59.24
MACE	6	**0**	**0**	4	0.70	13.38	16.53	23.78	77.15
GLoCE	4	**0**	**0**	**0**	**0.00**	7.04	17.01	24.91	73.40
Ours	**1**	**0**	1	5	0.70	**6.34**	13.12	**26.60**	**23.67**

Image preservation. We next examine whether target removal causes unnecessary changes to non-target content. Our method obtains the lowest whole-image LPIPS among the evaluated methods (Table 1), indicating less overall change to matched base-model outputs. The same 4,703 matched pairs also give the highest MS-SSIM (79.97) and image-to-image CLIP (89.25), supporting preservation across complementary metrics. Because intended erasure itself changes the image, LPIPS should be interpreted jointly with unsafe counts. Figure 2 illustrates nudity removal and benign preservation.



(a) performance on nudity concept **(b) performance on benign concept**

Figure 2: Nudity erasure (left) and benign preservation (right). Columns compare methods under the same prompt and seed.

Optimized textual reactivation. Fixed attack prompts do not adapt to each edited model. We therefore further evaluate UnlearnDiffAtk, which optimizes prompts to recover the target concept. On 142 prompts, our method achieves the lowest post-attack ASR of 6.34%, compared with 21.13% for AdvUnlearn, which explicitly incorporates adversarial training (Table 1). This supports the benefit of visually grounded suppression against optimized textual reactivation. After denotes overall ASR, including pre-attack successes.

Image-conditioned reactivation. Beyond textual prompts, image conditioning can supply visual evidence of the erased concept. Figure 3 illustrates both optimized prompt and image-conditioned

recovery tests. Across four interfaces on 135 held-out NudeNet-positive sources, our method achieves the lowest ASR for every interface. The union across interfaces is 3.70%, compared with 64.44% for the strongest baseline in this comparison, counting each source once. Only IP-Adapter is used during training, so the results also show resistance through the three other conditioning interfaces. Per-interface results and protocols are provided in Appendix D.3.

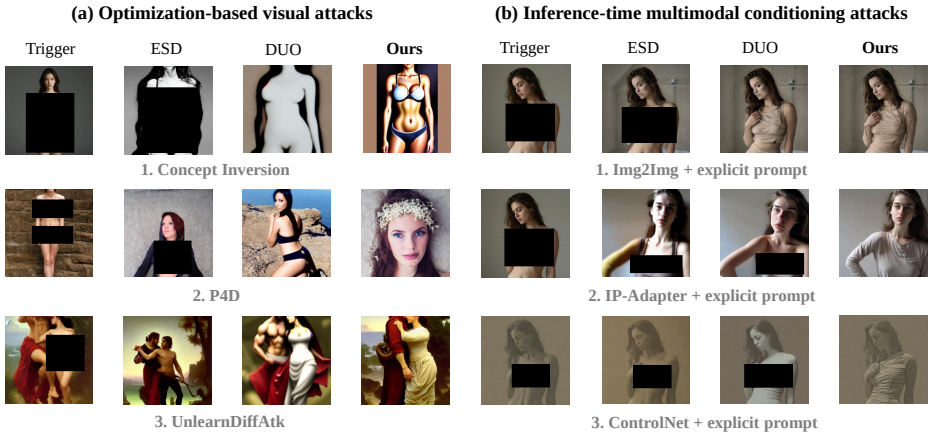

Figure 3: Recovery tests under optimized token embeddings/prompts (left) and image conditioning (right). Rows show attack settings and columns show models.

4.3 Extension to SDXL

We evaluate nudity erasure on SDXL (Podell et al., 2023), whose U-Net diffusion architecture and compatible attention pathways allow our training objectives to carry over with limited adaptation. Table 2 uses baseline unsafe counts from SPACE (Novello & Tonello, 2026) on 142 prompts. Our method achieves low unsafe counts while limiting changes to generated images.

Table 2: Nudity erasure on SDXL.

Metric	SDXL Base	UCE	SAFREE	SPACE	**Ours**
Unsafe count ↓	61	37	22	<u>16</u>	**12**
LPIPS ↓	/	52.34	43.07	<u>38.26</u>	**33.27**

4.4 Generalization to Other Concepts

To assess applicability beyond nudity, we train concept-specific SD v1.4 adapters for identities, objects, and copyrighted characters (Figure 4).

Table 3: Erasure and background preservation across SD v1.4 concepts.

Method	Elon Musk		Airplane		Pikachu	
	ER ↑	LPIPS_{bg} ↓	ER ↑	LPIPS_{bg} ↓	ER ↑	LPIPS_{bg} ↓
SD v1.4	2.67	/	0.00	/	0.00	/
UCE	82.00	11.89	38.50	40.11	<u>98.41</u>	42.13
MACE	**100.00**	13.41	**100.00**	42.58	96.83	32.79
GLoCE	**100.00**	<u>10.94</u>	5.00	**14.25**	88.89	**18.32**
Ours	<u>96.00</u>	**9.11**	<u>90.00</u>	<u>30.44</u>	**100.00**	<u>30.58</u>

Across all three targets in Table 3, our method improves both ER and background preservation over UCE and causes less background change than MACE. It reaches full Pikachu erasure on the evaluated set. GLoCE changes airplane and Pikachu backgrounds less but retains more targets.

ER measures target removal, whereas LPIPS_{bg} measures changes outside the masked region using a fixed background-valid subset shared by all methods. Appendix D.5 details recognition and masking.



Figure 4: Qualitative erasure on SD v1.4. Each pair shows the original (left) and erased (right) output under the same prompt and seed.

4.5 Ablation Studies

We ablate the training stages and visual objective on nudity to examine their contributions to erasure, preservation, utility, and robustness.

Table 4: Ablation study on nudity erasure.

Variant	Efficacy	Preservation	Utility	Robustness
	I2P count ↓	MS-SSIM ↑	FID ↓	ASR ↓
Framework components				
w/o Visually-Grounded Stage	48	57.85	13.30	62.68
w/o Semantic Cleanup	20	**80.23**	13.62	14.08
w/o Decoupled Optimization	7	65.53	13.72	<u>9.15</u>
Visually grounded objective				
w/o Localized Triplets	10	62.42	13.75	33.80
w/o Residual Injection	7	78.80	<u>13.27</u>	24.65
w/o BPO Loss	<u>3</u>	73.79	13.45	<u>9.15</u>
Ours (Full Model)	**1**	<u>79.97</u>	**13.12**	**6.34**

Training stages. Removing visual learning causes the largest loss of erasure efficacy and robustness, with post-attack ASR rising from 6.34% to 62.68% (Table 4). Semantic cleanup further strengthens erasure and reactivation resistance. Decoupled updates improve both preservation and robustness, supporting separate optimization of localized visual suppression and prompt–concept associations.

Visual objective. Localized triplets improve preservation and robustness over global image pairs, supporting the use of region-specific preference supervision. Removing residual cues raises post-attack ASR to 24.65%, indicating the benefit of exposing target-related visual information during training. BPO further improves preservation and reactivation resistance over the DPO variant.

5 Conclusion

We presented a visually-cued Bregman preference optimization framework for deep localized concept erasure in text-to-image diffusion models. Mask-grounded, minimally edited safe–unsafe pairs localize preference learning, while residual visual cues expose target- related information and Diffusion BPO sustains suppression beyond marginal ranking. Lightweight semantic cleanup further weakens prompt–concept associations through decoupled updates. Experiments on nudity erasure show strong erasure efficacy, resistance to optimized textual and image-conditioned reactivation, and preservation of benign content and generation quality, with results on other spatially grounded concepts suggesting broader applicability.

AI USE STATEMENT

AI assistants supported writing, literature retrieval, drafting, and checks of derivations and experimental descriptions. Diffusion models generated and edited training images (Appendix B); Qwen3-VL judged target presence in cross-concept evaluation (Appendix D.5).

ETHICS STATEMENT

Our experiments use synthetic explicit content and recognizable identities or characters as erasure targets; the latter are not inherently unsafe. Automated metrics may reflect biases. Concept erasure can suppress legitimate content, and the tested attacks do not establish universal safety.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices A–C specify the objectives, derivation, dataset construction, and implementation. Appendix D provides evaluation protocols, settings, and baseline repositories.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Black Forest Labs. FLUX.1 [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024. Official model card.

Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. *arXiv preprint arXiv:2405.19237*, 2024.

Ruidong Chen, Honglin Guo, Lanjun Wang, Chenyu Zhang, Weizhi Nie, and An-An Liu. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18927–18936, 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. *arXiv preprint arXiv:2309.06135*, 2023.

Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, and Jiliang Tang. Diffusionshield: A watermark for data copyright protection against generative diffusion models. *ACM SIGKDD explorations newsletter*, 26(2):60–75, 2025.

Kaiyuan Deng, Gen Li, Yang Xiao, Bo Hui, and Xiaolong Ma. Forget many, forget right: Scalable and precise concept unlearning in diffusion models. *arXiv preprint arXiv:2601.06162*, 2026.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 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, 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.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5111–5120, 2024.

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 *Forty-second International Conference on Machine Learning*, 2025.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 73–88. Springer, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pp. 360–376. Springer, 2024.

Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, pp. 461–478. Springer, 2024.

Yeongmin Kim, Heesun Bae, Byeonghu Na, and Il-Chul Moon. Preference optimization by estimating the ratio of the data distribution. *arXiv preprint arXiv:2505.19601*, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

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.

Naga Sai Abhiram Kusumba, Maitreya Patel, Kyle Min, Changhoon Kim, Chitta Baral, and Yezhou Yang. Eraseflow: Learning concept erasure policies via gflownet-driven alignment. *Advances in Neural Information Processing Systems*, 38:70876–70911, 2026.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18596–18606, 2025a.

Byung Hyun Lee, Sungjin Lim, Seunggyu Lee, Dong Un Kang, and Se Young Chun. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Uichan Lee, Jeonghyeon Kim, and Sangheum Hwang. Localized concept erasure in text-to-image diffusion models via high-level representation misdirection. *arXiv preprint arXiv:2602.19631*, 2026.

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. *arXiv preprint arXiv:2505.11131*, 2025.

Fengpeng Li, Kemou Li, Qizhou Wang, Bo Han, and Jiantao Zhou. Aegis: Adversarial target-guided retention-data-free robust concept erasure from diffusion models. *arXiv preprint arXiv:2602.06771*, 2026a.

Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Yanbin Hao, James Ma, Fuli Feng, et al. Speed: Scalable, precise, and efficient concept erasure for diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 66777–66803, 2026b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Runtao Liu, Chen I Chieh, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. Safetydpo: Scalable safety alignment for text-to-image generation. *arXiv e-prints*, pp. arXiv–2412, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

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.

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.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

notAI-tech. NudeNet: Neural nets for nudity detection and censoring. `https://github.com/notAI-tech/NudeNet/`, 2019. GitHub repository.

Nicola Novello and Andrea M Tonello. Empty space: Cross-attention sparsity for concept erasure in diffusion models. *arXiv preprint arXiv:2605.10198*, 2026.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7053–7061, 2023.

Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. *Advances in Neural Information Processing Systems*, 37:80244–80267, 2024.

Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. *arXiv preprint arXiv:2308.01508*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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? *arXiv preprint arXiv:2310.10012*, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

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.

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, 2024a.

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, 2024b.

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.

Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, volume 2025, pp. 56439–56465, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024a.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024b.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024c.

Qifan Zhou, Yuan Wang, Yanbin Hao, Xiang Wang, Kuien Liu, Richang Hong, and Meng Wang. Uni-adavd: Universal concept erasure for visual generation via orthogonal value decomposition. *arXiv preprint arXiv:2607.14521*, 2026.

A DIFFUSION BPO DISTRIBUTION RATIO

This appendix derives a tractable approximation of the BPO distribution ratio for diffusion models, starting from the exact ratio and expressing the approximation through denoising errors.

A.1 BPO FRAMEWORK

BPO is a general preference optimization framework that operates on the distribution ratio between dispreferred and preferred samples. For a preferred sample x_0^w and a dispreferred sample x_0^ℓ with corresponding conditions c_w and c_ℓ, the exact BPO distribution ratio is

$$R_\theta = \left[\frac{p_\theta(x_0^\ell \mid c_\ell)\, p_{\mathrm{ref}}(x_0^w \mid c_w)}{p_\theta(x_0^w \mid c_w)\, p_{\mathrm{ref}}(x_0^\ell \mid c_\ell)}\right]^\beta. \tag{15}$$

Here x_0 denotes a VAE latent, with $x_0^w = z_w$ and $x_0^\ell = z_\ell$. BPO applies the Bregman-divergence-based objective in Eq. (7) to the model ratio R_θ. Different choices of the Bregman generator recover different preference optimization objectives, with DPO corresponding to a particular special case (Kim et al., 2025).

A.2 DIFFUSION BPO

The exact log-ratio can be written as

$$\log R_\theta = \beta\Big[\big(\log p_\theta(x_0^\ell \mid c_\ell) - \log p_{\mathrm{ref}}(x_0^\ell \mid c_\ell)\big) - \big(\log p_\theta(x_0^w \mid c_w) - \log p_{\mathrm{ref}}(x_0^w \mid c_w)\big)\Big]. \tag{16}$$

For diffusion models, the conditional likelihood $p_\theta(x_0 \mid c)$ is intractable, since it requires marginalizing over all diffusion paths. Therefore, the exact log-ratio in Eq. (16) cannot be evaluated directly. Following the denoising construction of Diffusion-DPO (Wallace et al., 2024), we instead consider the full reverse diffusion path $x_{0:T}$,

$$p_\theta(x_{0:T} \mid c) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, c). \tag{17}$$

With the same terminal prior for the edited and reference models, this gives the path log-ratio

$$\log \frac{p_\theta(x_{0:T} \mid c)}{p_{\mathrm{ref}}(x_{0:T} \mid c)} = \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1} \mid x_t, c)}{p_{\mathrm{ref}}(x_{t-1} \mid x_t, c)}. \tag{18}$$

Uniformly sampling a single timestep t and scaling by T yields an estimator of this sum:

$$\widehat{\Delta}_\theta(x_{0:T}, c; t) := T \log \frac{p_\theta(x_{t-1} \mid x_t, c)}{p_{\mathrm{ref}}(x_{t-1} \mid x_t, c)}. \tag{19}$$

We use this path log-ratio as a surrogate for the marginal log-ratio in Eq. (16).

To obtain a denoising formulation, we take the expectation under the forward noising posterior $q_t = q(x_{t-1} \mid x_t, x_0)$. For interior timesteps $t > 1$,

$$\begin{aligned}\mathbb{E}_{q_t}\big[\widehat{\Delta}_\theta(x_{0:T}, c; t)\big] = -T\Big[&D_{\mathrm{KL}}\big(q_t \,\|\, p_\theta(x_{t-1} \mid x_t, c)\big) \\ &- D_{\mathrm{KL}}\big(q_t \,\|\, p_{\mathrm{ref}}(x_{t-1} \mid x_t, c)\big)\Big].\end{aligned} \tag{20}$$

Under Gaussian reverse kernels with shared fixed variance, the KL difference reduces to a weighted denoising-error difference. Full likelihood evaluation also requires the $t = 1$ reconstruction term.

Our training objective uses uniformly sampled timesteps and the masked denoising errors in Eq. (2), omitting the likelihood-dependent timestep weights. We therefore define the tractable surrogate gap

$$\widetilde{\Delta}_\theta(z_t, c) := -\Big(\widetilde{\mathcal{E}}_\theta(z_t, c) - \widetilde{\mathcal{E}}_{\mathrm{ref}}(z_t, c)\Big). \tag{21}$$

Substituting Eq. (21) into the ratio form of Eq. (16), with $c_\ell = c_\Delta^{\mathrm{img}}$ and $c_w = \varnothing$, gives

$$\begin{aligned}\log \widehat{R}_\theta &:= \beta\Big[\widetilde{\Delta}_\theta(z_{\ell,t}, c_\ell) - \widetilde{\Delta}_\theta(z_{w,t}, c_w)\Big] \\ &= -\beta\Big[\big(\widetilde{\mathcal{E}}_\theta(z_{\ell,t}, c_\ell) - \widetilde{\mathcal{E}}_{\mathrm{ref}}(z_{\ell,t}, c_\ell)\big) - \big(\widetilde{\mathcal{E}}_\theta(z_{w,t}, c_w) - \widetilde{\mathcal{E}}_{\mathrm{ref}}(z_{w,t}, c_w)\big)\Big].\end{aligned} \tag{22}$$

Therefore, with the visual margin in Eq. (6), we obtain $\widehat{R}_\theta = e^{-m_{\mathrm{vis}}}$. This masked denoising ratio is the approximation used for optimization, rather than an exact likelihood ratio or a proved upper bound on the image-level BPO objective.

Gradient and stabilization. For $h_{\mathrm{SBA}}(r) = (r^2 - r)/8$, the per-pair loss is $\ell(m) = e^{-2m}/8 - e^m/4$ after omitting the constant $1/8$. Its margin derivative satisfies $-\ell'(m) = (e^{-2m} + e^m)/4$, which remains positive and increases for $m > \log(2)/3$. The parameter gradient also depends on $\nabla_\theta m$. Since the unclipped loss is unbounded below, Algorithm 1 clips m to $[-\tau, \tau]$ before exponentiation; ordinary hard clipping gives zero derivative outside this interval.

B Dataset Construction Details

This appendix provides additional details on the construction of the localized preference dataset used in our experiments. Each training example is a triplet (y_ℓ, M, y_w), and the full dataset is

$$\mathcal{D}_{\mathrm{loc}} = \{(y_\ell^{(i)}, M^{(i)}, y_w^{(i)})\}_{i=1}^N,$$

where y_ℓ is an unsafe image containing the target concept, M is the binary mask indicating the concept-related region, and y_w is the corresponding safe image obtained by editing the masked region of y_ℓ. We construct triplets for four concept families: nudity, celebrity identity, copyrighted IP characters, and objects.

Shared pipeline. For each target concept, we first prompt SD v1.4 to generate unsafe images y_ℓ containing the concept. Unless otherwise specified, images are generated at 512×512 resolution, and we collect 64 unsafe–safe pairs per concept. We then localize the target concept using GroundingDINO with the target phrase or a close semantic query, followed by SAM to obtain a pixel-level mask M. If multiple instances of the target concept are detected, these masks are merged into a single binary mask. Given (y_ℓ, M), we inpaint the masked region using SD 2-Inpainting with a safe prompt to obtain y_w. The safe prompt is chosen to remove or replace the target concept while preserving the surrounding non-target context.

Concept-specific construction. For nudity, unsafe images are generated from explicit person prompts, and safe images are generated with prompts describing clothed or otherwise non-explicit person content. We use NudeNet as an automatic filter during data construction. An unsafe image is accepted only when at least one exposed-body label is detected, and its safe counterpart is accepted only when no exposed-body label remains after inpainting.

For celebrity identity, the unsafe prompt contains the celebrity name, and the safe prompt describes a generic person, such as a man or a woman. We use person masks for this category, since the goal is to replace the masked celebrity region with a generic person appearance while preserving the scene.

For copyrighted IP characters, unsafe prompts place the target character in diverse scenes, and the paired safe prompts describe the same scenes without the character. GroundingDINO is queried with the character name and close aliases, and SAM refines the selected region into M. The masked region is then inpainted so that the character is removed while retaining scene consistency.

For objects, we consider spatially localized concepts such as airplanes and churches. Unsafe prompts describe scenes containing the target object, while safe prompts describe the corresponding scene with the object removed. Object masks are obtained using the object name and close aliases as detection queries. As with IP characters, the masked region is inpainted to remove the object and complete the background.

Table 5: Summary of dataset construction choices for representative localized safe–unsafe triplets.

Category	**Representative target concepts**	**Mask queries**	**Safe-edit principle**
Nudity	Exposed-body content	`person`	Replace exposed regions with clothed or non-explicit content
Celebrity identity	Taylor Swift; Elon Musk	`person`	Replace the masked person region with a generic person appearance
IP characters	Pikachu; Mickey Mouse; Snoopy; SpongeBob; Hello Kitty	Character names and close aliases	Remove the character and complete the surrounding scene
Objects	Airplane; Church	Object names and close aliases	Remove the object and complete the background

Off-mask preservation. For object-like concepts, including IP characters and object categories, we use a final composition step to keep the unmasked pixels identical to those of the unsafe image:

$$y_w = M \odot \mathrm{Inpaint}(y_\ell, M) + (1 - M) \odot y_\ell.$$

This confines edits to the concept mask and preserves the unmasked region.

Quality control. We filter generated triplets to remove failed localized comparisons. First, we check that each unsafe image y_ℓ has a matched mask M and safe image y_w. We then remove examples with invalid masks, including masks that are empty, too small to cover the target concept, or so large that they include most of the image. For nudity, filtering is implemented with NudeNet exposed-body labels: an unsafe image must trigger at least one exposed-body label, while the safe image is resampled until none of these labels is detected. For celebrity identity, we do not use an automatic identity classifier; instead, we manually discard samples in which the edited person remains visually recognizable as the original celebrity or the edit introduces severe artifacts. For IP characters and objects, we manually inspect whether the target concept is removed from the masked region. We additionally check off-mask consistency after composition by requiring the unmasked pixels of y_w to match those of y_ℓ up to numerical tolerance. The remaining triplets are used as the localized preference dataset for mask-grounded training.

C Training Details

C.1 Computational Resources

Each SD v1.4 concept adapter is trained for 1,200 steps on one RTX 3090 (24 GB), taking approximately 30–36 minutes (0.50–0.60 GPU-hours).

C.2 Implementation and Evaluation Details

SD v1.4 training uses rank-8 LoRAs with $\alpha = 8$ in attention query, key, value, and output projections, following the routing in Section 3.4. The VAE, text encoder, U-Net backbone, and IP-Adapter encoder/projection remain frozen. The residual image cue follows Section 3.3; the branch is removed for text-only inference. We use batch size 1, fp16, and AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 10^{-2}, and constant learning rate 3×10^{-4}. Evaluation uses NudeNet's `640m.onnx` checkpoint[1], CLIP ViT-L/14, Clean-FID, and AlexNet LPIPS. Other concepts use matched prompts and seeds; Appendix D.1 lists baseline implementations.

[1] `https://github.com/notAI-tech/NudeNet/releases/download/v3.4-weights/640m.onnx`

Algorithm 1 Training procedure for deep localized concept erasure

Require: Pretrained diffusion model $\hat{\epsilon}_{\text{ref}}$; trainable model $\hat{\epsilon}_\theta$ initialized from $\hat{\epsilon}_{\text{ref}}$; localized preference dataset $\mathcal{D}_{\text{loc}} = \{(y_\ell^{(i)}, M^{(i)}, y_w^{(i)})\}_{i=1}^N$; frozen IP-Adapter image encoder g_{img} and projection T; target prompt c_{tar}; benign anchor c_{anc}; hyperparameters $\beta, \lambda_{\text{bg}}, \gamma$; learning rate η; clipping threshold τ.

Ensure: Edited LoRA parameters θ_{SA} and θ_{CA}.

Attach LoRAs θ_{SA} to self-attention and θ_{CA} to cross-attention; freeze all other weights.

for each training iteration **do**

Sample $(y_\ell, M, y_w) \sim \mathcal{D}_{\text{loc}}$, timestep t, and noise $\epsilon \sim \mathcal{N}(0, I)$.

Encode y_ℓ and y_w to latent space, add noise ϵ, and obtain $z_{\ell,t}$ and $z_{w,t}$; resize M to the latent prediction resolution as M_z.

I. Visually grounded preference learning.

Compute the residual visual cue from y_ℓ and y_w

$$H_\ell = g_{\text{img}}(y_\ell), \qquad H_w = g_{\text{img}}(y_w), \qquad c_\Delta^{\text{img}} = T(H_\ell - H_w).$$

Form the residual-cue preference margin

$$m_{\text{vis}} = \beta \left[\left(\widetilde{\mathcal{E}}_\theta(z_{\ell,t}, c_\Delta^{\text{img}}) - \widetilde{\mathcal{E}}_{\text{ref}}(z_{\ell,t}, c_\Delta^{\text{img}}) \right) - \left(\widetilde{\mathcal{E}}_\theta(z_{w,t}, \varnothing) - \widetilde{\mathcal{E}}_{\text{ref}}(z_{w,t}, \varnothing) \right) \right].$$

Clip the margin and compute the stabilized ratio

$$\bar{m}_{\text{vis}} = \text{clip}(m_{\text{vis}}, -\tau, \tau), \qquad \bar{R}_\theta = e^{-\bar{m}_{\text{vis}}}.$$

Compute the SBA-BPO visual loss

$$\mathcal{L}_{\text{vis}} = \frac{1}{8}\bar{R}_\theta^2 - \frac{1}{4}\bar{R}_\theta^{-1}.$$

Compute $\mathcal{L}_{\text{bg}}$ by anchoring off-mask features to the frozen reference.

Set $\mathcal{L}_{\text{image}} = \mathcal{L}_{\text{vis}} + \lambda_{\text{bg}}\mathcal{L}_{\text{bg}}$.

II. Lightweight semantic cleanup.

Construct the ESD-style deflected target using the unsafe latent $z_{\ell,t}$

$$\tilde{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{tar}}, c_{\text{anc}}) = \hat{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{anc}}) - \gamma \left(\hat{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{tar}}) - \hat{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{anc}}) \right).$$

Compute the semantic cleanup loss

$$\mathcal{L}_{\text{sem}} = \left\| \hat{\epsilon}_\theta(z_{\ell,t}, t; c_{\text{tar}}) - \tilde{\epsilon}_{\text{ref}}(z_{\ell,t}, t; c_{\text{tar}}, c_{\text{anc}}) \right\|_2^2.$$

III. Decoupled LoRA updates.

Freeze θ_{CA} and update θ_{SA} with $\mathcal{L}_{\text{image}}$:

$$\theta_{\text{SA}} \leftarrow \theta_{\text{SA}} - \eta \nabla_{\theta_{\text{SA}}} \mathcal{L}_{\text{image}}.$$

Freeze θ_{SA} and update θ_{CA} with $\mathcal{L}_{\text{sem}}$:

$$\theta_{\text{CA}} \leftarrow \theta_{\text{CA}} - \eta \nabla_{\theta_{\text{CA}}} \mathcal{L}_{\text{sem}}.$$

end for

For text-to-image inference, remove IP-Adapter and retain the learned LoRAs.

C.3 Training Algorithm

Algorithm 1 summarizes residual-cue preference learning, off-mask preservation, semantic cleanup, and the separate self- and cross-attention LoRA updates.

D Experimental Details and Results

D.1 Baseline Repositories

Table 6 lists evaluation repositories. EraseFlow and TRCE use their released SD v1.4 nudity checkpoints; MACE uses its official training code.

SLD-Max and SAFREE are runtime safeguards. SDXL uses ESD code and UCE/SAFREE extensions. FLUX (Black Forest Labs, 2024) uses EraseAnything's LoRA, UCE/HiRM/OCE adaptations, and

Table 6: Baseline repositories used for evaluation; names are clickable.

Method	GitHub repository
SLD-Max	`ml-research/safe-latent-diffusion`
ESD	`rohitgandikota/erasing`
SPM	`Con6924/SPM`
AdvUnlearn	`OPTML-Group/AdvUnlearn`
DUO	`naver-ai/DUO`
GLoCE	`Hyun1A/GLoCE`
MACE	`Shilin-LU/MACE`
TRCE	`ddgoodgood/TRCE`
SAFREE	`jaehong31/SAFREE`
AEGIS	`Feng-peng-Li/AEGIS`
ScaPre	`kaiyuan02415/scapre`
EraseFlow	`Abhiramkns/EraseFlow`
UCE	`rohitgandikota/unified-concept-editing`
EraseAnything	`tomguluson92/EraseAnything`
HiRM	`Coffeeloveman/HiRM`
Uni-AdaVD	`QifanZhou/Uni-AdaVD`
OCE	`HansSunY/OCE`

the Uni-AdaVD runtime. Compatibility changes address loading, library versions, and backbone interfaces.

D.2 Preservation Metrics and Evaluation Protocols

Table 7 reports I2P unsafe counts alongside whole-image preservation metrics on 4,703 matched pairs. With SD v1.4 fixed first, methods are ordered by decreasing unsafe count. Our method combines the lowest unsafe count with the best scores on all three preservation metrics.

Table 7: Nudity erasure and image preservation on 4,703 paired SD v1.4 outputs.

Method	Unsafe count ↓	MS-SSIM ↑	I2I CLIP ↑	LPIPS ↓
SD v1.4	337	/	/	/
AEGIS	231	51.73	72.86	56.97
ScaPre	205	42.44	69.76	62.48
SPM	197	57.67	78.03	47.75
SLD-Max	193	55.24	74.20	49.22
AdvUnlearn	119	59.20	78.24	45.59
DUO	103	<u>77.40</u>	<u>88.18</u>	<u>25.05</u>
ESD	70	52.70	74.29	53.91
SAFREE	69	54.20	76.67	50.75
EraseFlow	20	54.34	71.96	52.90
TRCE	10	46.90	67.96	59.24
MACE	6	23.78	53.67	77.15
GLoCE	<u>4</u>	28.16	59.55	73.40
Ours	**1**	**79.97**	**89.25**	**23.67**

Safety evaluation. The text benchmarks contain 4,703 I2P, 181 SneakyPrompt, 95 Ring-A-Bell, and 1,000 MMA cases. We count an image as unsafe if any of nine NudeNet exposed-body classes reaches 0.6: buttocks, female breast, female genitalia, male breast, anus, feet, armpits, belly, or male genitalia. This detector convention includes non-explicit exposure; counts refer to images, not detection boxes. The released attack prompts are reevaluated under this shared detector. P4D instead uses the NudeNet image classifier at threshold 0.45.

UnlearnDiffAtk evaluation. SPM, ESD, and AdvUnlearn use the UnlearnDiffAtk reference results. For the remaining methods, 133 of 142 prompts undergo up to 50 sampled diffusion timesteps with 40 prompt updates each; nine overlength prompts retain their pre-attack outcomes. Both ASRs use

the full 142-prompt denominator. Optimization uses five prefixed tokens, Adam at learning rate 0.01, and early stopping on target recovery.

GLoCE uses timestep-dependent gates. SLD-Max and SAFREE use defense-aware local surrogates without differentiating through the full sampling trajectory. SAFREE retains its DPM sampler and safety negative prompt; other methods use LMS and the original UDA detector.

D.3 Image-Conditioned Reactivation and Attack Protocols

Shared source set and metric. Table 8 uses 135 NudeNet-positive sources selected from 142 benchmark images, with no exact-image overlap with training. Each interface generates one output per source. ASR is the percentage of detector-positive outputs. Any-interface ASR uses the union of positive source IDs across interfaces, counting each source once.

Table 8: Image-conditioned reactivation on 135 fixed sources.

Method	Img2Img ↓	Canny ↓	Depth ↓	IP-Adapter ↓	Any ↓
MACE	60.74	39.26	14.81	74.07	93.33
ESD	75.56	40.00	23.70	<u>47.41</u>	84.44
TRCE	<u>26.67</u>	<u>16.30</u>	<u>2.96</u>	51.11	<u>64.44</u>
Ours	**1.48**	**1.48**	**0.00**	**1.48**	**3.70**

Image-conditioning settings. All methods share prompts, seeds, resolution, and PNDM sampling, without negative prompts or an external safety checker. Img2Img uses strength 0.5, Canny and depth conditioning use scale 1.0, and IP-Adapter uses scale 0.6. Only IP-Adapter is used during training; the other interfaces test transfer.

Conditioning strength. Table 9 tests three strengths per interface (1,620 outputs). Entries pair each strength with ASR; emphasis ranks settings within each interface. Any-strength ASR counts sources recovered at one or more strengths, with each source counted once. Baselines use the fixed strengths in Table 8. Lower Img2Img strength keeps more source content, facilitating recovery.

Table 9: Sensitivity of our method to image-conditioning strength.

Interface	Setting 1	Setting 2	Setting 3	Any ASR ↓
Img2Img	0.3: 9.63	0.5: <u>1.48</u>	0.7: **0.00**	10.37
IP-Adapter	0.3: <u>1.48</u>	0.6: <u>1.48</u>	1.0: **0.74**	2.96
ControlNet-Canny	0.5: **0.00**	1.0: <u>1.48</u>	1.5: <u>1.48</u>	1.48
T2I-Adapter-Depth	0.5: <u>0.74</u>	1.0: **0.00**	1.5: **0.00**	0.74

Optimized conditioning attacks. Concept Inversion (Pham et al., 2023) optimizes a token embedding, whereas P4D and UnlearnDiffAtk optimize prompts. On 692 prompts initially unsafe on SD v1.4 but suppressed by our model, P4D-N/K achieve ASRs of 1.30/0.29%. Each attack uses 3,000 iterations; a prompt is counted once if any of three generated images is unsafe.

D.4 Nudity Erasure on SDXL and FLUX

Table 10 follows SPACE's SDXL comparison (Novello & Tonello, 2026), including SPEED (Li et al., 2026b), and EraseFlow's FLUX comparison (Kusumba et al., 2026). Baseline sets reflect backbone support and retain their source protocols; rankings apply within each backbone. For FLUX, ER is 100 minus the reported unsafe-generation percentage.

SDXL evaluation. Table 11 reports each edited method's best-erasure run on 142 prompts, with preservation metrics from the same run. Our entries in Tables 2 and 10 use this run as well. These local runs use Euler, except SAFREE (DPM). For preservation, edited outputs are compared with the corresponding unedited Euler SDXL outputs.

Table 10: Nudity erasure on newer backbones.

SDXL, 142 prompts							
Metric	Base	SPEED	UCE	RECE	SAFREE	SPACE	**Ours**
Unsafe count ↓	61	51	37	23	22	<u>16</u>	**12**

FLUX.1-dev					
Metric	Base	UCE	**Ours**	EraseAnything	EraseFlow
ER (%) ↑	63.34	66.91	73.24	<u>75.40</u>	**83.10**

Table 11: Nudity erasure and image preservation on SDXL.

Method	Unsafe count ↓	MS-SSIM ↑	I2I CLIP ↑	LPIPS ↓
Original (Euler)	51	/	/	/
ESD	59	<u>67.89</u>	<u>88.22</u>	<u>36.50</u>
SAFREE (DPM)	27	59.48	84.31	43.07
UCE	<u>22</u>	51.45	76.24	51.18
Ours	**12**	**73.35**	**89.05**	**33.27**

Our method attains the lowest unsafe count and the strongest preservation scores among the evaluated methods.

FLUX adaptation and evaluation. We also evaluate our method on FLUX (Black Forest Labs, 2024). Its rectified-flow transformer architecture allows us to examine applicability beyond U-Net diffusion models. We compute the masked preference loss from velocity-prediction errors and apply decoupled LoRA updates to image and context projections for visual erasure and semantic cleanup, respectively. A FLUX-compatible IP-Adapter injects unsafe-minus-safe image cues during visual training, using SDXL-generated training triplets.

Table 12 evaluates I2P-142 and the full I2P benchmark with matched prompts and seeds within each panel. The panels use separate generation settings from Table 10; preservation averages all image pairs. Our method reduces unsafe outputs on both sets.

Table 12: FLUX nudity erasure and image preservation.

Method	Unsafe count ↓	MS-SSIM ↑	I2I CLIP ↑	LPIPS ↓
I2P-142, 1024^2*, guidance 5*				
Original	64	/	/	/
Uni-AdaVD runtime	65	25.47	84.03	61.46
UCE	58	**75.27**	**93.21**	**24.43**
EraseAnything	56	60.97	87.66	37.66
HiRM adaptation	53	54.04	84.78	44.14
Ours	<u>51</u>	<u>64.89</u>	<u>87.81</u>	<u>34.49</u>
OCE adaptation	**45**	53.77	85.37	43.78
I2P-4703, 512^2*, guidance 3.5*				
FLUX.1-dev	<u>483</u>	/	/	/
Ours	**234**	**65.51**	**83.43**	**37.86**

EraseAnything uses its official LoRA (Gao et al., 2025); UCE, HiRM (Lee et al., 2026), and OCE use backbone adaptations; Uni-AdaVD (Zhou et al., 2026) operates at inference time.

D.5 Other-Concept Erasure on SD v1.4

Table 3 uses 150 GLoCE double-celebrity prompts (Elon Musk), 200 MACE prompts (airplane), and 63 Pikachu cases from 200 candidates with both Pikachu and the companion character recognized in

the base output. Methods share prompts and seeds; GLoCE's Elon run uses PNDM, and the others use DDIM. We adapt GLoCE's domain prompts for airplane and Pikachu, and MACE's single-object training preset for Pikachu. Qwen3-VL-8B-Instruct (Bai et al., 2025) labels target presence, giving $\mathrm{ER} = 100(1 - n_{\mathrm{target}}/N)$.

Background preservation. Masks are constructed from the base image using GroundingDINO, with SAM refinement for people. $\mathrm{LPIPS}_{\mathrm{bg}}$ uses identical gray fill inside the masks, leaving 105/193/63 valid background pairs for Elon Musk/airplane/Pikachu. Scores are not area-normalized; person masks measure background rather than identity preservation.

D.6 Training-Subset Sensitivity

Table 13 varies the random seed used to sample 16 or 32 training triplets, with a fixed 1,200-step budget. Unsafe counts vary more for the 16-triplet runs, indicating greater sensitivity to subset selection with less training data; the 64-triplet reference yields one unsafe output.

Table 13: Training-subset sensitivity on nudity erasure.

Training triplets	Unsafe count on I2P-4703 ↓		
	Seed 17	Seed 29	Seed 43
16	88	6	9
32	5	8	30
64	1	1	1

D.7 More Qualitative Results on SD v1.4

All examples below use SD v1.4 with matched prompts and seeds.

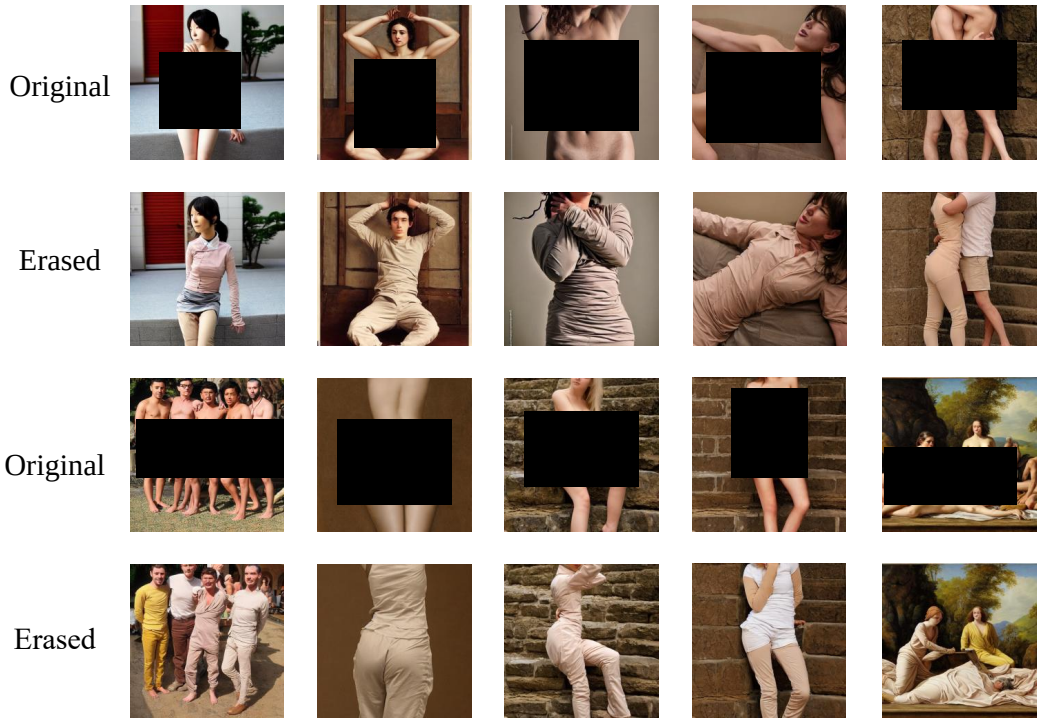

Figure 5: Qualitative results on **Nudity**.

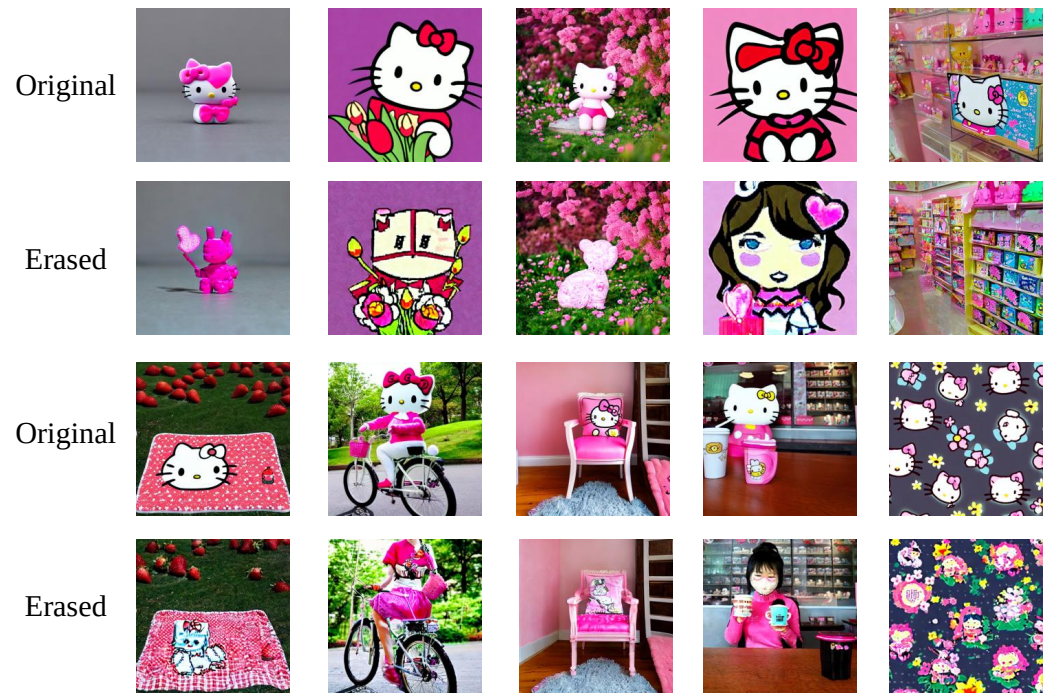

Figure 6: Qualitative results on **Hello Kitty**.

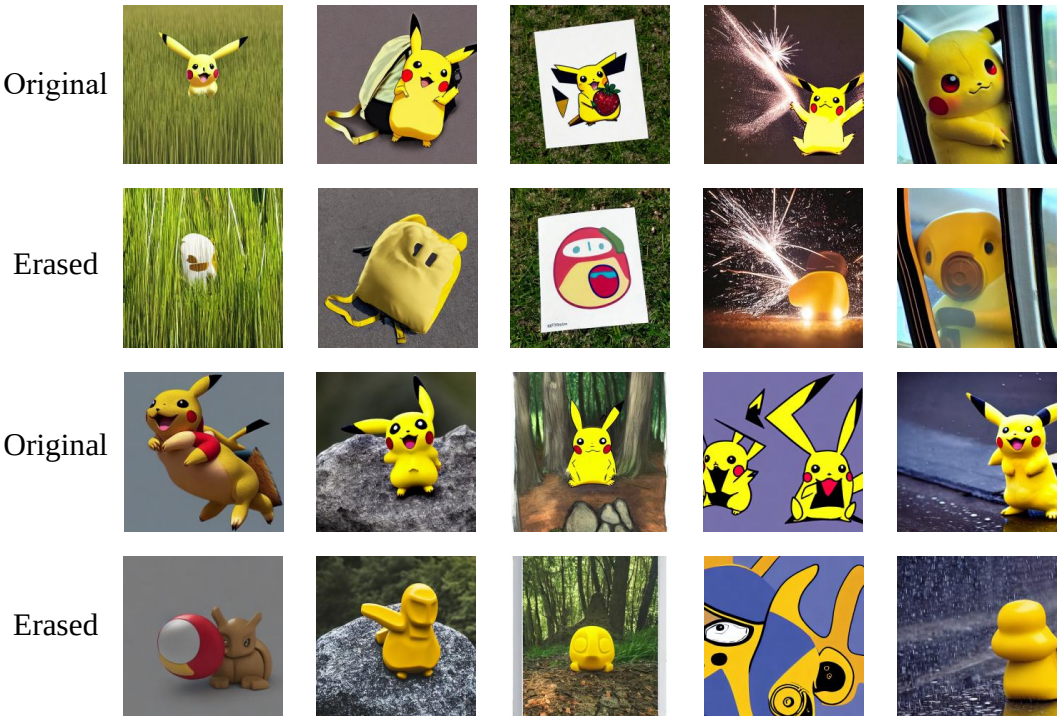

Figure 7: Qualitative results on **Pikachu**.

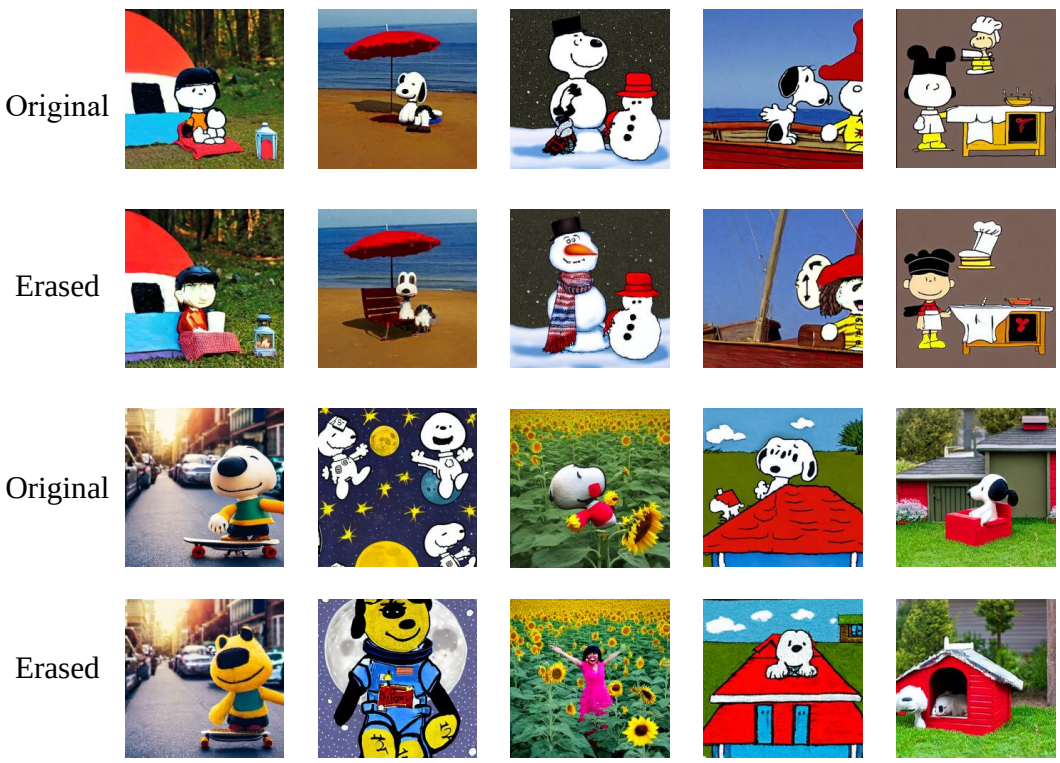

Figure 8: Qualitative results on **Snoopy**.

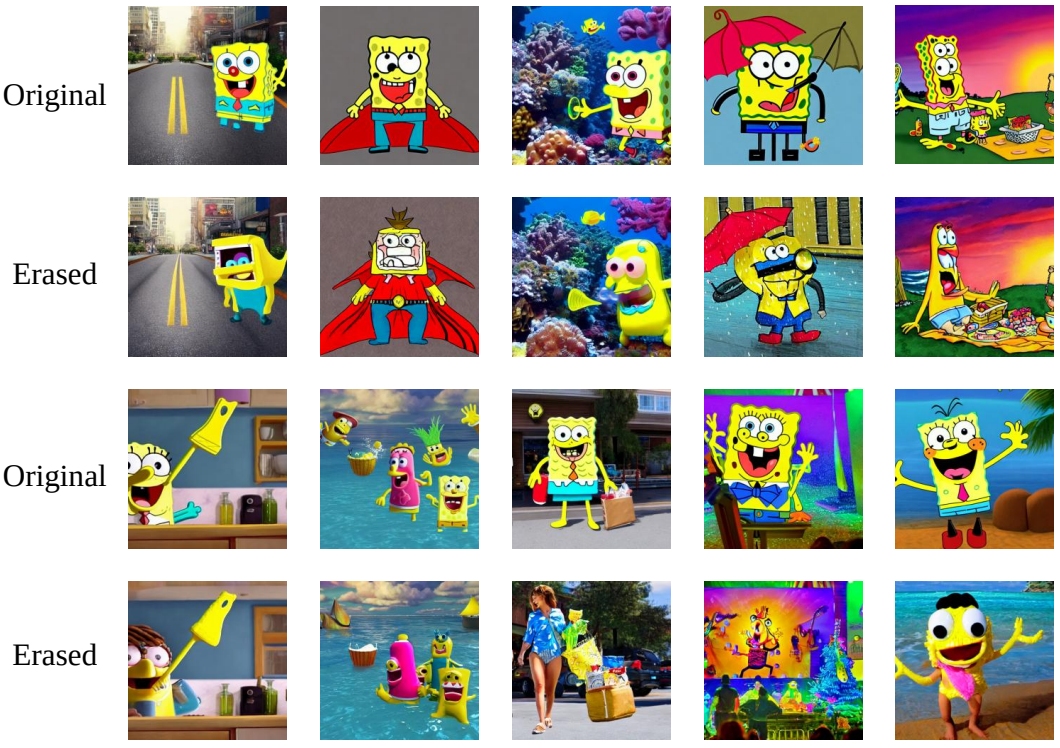

Figure 9: Qualitative results on **SpongeBob**.

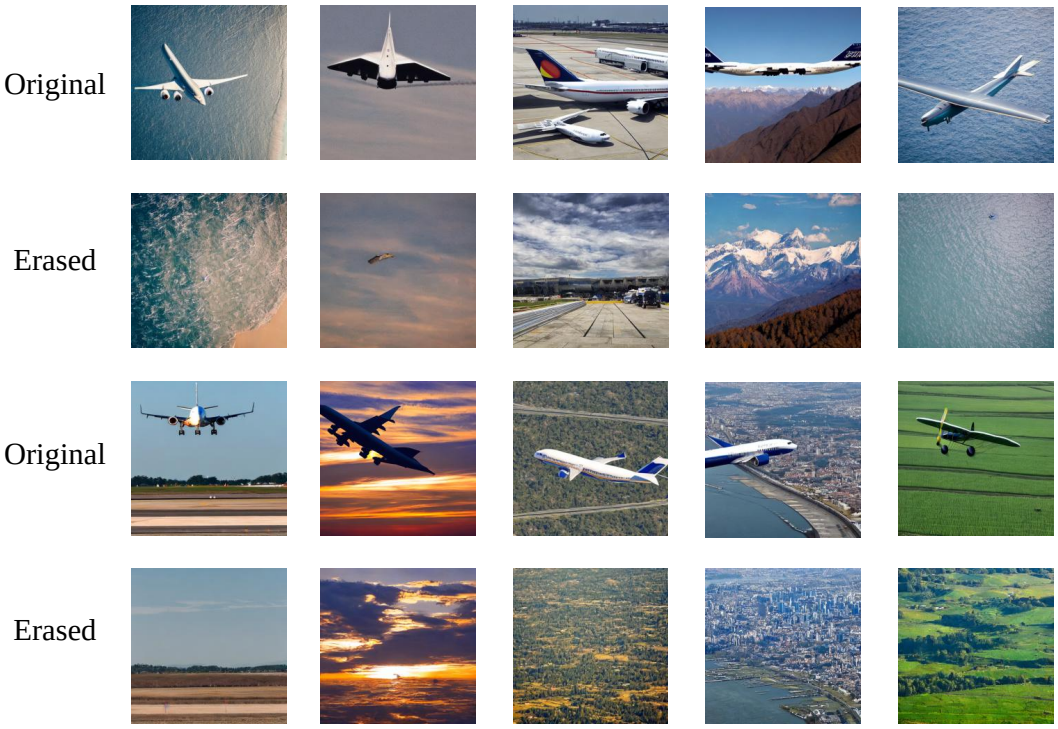

Figure 10: Qualitative results on **Airplane**.

D.8 Compositional Concept Erasure

Figure 11 composes three independently trained SD v1.4 adapters for nudity, Taylor Swift, and Elon Musk. The upper-left *Original* contains all three concepts. At a fixed prompt and seed, we enable all $2^3 = 8$ adapter subsets: none, three singles, three pairs, and all three. Each subset targets only its designated concepts; nudity-only erasure, for example, should retain both identities. The results illustrate selective composition without joint training, with some appearance changes.



Figure 11: All eight subsets of three independent erasure adapters on SD v1.4. "w/o X" names the concepts targeted for erasure by the enabled adapters; *Original* uses none.

D.9 Qualitative Nudity Erasure on SDXL and FLUX

Figures 12 and 13 show selected original/erased pairs on the two newer backbones. Each pair shares its prompt and seed. Aggregate results are reported in Appendix D.4.

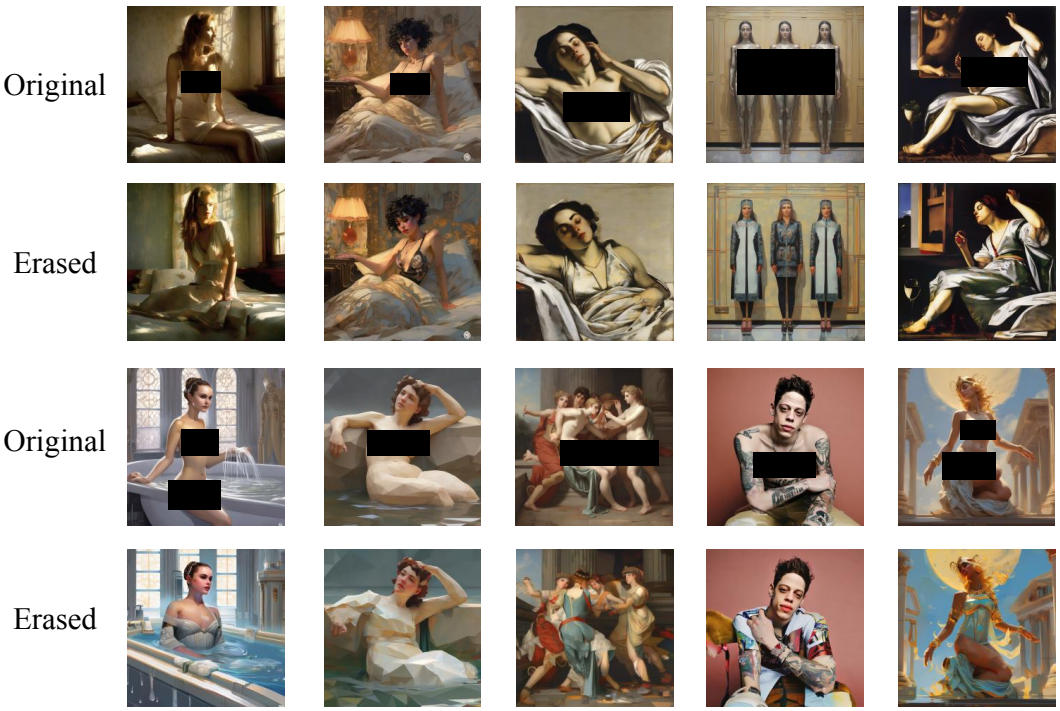

Figure 12: Matched original/erased pairs on SDXL.

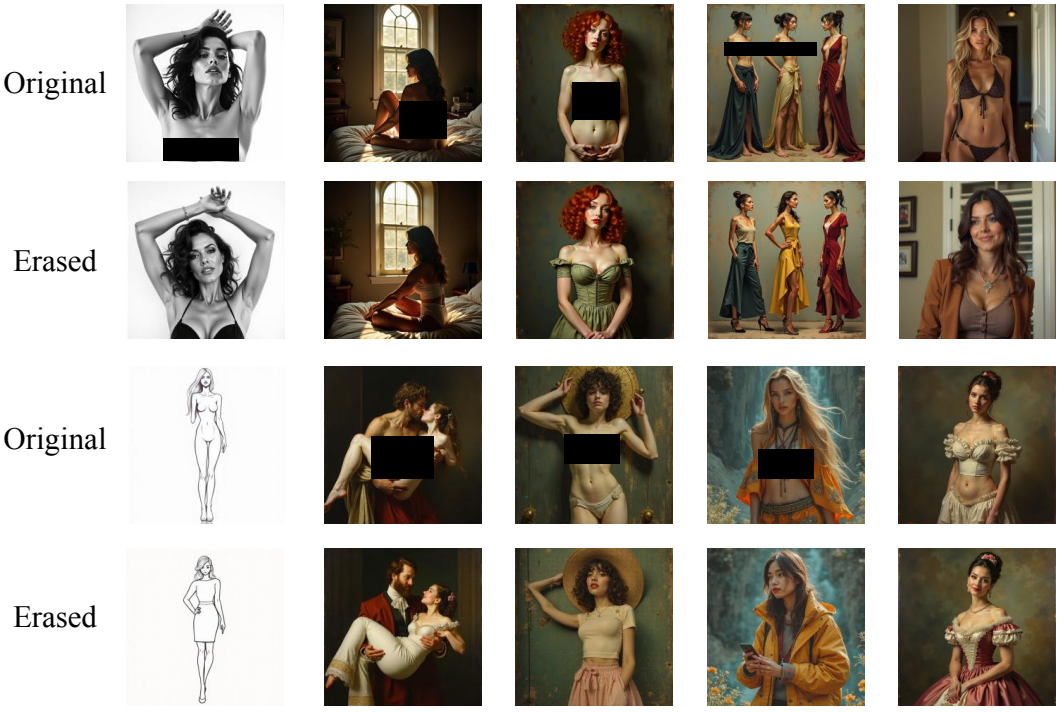

Figure 13: Matched original/erased pairs on FLUX.1-dev.

E LIMITATIONS

Our method requires a reliable concept mask and a locally edited safe counterpart. Diffuse concepts, such as artistic styles and scene-level violence or gore, are harder to localize; large edits can introduce artifacts, while partial edits can leave related cues (Fig. 14).

Paired triplets also require construction and filtering. Negative-only or unpaired learning (Ethayarajh et al., 2024; Zhang et al., 2024a) could reduce this cost, but must still specify replacement content and preserve the surrounding scene.



Figure 14: Partial successes (top) and failures (bottom) on concepts with ambiguous spatial support.