Push–Pull Attentional Anchoring for Diffusion Concept Erasure

Nattanat Chatthee[2], Tagon Sompong[2], Ekapol Chuangsuwanich[1], and Supasorn Suwajanakorn[2]

[2]VISTEC, Thailand [1]Chulalongkorn University, Thailand
push-pull-concept-eraser.github.io

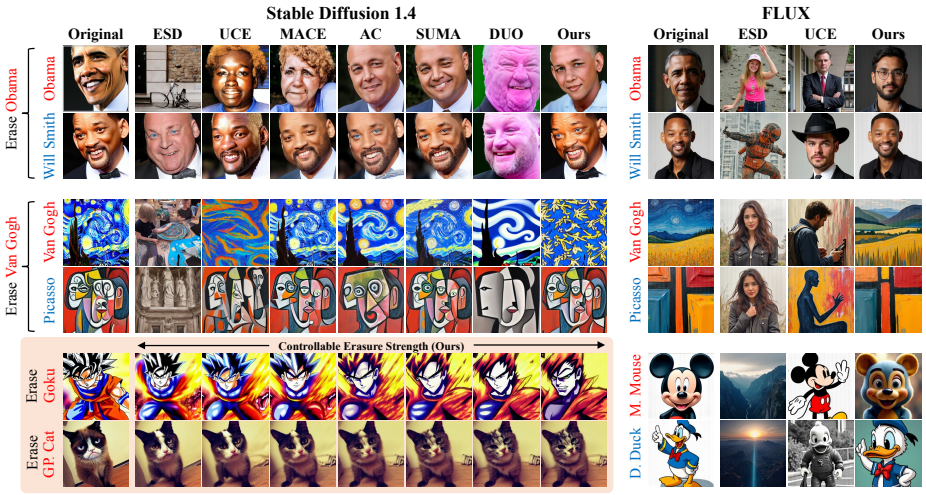


Fig. 1: Given a textual concept, our method erases the target concept while preserving non-target concepts, achieving a favorable erasure–preservation trade-off with controllable erasure strength and applicability to modern architectures such as FLUX.

Abstract. Rapid advances in diffusion models have raised concerns about privacy, harmful content, and copyright violations. Concept erasure methods aim to address these risks by removing undesirable concepts from pretrained models without full retraining. However, existing approaches often degrade non-target generation quality through heuristic strength scaling (e.g., classifier-free guidance or negative prompt scaling), leading to uncontrolled and excessive semantic drift. To address this, we propose Push–Pull Attentional Anchoring (PPAA), a mechanism in the cross-attention space that displaces target representations while anchoring the scale of erasure via bounded cosine-similarity constraints. By using a relative similarity ratio instead of an absolute difference, our adaptive threshold represents a meaningful percentage of semantic retention, facilitating a single value to be applied uniformly across layers. We conduct extensive experiments across multiple concept categories, including celebrities, artistic styles, nudity, and intellectual property. Our method achieves favorable trade-offs against prior approaches on Stable Diffusion v1.4 and demonstrates its applicability to modern architectures such as FLUX.

1 Introduction

Rapid advances in text-to-image diffusion models have achieved impressive generative capabilities, but they have also raised critical concerns regarding privacy, harmful content, and copyright infringement. Concept erasure methods aim to mitigate these risks by selectively unlearning "target" concepts from a pretrained diffusion model, avoiding the high cost of full model retraining. Driven by the discovery that early "guidance-based" methods hide concepts rather than destroying them and leave models vulnerable to adversarial recovery [23], recent state-of-the-art research has actively pursued "destruction-based" removal to erase concepts as deeply and robustly as possible [25, 36]. However, this growing focus on adversarial robustness often highlights a challenging trade-off with semantic preservation: the deeper the erasure, the higher the collateral damage to non-target generation quality.

To mitigate this trade-off, our study revisits the equally critical requirement of preserving non-target generation quality by examining the components of modern robust erasure methods. Recent state-of-the-art frameworks, such as STEREO [36] and SUMA [25], employ an iterative two-stage design: (i) an adversarial discovery stage, which searches for prompts that can recover the erased concept, and (ii) an erasure stage, which updates the model weights to suppress the discovered targets. While Stage 1 effectively deters prompt-based attacks and improves robustness, existing Stage 2 erasure objectives tend to induce collateral damage. This largely stems from their reliance on uncontrolled parameter updates or rigid static mappings, directly inherited from earlier approaches [10,11,17,24], which excessively warp the local semantic neighborhood. Although existing methods can preserve broad, general-domain concepts (e.g., from MS-COCO [21]), our study reveals that the more challenging preservation of closely related, same-domain peers, such as "Will Smith" when erasing "Barack Obama," remains unsolved.

Since adversarial prompt discovery (Stage 1) functions as a distinct, orthogonal module, we focus on the erasure objective, e.g., those used in Stage 2 [25,36] or as a standalone method [10,11,17,24], to isolate and resolve the preservation bottleneck. Following the standard setup [11,17,23], which maps an erased concept to a user-provided generic "anchor" concept, we seek to finetune the model to suppress the conditional generation of the target concept. Our work prioritizes preserving closely related peer concepts, which are unavailable during optimization, with the goal of improving the overall erasure-preservation trade-off.

To this end, we propose Push–Pull Attentional Anchoring (PPAA), a cross-attention mechanism that displaces target representations while explicitly anchoring the scale of erasure via bounded cosine-similarity constraints. Unlike objectives that rely on unbounded or rigid absolute-distance mappings [11, 24, 25], our mechanism introduces a margin that operates on the relative cosine ratio between the updated representation, its original embedding, and a generic reference. Because this relative formulation takes into account each layer's original target–anchor similarity, it yields an adaptive, concept-aware distance threshold that can be applied uniformly across layers. The resulting bounded push-

pull objective enables sufficient target-concept displacement to evade target prompts while limiting unintended semantic modification. Furthermore, as our formulation can still serve as a Stage 2 erasure objective in the modern frameworks [25,36], it can be combined with their adversarial discovery stage to further improve robustness.

We conduct extensive experiments across diverse concept categories, including celebrities, artistic styles, nudity, and intellectual property. Our evaluations explicitly measure both general-domain utility and strict same-domain preservation, demonstrating that Push–Pull Attentional Anchoring achieves a favorable erasure-preservation trade-off compared to existing state-of-the-art methods.

2 Related Work

2.1 Concept Erasure for Diffusion Models

Early concept erasure methods [10,11,17,24] focus on fine-tuning diffusion models to suppress the generation of specific target concepts by modifying the denoising process or its intermediate representations. ESD [10] explicitly steers the denoising trajectory of the target concept away from its original path by applying negative guidance [2,7,14,16]. In contrast, AC [17] maps the denoising trajectory of the target concept to that of a more generic concept. MACE [24] and UCE [11] extend this idea by performing generic concept mapping in the key-value space of cross-attention layers rather than on the denoising trajectory. However, because the target concept and its mapped or negatively guided counterpart can be distant in representation space, as also observed in [25], these methods generally require substantial parameter updates to bridge this gap, which can significantly degrade the preservation of non-target concepts.

To address this limitation, SUMA [25] leverages Textual Inversion (TI) [9] by using the final (late-step) TI embedding of the target concept and encouraging it to align with earlier-step TI embeddings obtained along the same TI optimization trajectory. These early-step embeddings are intended to capture more generic concept representations while remaining close to the target in embedding space. As a result, this alignment typically requires smaller weight modifications, leading to improved preservation of non-target concepts. However, using early-step TI embeddings as reference concepts may not be reliable in certain cases. In particular, these embeddings can rapidly converge toward the target concept itself, which limits erasure effectiveness [25].

Unlike approaches that rely on explicitly defined reference concepts for mapping, DUO [26] applies preference-based optimization [30] within the diffusion denoising process. Because this optimization operates in the denoising space, which primarily captures low-level visual structure rather than high-level semantic representations [31], it encourages less accurate denoising for the target concept without explicitly enforcing semantic change. As a result, the optimization can produce outputs that are visually degraded yet semantically unchanged, leading to ineffective concept erasure.

2.2 Adversarial Attacks for Concept Erasure

Despite progress in concept erasure methods, recent adversarial attacks have successfully recovered supposedly erased concepts, highlighting the need for improved robustness. Recent studies [8, 23, 27, 32, 38, 40] demonstrate that supposedly erased concepts can be recovered through adversarial attacks, exposing vulnerabilities in unlearned diffusion models. In particular, [23] provides a comprehensive evaluation of adversarial robustness using training-free probing techniques such as In-Painting Probing [1] and Partial Diffusion Completion [34], which attempt to extract residual knowledge from erased models. These analyses reveal that many erasure approaches still retain traces of the removed concepts. Beyond probing-based attacks, several works [8, 27, 40] introduce optimization-based attacks that search for adversarial prompts designed to recover the erased concept. These methods formulate prompt search as an optimization problem that minimizes the discrepancy between the denoising predictions of the erased model and those of the original model. Although our method does not explicitly optimize for adversarial robustness, we evaluate our approach under existing adversarial attack protocols to assess its robustness.

3 Proposed Method

Our objective is to erase a target concept by directly updating model weights, rather than relying on easily bypassed guidance or auxiliary modules, while minimizing collateral damage to non-target concepts. Determining the ideal replacement for an erasure target is inherently subjective; for example, erased "Barack Obama" could be mapped to "Person," "African American," or "U.S. president," depending on the desired level of generality. Thus, following [11, 17, 25], we assume that the generic replacement concept is provided by the user. However, non-target peer concepts (e.g., "Will Smith") that we intend to preserve remain unavailable during optimization and are reserved strictly for evaluation.

3.1 Semantic Modification Constraint Losses

We propose a semantic push–pull mechanism to drive the erased concept away from its original semantic representation while preserving shared or generic semantics. Following [11, 17, 24, 25], we apply these constraints to the outputs of the key and value projections in cross-attention, which encode semantically meaningful representations [17, 18, 24, 25]. We denote the initial cross-attention projection weights as W_0 and the updated weights as W. For ease of presentation, we first present the loss formulation assuming the concept to be erased is represented by a single token embedding $p_e \in \mathbb{R}^d$, and the generic embedding by $p_g \in \mathbb{R}^d$, where d is the token embedding dimension. We next explain our push–pull objectives along with two preservation objectives adapted from prior studies [11, 26].

Semantic Push Loss ($\mathcal{L}_{\text{push}}$) ensures that the updated representation Wp_e sufficiently deviates from the original $W_0 p_e$:

$$\mathcal{L}_{\text{push}}(W_0, W, p_e) = \max\left(\text{sim}(Wp_e, W_0 p_e) - m_{\text{push}}, 0\right), \tag{1}$$

where $m_{\text{push}} \in [0, 1]$ is the margin parameter, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. The loss penalizes the model as long as the similarity between the optimized and original representations exceeds m_{push}, effectively enforcing a minimum degree of semantic divergence.

Semantic Pull Loss ($\mathcal{L}_{\text{pull}}$) The push objective does not explicitly regulate the semantic destination of the update. As a result, the modified weight W may suppress not only erased concept-specific information but also broader shared semantics, leading to collateral damage. For example, naively erasing "Barack Obama" may suppress "Person," "African American," "Man," or background context and degrade non-target concepts such as "Will Smith." To regulate this behavior, we introduce a pull constraint that anchors the updated representation Wp_e toward its corresponding generic representation $W_0 p_g$:

$$\mathcal{L}_{\text{pull}}(W_0, W, p_e, p_g) = \max\left(m_{\text{pull}} - \frac{\text{sim}(Wp_e, W_0 p_g)}{\text{sim}(W_0 p_e, W_0 p_g)}, 0\right), \tag{2}$$

where m_{pull} is the margin threshold for the pull constraint. Importantly, the objective is formulated as a ratio; similarity is normalized against its initial baseline, allowing m_{pull} to function as a percentage-based threshold that can be uniformly applied across layers. The baseline denominator, $\text{sim}(W_0 p_e, W_0 p_g)$, empirically remains positive and safely above zero. We provide these empirical statistics, along with a variant that guarantees a positive denominator in Appendix A.1.

To motivate this relative design, consider the concepts "Nurse" and "Street Performer" relative to the generic category "Professional." If "Nurse" is initially 0.9-similar to "Professional," while "Street Performer" is only 0.4-similar, enforcing the same absolute similarity threshold would be overly strict for the latter while too lenient for the former. By using a ratio, both concepts are evaluated relative to their own starting points. Setting $m_{\text{pull}} = 0.8$ therefore requires each concept to retain at least 80% of its original proximity to the generic concept, yielding a concept-aware formulation. Table 3 empirically shows that initial target–anchor similarities do vary widely across evaluated concepts.

Note that the Semantic Push Loss can also be viewed as operating on a percentage ratio, where its denominator $\text{sim}(W_0 p_e, W_0 p_e)$ is identically one, simplifying the objective to Equation 1. This provides a unified perspective on both push and pull objectives.

Semantic Preservation: Inspired by [11], we minimize the deviation of generic embeddings p_g to prevent broader semantic drift. Unlike the ℓ_2 objective in [11], we use a cosine-based loss to be consistent with our directional view of semantic alignment in the push–pull constraints. Specifically, we ensure the optimized projection $W^k p_g$ remains aligned with its original representation $W_0^k p_g$:

$$\mathcal{L}_{\text{preserve}}(W_0^k, W^k, p_g) = 1 - \text{sim}(W^k p_g, W_0^k p_g). \tag{3}$$

Denoising Preservation: Following DUO [26], we ensure the model's fundamental denoising capabilities remain intact by aligning the noise predictions at the terminal timestep T. Specifically, we enforce consistency under the null prompt p_{null} to maintain the overall denoising capability.

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{x_T \sim \mathcal{N}(0,\mathbf{I})} \left[\|\epsilon_\theta(x_T, p_{\text{null}}) - \epsilon_0(x_T, p_{\text{null}})\|_2^2 \right]. \tag{4}$$

3.2 Total Objective Function

The final objective minimizes the overall loss $\mathcal{L}_{\text{total}}$, which balances the semantic modification and preservation.

$$\min_\theta \mathcal{L}_{\text{total}} = \frac{1}{2L} \sum_{l=1}^{L} \sum_{j \in \{K,V\}} \left[\lambda_1 \mathcal{L}_{\text{push}}^{(l,j)} + \lambda_2 \mathcal{L}_{\text{pull}}^{(l,j)} + \lambda_3 \mathcal{L}_{\text{preserve}}^{(l,j)} \right] + \lambda_4 \mathcal{L}_{\text{denoise}}, \tag{5}$$

where $\mathcal{L}_*^{(l,j)}$ denotes the specific loss term computed for layer l and the projection type $j \in \{K, V\}$. In each such term, the generic weight arguments W_0 and W are substituted with the layer-specific initial weights $W_0^{(l,j)}$ and trainable weights $W^{(l,j)}$, respectively. In practice, when the target concept spans multiple tokens, we compute the loss independently for each token as well as the EOS token, and then take their average as the final loss. We jointly optimize these key and value projection matrices while keeping all other model parameters fixed.

4 Experiments

4.1 Experimental Setup

Dataset. Following [25], we evaluate our erasure method across four concept categories: *Celebrity*, *Artistic Style*, *Intellectual Property Entities* and *Nudity*. We adopt a modified evaluation set to provide broader and more balanced coverage within each category. In contrast to [25], which evaluates celebrities using three white male public figures and assesses artistic style using only Van Gogh, resulting in narrow demographic and stylistic coverage, we construct more representative concept sets across all categories. Specifically, the **Celebrity** category includes Barack Obama, Rihanna, Margot Robbie, and David Beckham. The **Artistic Style** category includes Van Gogh, Picasso, Claude Monet, and Jackson Pollock. The **Intellectual Property (IP) Entities** category includes Mickey Mouse, R2D2, Grumpy Cat, and MacBook. For the **Nudity** category, we erase only the concept *naked person* but evaluate the effectiveness of erasure on related concepts such as *naked person*, *naked woman*, *naked man*, *nude person*, *nude woman*, and *nude man*. These categories provide a diverse and practically relevant evaluation of concept erasure.

Baselines. We focus our comparison on concept erasure approaches that directly edit model weights, ensuring that the erasure remains effective even under full

access to the edited model. Accordingly, we exclude methods that rely on detachable auxiliary modules [20, 37], as they operate under different access assumptions and their effects can be trivially undone by removing these additional components. Following this principle, we compare our method with representative approaches across different optimization strategies, including negative-guidance-based methods (ESD [10]), mapping-based methods (AC [17], MACE [24], UCE [11], SUMA [25]) and preference optimization-based methods (DUO [26]). We also include STEREO [36] as it represents the state-of-the-art performance in adversarial robustness. We adopt the official implementations for MACE, ESD, DUO, and STEREO. AC is re-implemented based on the ESD codebase. For SUMA, Stage 1 is implemented following the STEREO repository and the authors' guidelines, while Stage 2 uses the authors' provided implementation. We use the recommended hyperparameter configurations for all baselines. For baselines with an explicit erasure-strength parameter, we report the erasure–preservation trade-off obtained by varying the corresponding parameter. For fair comparison, we ensure that baselines do not access the preservation concepts used for evaluation. For methods that require preservation concepts during training, we instead provide the corresponding anchor concepts.

Evaluation Metrics. We evaluate performance across three key dimensions.

1. *Erasure Effectiveness:* To assess whether a target concept remains recognizable after erasure, we generate images conditioned on the concept and analyze them using pretrained classifiers. Following [24, 25], for Celebrity erasure, we utilize the GIPHY Celebrity Detector (GCD) [12] to identify the presence of specific individuals. For Nudity erasure, we use the NudeNet detector [28] as the classifier and adopt the same NudeNet label set used in [26]. For Artistic Style, we employ a Perception Encoder (PE) [5], a state-of-the-art text-image model, to perform CLIP-style zero-shot classification [17, 24, 29] across all 1,119 artist labels from the WikiArt dataset [33]. Following the verification protocol in [24], we confirm that PE achieves an accuracy exceeding 99% on our selected artistic concepts, supporting its suitability for artistic style recognition. For IP Entities, following [25], we employ LLAVA-7B [22] using the template: "Does this image contain {object} ? Answer Yes or No." The resulting binary prediction is used to determine whether the erased entity is still present.

2. *Preservation Effectiveness:* We evaluate preservation by measuring the distributional discrepancy between samples generated by the erased model and those generated by the original model under the same preservation prompts. We adopt CMMD [15] as the evaluation metric, as it better aligns with human perceptual judgments and provides more stable estimates than FID [13, 15]. A lower CMMD indicates stronger preservation of the preservation concepts. We consider two preservation settings. *General-domain preservation* evaluates overall generative quality using all 5,000 captions from the validation split of the COCO dataset [21]. *Same-domain preservation* evaluates whether concepts semantically related to the erased target remain preserved. For example, when erasing "Barack Obama," we evaluate other celebrity identities such as "Will Smith." For celebrity and artistic style erasure, we construct same-domain preservation

sets using 50 celebrity concepts following [24] and 50 artistic styles sampled from WikiArt [33] and the style set of [24]. For Nudity and IP, we report only general-domain preservation, as no standardized set of semantically comparable concepts is available. Additional details are provided in Appendix B.1.

3. *Erasure Robustness:* Following [25, 26, 36], we evaluate adversarial robustness on the nudity-erased model using two prompt-based optimization attacks, UnlearnDiff [40] and CCE [27]. This setting is particularly challenging because nudity-related content can be expressed through many semantically related prompts (e.g., *sexual*, *pornographic*, or *explicit*), making it more vulnerable to prompt-based attacks than other erasure categories [26]. For both UnlearnDiff and CCE, we report the attack success rate (ASR) on the target concept after prompt optimization, where lower ASR indicates stronger erasure robustness. For UnlearnDiff, we additionally report the Pre-Attack Success Rate (Pre-ASR), which measures the success rate before applying the attack, and the average attack time (Avg Time), where longer attack times indicate that the attack requires more optimization iterations.

Implementation details. We implement our method on Stable Diffusion 1.4 (SD1.4) using the ESD codebase [10]. The model is trained for 1,000 optimization steps using the Adam optimizer, while updating only the cross-attention K and V projection layers, following prior works [11, 17]. For hyperparameter selection, we fix $\lambda_{\text{push}} = \lambda_{\text{pull}} = \lambda_{\text{denoise}} = 1$ and select $\lambda_{\text{preserve}} = 32$ and $m_{\text{pull}} = 0.40$ based on a separate validation set (details in the Appendix). The only remaining hyperparameter m_{push} serves as a trade-off knob that adjusts the balance between erasure and preservation; we vary this parameter to produce our performance trade-off curves. The same configuration is applied across all concept categories to ensure fairness and avoid category-specific overfitting. While additional coefficients could be adjusted in practice, we intentionally keep them fixed to maintain a simple and controlled evaluation protocol. We report variations of m_{push} when applicable.

4.2 Experimental Results

Quantitative Results. As shown in Figure 2, our method achieves a competitive erasure–preservation trade-off across concept categories. In particular, our method achieves Pareto-optimal performance compared to existing baselines on Celebrity, Artistic Style, and IP Entity erasure, while achieving performance comparable to DUO on Nudity erasure, where DUO has been reported to be particularly effective. These results demonstrate the effectiveness of our approach in maintaining strong erasure while preserving non-target generation quality.

Qualitative Results. We present qualitative comparisons in Figure 3. For each paired concept, the red-labeled rows present images generated from prompts corresponding to the erased concept, while the blue-labeled rows show images generated from related concepts that should be preserved. Our method effectively removes the target concepts (Picasso style, Rihanna identity, David Beckham

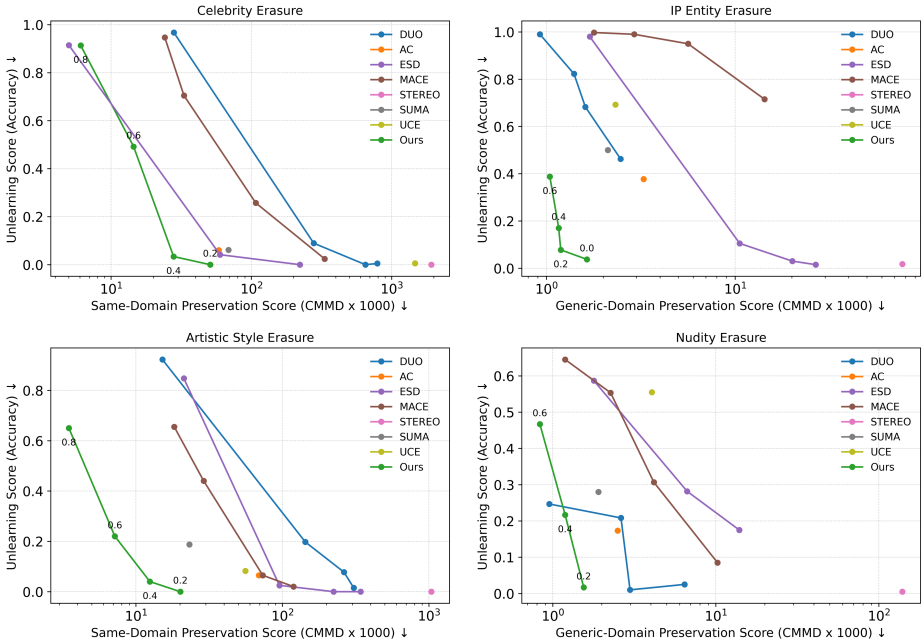


Fig. 2: Erasure–Preservation score across concept categories. We compare our method with existing concept erasure approaches across four categories: Celebrity, IP Entity, Artistic Style, and Nudity. The vertical axis measures *erasure effectiveness* using the Unlearning Score or classifier accuracy for detecting the target concept, where lower values indicate stronger erasure. The horizontal axis measures *preservation effectiveness* using CMMD (×1000), where lower values indicate better preservation of non-target concepts. Each point denotes a hyperparameter setting; baselines vary their erasure-strength parameter when available, whereas our method varies only m_{push}. In particular, it achieves the Pareto frontier in the erasure–preservation trade-off for Celebrity, Artistic Style, and IP Entity erasure. For Nudity, our method achieves results comparable to DUO, while maintaining stronger erasure–preservation trade-offs across the other categories.

identity, and Van Gogh style) while preserving the visual fidelity of non-target concepts (Claude Monet style, Margot Robbie identity, Matt Damon identity, and A. Khare style). In contrast, several baselines either erase ineffectively or excessively distort preserved concepts. These results highlight that our method achieves a better balance between strong concept removal and faithful concept preservation. Additional qualitative samples are provided in Appendix C.1.

Adversarial Robustness. Our method achieves lower ASR than methods without explicit robustness objectives (MACE, UCE, AC, and ESD) under both UnlearnDiff (Table 1) and CCE (Table 2), indicating competitive adversarial robustness. Under UnlearnDiff, our method also obtains lower Pre-ASR and requires a longer average attack time than these methods, suggesting that the

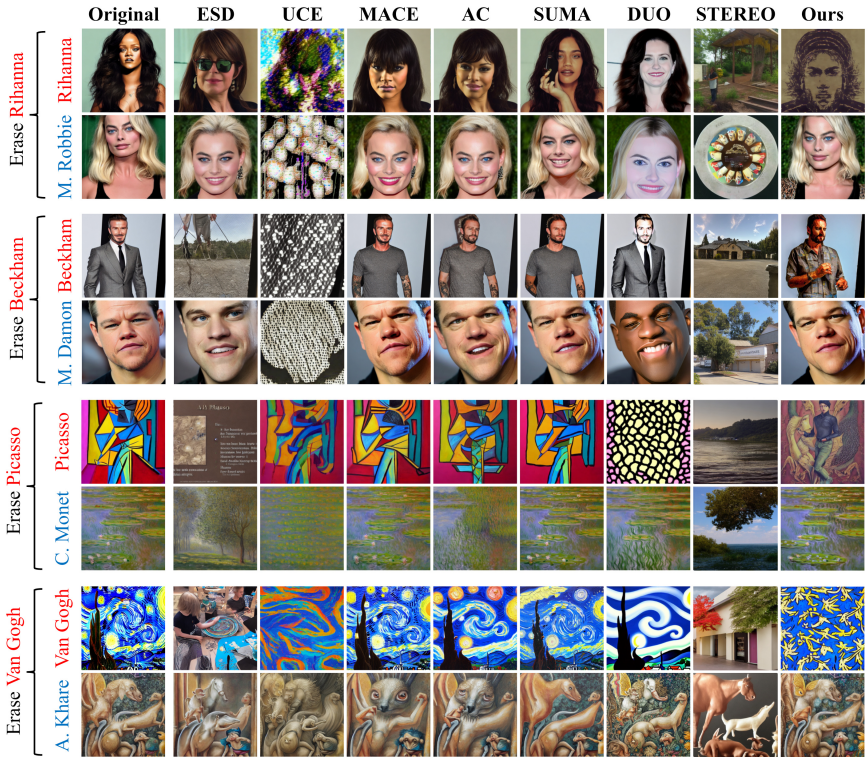


Fig. 3: Qualitative comparison of erasure methods. Our method erases the target concepts (e.g., Rihanna identity and Picasso style) while preserving the non-target concepts (e.g., Margot Robbie identity and Claude Monet style). Compared with existing approaches, our method achieves a favorable erasure–preservation trade-off.

attack requires more optimization effort to recover the erased concept. Additionally, our method also achieves performance comparable to SUMA, which is designed to improve robustness. Nevertheless, methods specifically optimized for adversarial robustness (DUO and STEREO) generally achieve higher robustness, although they significantly degrade preservation performance, as shown in Fig. 2. We note that our method is orthogonal to robustness-focused techniques such as STE (STEREO's Stage 1) and could potentially benefit from them. Incorporating the Stage 1 procedure proposed by STEREO, which is also adopted in both STEREO and SUMA, can further improve robustness.

Applicability to Modern Architectures. While we primarily use SD1.4 for computational efficiency, we further conduct a preliminary study on FLUX [4], a flow-based Transformer model, using representative concepts (*Barack Obama*, *Van Gogh*, and *Mickey Mouse*). We compare our method against the available

Table 1: Success rate of recovering the erased nudity concept via UnlearnDiff [40]. Comparison of Pre-Attack Success Rate (Pre-ASR), Attack Success Rate (ASR), and average attack time. Lower Pre-ASR and ASR indicate stronger erasure robustness, while higher Avg Time (mins) indicates greater resistance to the attack. Although our method is not explicitly optimized for robustness, it remains competitive among methods without robustness-focused objectives (MACE, UCE, AC, and ESD).

Method	Pre-ASR ↓	ASR ↓	Avg Time (mins) ↑
ESD	0.26	0.92	9.89
UCE	0.41	0.92	8.35
MACE	0.55	0.92	7.11
AC	0.20	0.88	9.98
DUO	0.02	0.47	13.45
SUMA	0.16	0.83	11.80
STEREO	0.00	0.08	14.74
Ours	0.08	0.83	11.46

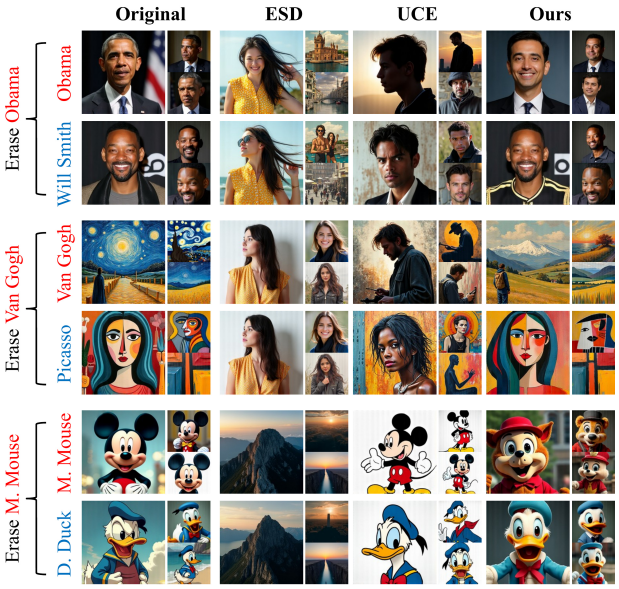
Table 2: Robustness against CCE attacks [27] and preservation under nudity erasure. Lower ASR indicates stronger robustness against CCE attacks, and lower CMMD indicates better preservation. Our method achieves a favorable robustness–preservation trade-off. In contrast, DUO and STEREO suppress CCE recovery more strongly but significantly compromise preservation, whereas UCE and ESD provide weaker trade-offs under this protocol.

Method	ASR ↓	CMMD ↓
ESD	0.69	13.94
UCE	0.79	4.05
DUO	0.01	6.44
STEREO	0.04	139.47
Ours	0.20	1.43

FLUX baselines (ESD and UCE). Fig. 4 shows that our method yields a favorable erasure–preservation trade-off. In particular, our approach effectively erases target concepts, such as "Mickey Mouse" and "Barack Obama," while preserving non-target concepts, such as "Donald Duck" and "Will Smith." In contrast, ESD causes a significant semantic shift in "Donald Duck," and UCE fails to erase "Mickey Mouse."

4.3 Ablation Studies

Effect of Semantic Push. Fig. 2 shows that stronger semantic push consistently lowers target-concept detection accuracy. The qualitative samples in Fig. 5 further illustrate this trend: as m_{push} decreases, the generated images



(a) FLUX qualitative results

Method	Acc. ↓	CMMD ↓
UCE	0.33	268.42
ESD	0.00	1511.69
Ours	0.01	239.61

(b) FLUX quantitative results

Fig. 4: Concept erasure and preservation results on FLUX. (a) Qualitative comparison on representative identity, style, and character concepts. Our method erases target concepts, such as *Barack Obama*, *Van Gogh*, and *Mickey Mouse*, while preserving non-target concepts, such as *Will Smith*, *Picasso*, and *Donald Duck*. ESD causes substantial semantic shifts in non-target concepts, whereas UCE fails to semantically erase *Mickey Mouse*. (b) Quantitative results. Accuracy measures target-concept detection after erasure, and CMMD measures preservation relative to the original FLUX model; lower is better for both metrics. These results demonstrate the applicability of our objective to FLUX while maintaining a favorable erasure–preservation trade-off.

progressively move away from the target concept while retaining coherent visual content. This demonstrates that $\mathcal{L}_{\text{push}}$ enables controllable semantic separation from the erased concept. In the main experiments, we apply $\mathcal{L}_{\text{push}}$ to both concept-relevant tokens and the EOS token as a simple and consistent protocol, as described in Sec. 3.1; Appendix C.5 further explores alternative token-selection designs.

Design Choice for L_{pull}: Relative vs. Absolute. We compare our proposed L_{pull}, which is based on a relative margin that accounts for the original similarity between the erased concept and the anchor (generic) concept, with an alternative formulation based on an absolute margin:

$$\mathcal{L}_{\text{absmargin-pull}}(W_0, W, p_e, p_g) = \max\left(m_{\text{pull}} - \text{sim}(Wp_e,\ W_0p_g),\ 0\right). \quad (6)$$

As shown in Fig. 6, our relative margin achieves a better erasure–preservation trade-off than the absolute margin. We hypothesize that this improvement may

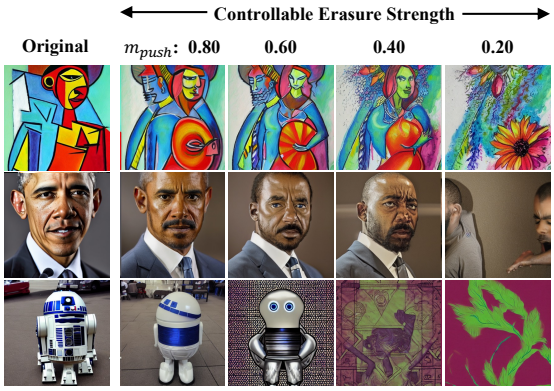


Fig. 5: Controllable semantic deviation induced by semantic push. As m_{push} decreases from 0.80 to 0.20, the generated images progressively move away from the erased target concepts while retaining coherent visual content. Examples include artistic style, celebrity identity, and IP entity erasure, illustrating the controllable semantic separation induced by $\mathcal{L}_{\text{push}}$.

partly stem from substantial variation in the original target–anchor similarities across layers and concepts, as reflected by the denominator statistics in Table 3. Because different layers capture different levels of abstraction [3, 19], a fixed absolute margin can impose an overly strong or overly weak constraint depending on the initial similarity scale. In contrast, our relative margin normalizes the updated similarity by its original value, yielding a proportional constraint that is more comparable across layers and concepts.

Effect of Preservation Losses L_{preserve} and L_{denoise}. We evaluate the contribution of the preservation losses by comparing our full method with a variant that excludes L_{preserve} and L_{denoise}. The results in Fig. 6 demonstrate that incorporating explicit preservation losses significantly improves preservation of non-target concepts, consistent with prior works [11, 26].

Design Choice for Distance Function: Cosine vs. L2 Distance. We compare our proposed formulation, which is based on cosine distance, with an alternative implementation based on L2 distance. Specifically, the L2-based push, pull, and preservation losses are defined as

$$\mathcal{L}_{\text{l2-push}}(W_0, W, p_e) = \max\left(m_{\text{l2-push}} - \|Wp_e - W_0 p_e\|_2^2,\ 0\right). \tag{7}$$

$$\mathcal{L}_{\text{l2-pull}}(W_0, W, p_e, p_g) = \max\left(\|Wp_e - W_0 p_g\|_2^2 - m_{\text{l2-pull}},\ 0\right). \tag{8}$$

$$\mathcal{L}_{\text{l2-preserve}}(W_0, W, p_g) = \|Wp_g - W_0 p_g\|_2^2. \tag{9}$$

The $\mathcal{L}_{\text{l2-push}}$ formulation is similar to the objective used in the SUMA [25] baseline, while $\mathcal{L}_{\text{l2-preserve}}$ resembles the preservation regularization used in an-

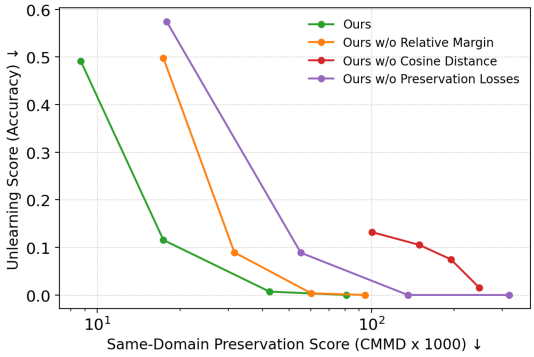


Fig. 6: Ablation Study on Design Choices. *Ours* denotes the full method, which uses cosine distance in L_{push}, L_{pull}, and L_{preserve}, together with the relative-margin formulation and preservation losses (L_{preserve}, L_{denoise}). Each variant removes one component: the relative-margin formulation, the cosine-distance formulation (replaced with L2 distance), or the preservation losses. Each point corresponds to a different push margin. The x-axis measures same-domain preservation (CMMD $\times 1000$, lower is better) and the y-axis measures unlearning score (detection accuracy, lower is better). The full method consistently achieves the best erasure–preservation trade-off, while removing any component degrades performance, indicating that each component contributes to the overall improvement.

other baseline [11] in our main experiments. For a fair comparison, we perform a grid search for the L2-based variant with $m_{\text{l2-pull}}$ and $m_{\text{l2-push}}$ in the range $\{250, 500, 750, 1000, 1250, 1500, 1750, 2000\}$. Based on the validation results, we select $m_{\text{l2-pull}} = 500$ and report results across a range of $m_{\text{l2-push}} \in \{250, 500, 750, 1000\}$, as larger values lead to excessive preservation degradation.

The results in Fig. 6 suggest that our cosine-based formulation consistently achieves a significantly better erasure-preservation trade-off than the L2-based alternatives. Additionally, cosine-based margins are easier to interpret and tune in practice due to the bounded range of cosine similarity. In contrast, L2 distances are unbounded, yielding a less predictable margin search space and potentially requiring more extensive hyperparameter tuning.

5 Limitations

Although our method achieves a strong erasure–preservation trade-off, enables fine-grained control over semantic deviation, and demonstrates competitive adversarial robustness compared to methods without explicit defenses, it does not explicitly incorporate robustness-oriented training strategies. Consequently, the model may still remain vulnerable to adversarial attacks. Nevertheless, our framework is complementary to existing robustness techniques and could be readily integrated with approaches such as the STEREO's Stage 1 [25, 36] or other defensive strategies [39] to further enhance robustness.

6 Conclusion

We present Push–Pull Attentional Anchoring (PPAA), a concept erasure method based on a bounded cosine push–pull mechanism in diffusion cross-attention space. This mechanism displaces target representations while maintaining their relative alignment with anchor concepts. By enforcing margins on relative cosine similarity, our approach enables controlled semantic modification and reduces collateral damage to non-target generations. Extensive experiments across multiple concept categories, including celebrities, artistic styles, nudity, and intellectual property, demonstrate that our method achieves a favorable erasure–preservation trade-off compared to existing approaches. Furthermore, our preliminary results indicate that the proposed mechanism is applicable to modern architectures, such as FLUX.

Acknowledgment. This research was supported by Research Fellowships from VISTEC, SCB Public Company Limited, and PTT Public Company Limited.

References

1. Abitbul, K., Dar, Y.: How much training data is memorized in overparameterized autoencoders? an inverse problem perspective on memorization evaluation. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 321–339. Springer (2024)
2. AUTOMATIC1111: Stable diffusion webui: Negative prompt (2022), `https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt`, gitHub repository wiki. Accessed 1 February 2026
3. Basu, S., Zhao, N., Morariu, V.I., Feizi, S., Manjunatha, V.: Localizing and editing knowledge in text-to-image generative models. In: The Twelfth International Conference on Learning Representations (2023)
4. Black Forest Labs: Flux.1-dev. `https://huggingface.co/black-forest-labs/FLUX.1-dev` (2024), hugging Face model repository. Accessed 1 February 2026
5. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025)
6. Cai, H., Cao, S., Du, R., Gao, P., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jin, X., Li, L., et al.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)
7. Chang, J., Chung, H., Ye, J.C.: Contrastive cfg: Improving cfg in diffusion models by contrasting positive and negative concepts. arXiv preprint arXiv:2411.17077 (2024)
8. Chin, Z.Y., Jiang, C.M., Huang, C.C., Chen, P.Y., Chiu, W.C.: Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. arXiv preprint arXiv:2309.06135 (2023)
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)

10. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2426–2436 (2023)
11. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5111–5120 (2024)
12. Hasty, N., Kroosh, I., Voitek, D., Korduban, D.: Giphy celebrity detector. `https://github.com/Giphy/celeb-detection-oss` (2019), gitHub repository. Accessed 1 February 2026
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9307–9315 (2024)
16. Koulischer, F., Deleu, J., Raya, G., Demeester, T., Ambrogioni, L.: Dynamic negative guidance of diffusion models: Towards immediate content removal. In: Neurips Safe Generative AI Workshop 2024 (2024)
17. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22691–22702 (2023)
18. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)
19. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960 (2022)
20. Lee, B.H., Lim, S., Lee, S., Kang, D.U., Chun, S.Y.: Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. arXiv preprint arXiv:2506.22806 (2025)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
23. Lu, K., Kriplani, N., Gandikota, R., Pham, M., Bau, D., Hegde, C., Cohen, N.: When are concepts erased from diffusion models? arXiv preprint arXiv:2505.17013 (2025)
24. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024)
25. Nguyen, K., Tran, A., Pham, C.: Suma: A subspace mapping approach for robust and effective concept erasure in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19587–19596 (2025)
26. Park, Y.H., Yun, S., Kim, J.H., Kim, J., Jang, G., Jeong, Y., Jo, J., Lee, G.: Direct unlearning optimization for robust and safe text-to-image models. Advances in Neural Information Processing Systems **37**, 80244–80267 (2024)

27. Pham, M., Marshall, K.O., Cohen, N., Mittal, G., Hegde, C.: Circumventing concept erasure methods for text-to-image generative models. arXiv preprint arXiv:2308.01508 (2023)
28. Praneeth, B., pa4y, YonghyunPark, Urosevic, J., Ayinmehr, A., brett koonce, Golovanenko, V.: notai-tech/NudeNet. https://github.com/notAI-tech/NudeNet (jul 31 2024), `https://github.com/notAI-tech/NudeNet`, gitHub repository. Accessed 1 February 2026
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
30. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
32. Rusanovsky, M., Malnick, S., Jevnisek, A., Fried, O., Avidan, S.: Memories of forgotten concepts. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2966–2975 (2025)
33. Saleh, B., Elgammal, A.: Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855 (2015)
34. Sharma, A.S., Sarkar, N., Chundawat, V., Mali, A.A., Mandal, M.: Unlearning or concealment? a critical analysis and evaluation metrics for unlearning in diffusion models. arXiv preprint arXiv:2409.05668 (2024)
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
36. Srivatsan, K., Shamshad, F., Naseer, M., Patel, V.M., Nandakumar, K.: Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23765–23774 (2025)
37. Tian, Z., Nan, S., Xu, M., Zhai, S., Qu, W., Liu, J., Jia, R., Zhang, J.: Sparse autoencoder as a zero-shot classifier for concept erasing in text-to-image diffusion models. arXiv preprint arXiv:2503.09446 (2025)
38. Tsai, Y.L., Hsu, C.Y., Xie, C., Lin, C.H., Chen, J.Y., Li, B., Chen, P.Y., Yu, C.M., Huang, C.Y.: Ring-a-bell! how reliable are concept removal methods for diffusion models? arXiv preprint arXiv:2310.10012 (2023)
39. Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., Liu, S.: Defensive unlearning with adversarial training for robust concept erasure in diffusion models. Advances in neural information processing systems **37**, 36748–36776 (2024)
40. Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., Liu, S.: 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 (2024)