CONCEPT UNLEARNING VIA CROSS-ATTENTION ACTIVATION SUBSPACES IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing closed-form methods for concept unlearning in text-to-image diffusion models typically derive editing directions from fixed text embeddings, which may not fully capture how concepts are expressed across latent states, timesteps, and layers. To capture this variation, we investigate cross-attention activations collected during denoising. In controlled probing experiments using the same anchor prompts, activation-derived bases achieve approximately five times the recall of text-derived bases on held-out prompts expressing the target concepts. Based on this finding, we propose Cross-Attention Subspace Erasure (`CASE`), a closed-form method that constructs layer-specific forget and retain subspaces from cross-attention activations. These subspaces define a retain-constrained linear operator incorporated directly into cross-attention weights, requiring no gradient-based fine-tuning or additional inference-time computation. Across ten concepts spanning four categories, `CASE` achieves the highest harmonic-mean score among evaluated baselines in all four categories, balancing suppression, retention, adversarial robustness, and generation quality. Further experiments demonstrate robustness to recovery attacks and a favorable suppression-retention trade-off when jointly unlearning up to 100 artistic styles. The benefits of activation-derived editing also extend to larger diffusion models, including SDXL and FLUX.

1 INTRODUCTION

Text-to-image diffusion models (Rombach et al., 2022) have become standard tools for synthesizing photorealistic images from natural-language prompts (Podell et al., 2024; Saharia et al., 2022; Ramesh et al., 2022). Trained on web-scale data, these models can also reproduce content that a deployed system may need to restrict, including specific artistic styles, depictions of public figures, and unsafe imagery (Schuhmann et al., 2022; Qu et al., 2023; Kumari et al., 2023). Concept unlearning aims to remove such concepts from a pretrained model without retraining it from scratch (Gandikota et al., 2023; Lu et al., 2024; Biswas et al., 2025; Wang et al., 2025a).

Concept unlearning requires both strong target suppression and selective preservation. The target concept should remain suppressed across diverse prompt formulations, including paraphrases and adversarial inputs. At the same time, the model's behavior on non-target concepts should remain close to that of the original model. These two objectives are inherently coupled because an edit that aggressively suppresses the target can also interfere with semantically related concepts. A practical unlearning method must therefore achieve a favorable balance between forgetting and retention.

Recent closed-form methods (Gandikota et al., 2024; Biswas et al., 2025; Orgad et al., 2023; Gong et al., 2024; Wang et al., 2025a) address this trade-off through a single deterministic weight update without gradient-based fine-tuning. However, a common design choice among closed-form methods is to derive editing directions from prompt-level text representations, which remain fixed throughout denoising. Cross-attention activations, in contrast, depend on the evolving latent state, timestep, and layer. As a result, a text-derived editing basis may fail to cover the different ways in which the same concept is expressed during generation.

This observation motivates a simple question: should the editing basis be constructed from the cross-attention activations observed during denoising? Using the same anchor prompts, we compare bases formed from text-encoder representations and from cross-attention activations collected during denoising. A controlled probing experiment shows that the activation basis achieves approximately

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

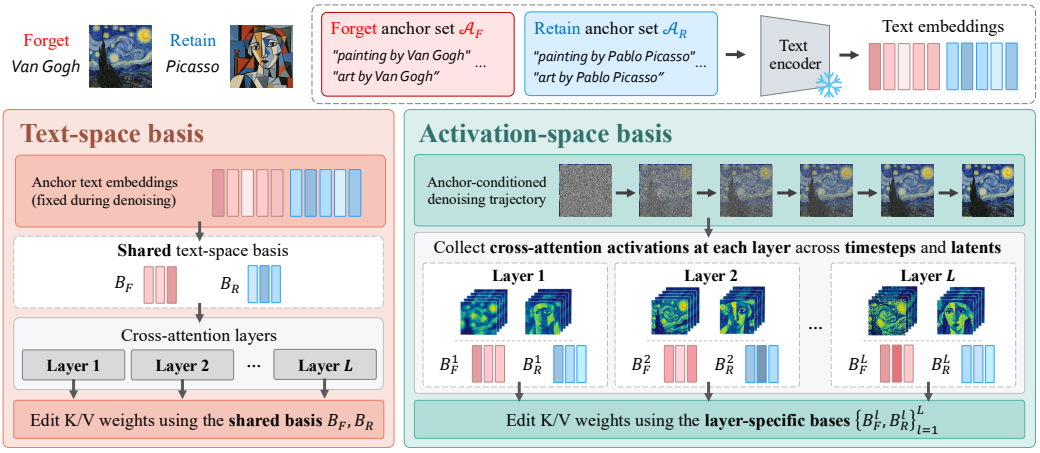


Figure 1: Comparison between text and activation bases. Prior closed-form methods build a shared text-space basis from anchor prompt embeddings. CASE instead captures layer-specific cross-attention activations during denoising and constructs an activation-space basis for each cross-attention layer.

five times the average recall of the text basis across the Style, IP, and Celebrity categories (Figure 2). This suggests that activation-space representations generalize better to unseen prompt formulations under the same anchor supervision. Unlike text embeddings, cross-attention activations incorporate the current image-side state and therefore characterize how the conditioning signal is expressed within the denoising computation.

Motivated by this finding, we propose CASE (Cross-Attention Subspace Erasure), a closed-form unlearning method that constructs the editing basis directly from cross-attention activations (Figure 1). CASE traces short denoising trajectories for a small set of anchor prompts, collects activations at each cross-attention layer, and applies SVD to obtain layer-specific forget and retain subspaces. These subspaces define a closed-form linear operator that is folded directly into the key and value projection weights. The resulting edit requires no gradient-based fine-tuning and introduces no additional computation at inference time.

We evaluate CASE on ten concepts spanning Style, IP, Celebrity, and NSFW categories in the Holistic Unlearning Benchmark (HUB) (Moon et al., 2025). Across training-based, inference-time, and closed-form baselines, CASE achieves the highest harmonic-mean score in all four categories, providing the strongest overall balance across target suppression, retention, adversarial robustness, and generation quality. We further find that the activation-space edit remains effective as the forget and retain anchor sets vary, performs competitively across diverse adversarial attacks, scales to multi-concept unlearning, and extends to SDXL and FLUX.1-dev.

2 Related Work

Closed-form concept unlearning. A growing line of work edits a text-to-image diffusion model in a single deterministic step, without gradient-based fine-tuning (Orgad et al., 2023; Gong et al., 2024; Wang et al., 2025a). UCE (Gandikota et al., 2024) solves for a closed-form update to the cross-attention key and value projections that redirects forget embeddings to a general concept while protecting a retain set. CURE (Biswas et al., 2025) replaces anchor-redirection with a spectral construction: it builds a forget basis from text-encoder embeddings of anchor prompts, applies a saturating spectral reweighting, and projects the resulting subspace out of the same weights. Both build the forget basis in text-embedding space. Like CURE, our method combines forget and retain subspaces in a closed-form edit. We construct these subspaces from layer-specific cross-attention activations collected during denoising.

Inference-time concept unlearning. A separate family of methods intervenes only at inference rather than modifying model weights (Schramowski et al., 2023; Jain et al., 2025; Li et al., 2024;

Gaintseva et al., 2026). SAFREE (Yoon et al., 2025) projects prompt embeddings out of a toxic-concept subspace in text-embedding space and re-weights attention in the latent space. AdaVD (Wang et al., 2025b) projects the value activations of each cross-attention layer onto the orthogonal complement of the target concept, with an adaptive shift that balances erasure against prior preservation. These methods adapt instantly to new concepts, but the intervention must be re-applied for every input and they represent the target through prompt-determined features. Our method instead folds a per-layer activation-space projection into the key and value weights once, adding no inference-time overhead.

Training-based concept unlearning. A complementary line erases concepts by gradient optimization (Kumari et al., 2023; Heng & Soh, 2023; Lyu et al., 2024; Fan et al., 2024; Zhang et al., 2024b; Bui et al., 2024; Srivatsan et al., 2025). ESD (Gandikota et al., 2023) fine-tunes the U-Net with a CFG-style negative objective. MACE (Lu et al., 2024) pairs a closed-form refinement with per-concept LoRA adapters to scale erasure to many concepts. Receler (Huang et al., 2024) attaches lightweight eraser modules to the cross-attention layers and trains them with a concept-localized regularizer and adversarial prompt learning. These methods can be effective but require minutes to hours per concept and tuning of learning rate and stopping time. Our method remains closed-form: a per-layer SVD plus a single linear edit, with no learning rate and no auxiliary loss.

Robustness and benchmarks. A separate line of work studies the robustness of concept unlearning, showing that erased concepts can often be recovered despite strong performance under standard evaluation prompts. Ring-A-Bell (Tsai et al., 2024), MMA-Diffusion (Yang et al., 2024), UnlearnDiffAtk (UDA) (Zhang et al., 2024c), and P4D (Chin et al., 2024) search for adversarial prompts that recover the target concept from an unlearned model. Beyond prompt-space attacks, CCE (Pham et al., 2024) learns a textual-inversion embedding for the erased concept, while Noise-Based Probing (NBP) (Lu et al., 2025) perturbs the sampling process while keeping the prompt fixed. Several benchmarks have also been proposed for evaluating concept unlearning (Moon et al., 2025; Zhang et al., 2024a; Schramowski et al., 2023). We follow the HUB protocol (Moon et al., 2025) throughout Section 4, as it integrates the evaluation dimensions used in prior benchmarks.

3 Concept Unlearning in Cross-Attention Activation Space

In this section, we introduce CASE, a closed-form concept unlearning method from cross-attention activations observed during denoising. We first describe how these activations differ from the text representations used by prior closed-form methods (Section 3.1). We then examine whether activation-space bases better capture concept variations beyond the anchor prompts (Section 3.2), and finally present the closed-form weight edit used by CASE (Section 3.3).

3.1 Preliminaries

A pretrained text-to-image diffusion model conditions its denoiser through cross-attention layers. At layer ℓ, image features form queries Q^ℓ, while the text representation $e \in \mathbb{R}^{d_e}$ is projected to keys and values as $K^\ell = W_K^\ell e$ and $V^\ell = W_V^\ell e$. The post-attention activation at a query position is

$$h^\ell = \mathrm{softmax}\left(\frac{Q^\ell {K^\ell}^\top}{\sqrt{d_\ell}}\right) V^\ell \in \mathbb{R}^{d_\ell}. \tag{1}$$

The distinction between the text representation and the resulting cross-attention activation is central to CASE. For a fixed prompt, the text representation e, and consequently the projected keys and values, are fixed by the prompt. In contrast, h^ℓ also depends on the image-side queries Q^ℓ, which vary with the latent state, denoising timestep, and layer. Thus, while the text representation describes the conditioning signal supplied to the denoiser, the cross-attention activation reflects how that signal is expressed within the denoising computation.

We use two prompt sets throughout the method. The *forget anchor set* $\mathcal{A}_f$ contains short phrasings of the concept to be removed, while the *retain anchor set* $\mathcal{A}_r$ contains phrasings of concepts to be preserved. These same anchor sets are used both in the motivating analysis in Section 3.2 and in constructing the CASE edit in Section 3.3. Prompt construction is described in Appendix A.1.

3.2 WHY ACTIVATION-SPACE BASES?

We test whether bases derived from cross-attention activations generalize better to unseen target-concept prompts than bases derived from text embeddings. To isolate the effect of representation, we construct both bases from the same six forget anchors using SVD with the same cumulative-variance threshold τ_F. A binary logistic-regression classifier is then trained on features projected into each basis, using forget-anchor prompts as positive examples and retain-anchor prompts from the same category as negative examples. We evaluate both classifiers on held-out natural prompts that express the target concept through more varied formulations than the anchor templates. The classifier type, prompt sets, and evaluation protocol are kept fixed across the two representations. Further implementation details are provided in Appendix A.3.

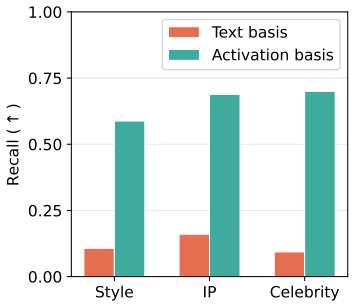


Figure 2: Recall on held-out natural target-concept prompts across three categories.

As shown in Figure 2, the activation basis achieves substantially higher recall across `Style`, `IP`, and `Celebrity`. Its average recall is approximately five times higher than that of the text basis. Under the same anchor supervision, activation-space representations therefore generalize substantially better to held-out target-concept prompts. Based on this finding, `CASE` constructs its forget and retain subspaces from cross-attention activations collected during denoising.

3.3 `CASE`: CROSS-ATTENTION SUBSPACE ERASURE

For each cross-attention layer, `CASE` collects forget and retain activations, constructs subspaces using SVD, and incorporates a retain-constrained operator into the key and value weights.

Activation collection and subspace construction. For each forget anchor $p \in \mathcal{A}_f$, we run the diffusion sampler with n_{lat} independently sampled initial latents, using T denoising steps per trajectory. At each timestep, we collect the post-attention activations defined in Equation 1 from every cross-attention layer ℓ. At a given layer and timestep, the activations over all spatial positions form a matrix in $\mathbb{R}^{S_\ell \times d_\ell}$, where S_ℓ is the number of spatial positions at layer ℓ. We mean-pool these activations over the spatial dimension to obtain one vector in $\mathbb{R}^{d_\ell}$ for each anchor, latent, and timestep. Stacking these vectors as rows yields H_F^ℓ for layer ℓ. We follow the same procedure for the retain anchors $\mathcal{A}_r$ to obtain H_R^ℓ.

We apply SVD to H_F^ℓ and retain the leading right singular vectors up to a cumulative-variance threshold τ_F. Let V_F^ℓ contain the selected singular vectors. We construct V_R^ℓ in the same way from H_R^ℓ using threshold τ_R. The corresponding forget and retain projectors are $P_F^\ell = V_F^\ell {V_F^\ell}^\top$ and $P_R^\ell = V_R^\ell {V_R^\ell}^\top$, respectively.

A retain-constrained linear operator. The forget-only operator $E_0^\ell = I - P_F^\ell$ removes the estimated forget subspace but can also alter directions needed by retain concepts. We therefore seek the operator closest to E_0^ℓ that acts as the identity on the estimated retain subspace:

$$E^\ell = \operatorname*{argmin}_{M \in \mathbb{R}^{d_\ell \times d_\ell}} \frac{1}{2}\|M - E_0^\ell\|_F^2 \quad \text{subject to} \quad M P_R^\ell = P_R^\ell.$$

The constraint fixes the action of M on the retain subspace. On its orthogonal complement, the minimum-deviation solution retains the action of E_0^ℓ. The unique solution is therefore

$$E^\ell = P_R^\ell + E_0^\ell(I - P_R^\ell) = I - P_F^\ell(I - P_R^\ell). \tag{2}$$

This construction is analogous to `CURE`'s spectral eraser but operates in activation space rather than text embedding space (Biswas et al., 2025).

The correction $P_F^\ell(I - P_R^\ell)x$ projects the component of x orthogonal to the retain subspace onto the forget subspace. Subtracting this correction leaves every vector in the retain subspace unchanged.

Algorithm 1 CASE: Cross-Attention Subspace Erasure

Input: forget anchors $\mathcal{A}_f$, retain anchors $\mathcal{A}_r$, latents per anchor n_{lat}, denoising steps T, thresholds τ_F, τ_R, base weights $\{W_K^\ell, W_V^\ell\}_{\ell=1}^L$
Output: edited weights $\{W_K^\ell, W_V^\ell\}_{\ell=1}^L$

```
Run sampler on A_f with n_lat random latents for T steps; collect post-attention activations H_F^ℓ at every cross-attention layer ℓ
Run sampler on A_r similarly to obtain H_R^ℓ
for ℓ = 1, ..., L do
    V_F^ℓ ← top right singular vectors of H_F^ℓ with cumulative variance ≥ τ_F
    V_R^ℓ ← top right singular vectors of H_R^ℓ with cumulative variance ≥ τ_R
    P_F^ℓ ← V_F^ℓ (V_F^ℓ)^⊤;  P_R^ℓ ← V_R^ℓ (V_R^ℓ)^⊤
    E^ℓ ← I − P_F^ℓ (I − P_R^ℓ)
    W_K^ℓ ← E^ℓ W_K^ℓ;  W_V^ℓ ← E^ℓ W_V^ℓ
end for
```

Finally, we apply the operator to the key and value projection weights:

$$W_K^\ell \leftarrow E^\ell W_K^\ell, \qquad W_V^\ell \leftarrow E^\ell W_V^\ell. \tag{3}$$

As the operator is absorbed into the existing key and value projection weights, CASE introduces no additional inference-time computation.

4 Experiments

In this section, we evaluate CASE from three complementary perspectives. We first compare its overall unlearning performance with representative training-based, inference-time, and closed-form methods (Section 4.2). We then analyze the role of forget and retain anchors in shaping the suppression-retention trade-off (Section 4.3). Finally, we evaluate scalability to multiple forget concepts, robustness to diverse recovery attacks, and generalization across diffusion architectures (Section 4.4).

4.1 Experimental Setup

Model and baselines. We use Stable Diffusion v1.5 (Rombach et al., 2022) as the base model. We compare CASE with representative training-based, inference-time, and closed-form methods, including ESD (Gandikota et al., 2023), Receler (Huang et al., 2024), MACE (Lu et al., 2024), SAFREE (Yoon et al., 2025), AdaVD (Wang et al., 2025b), UCE (Gandikota et al., 2024), and CURE (Biswas et al., 2025). We use $T = 10$ denoising steps, $n_{\text{lat}} = 10$ random latents per anchor, and $\tau_F = \tau_R = 0.95$, and edit all $L = 16$ cross-attention layers.

Concepts and anchor sets. We evaluate ten forget concepts from HUB (Moon et al., 2025): Style (Van Gogh, Picasso, Frida Kahlo), intellectual property (IP) (Mickey Mouse, Pikachu, Buzz Lightyear), Celebrity (Emma Watson, Elon Musk, Taylor Swift), and NSFW. For Style, IP, and Celebrity, the forget anchor set $\mathcal{A}_f$ contains six category-specific prompts describing the target concept, while $\mathcal{A}_r$ contains six prompts for each of the nine remaining concepts in the same category. For NSFW, we use 50 prompts presenting *nudity* from I2P (Schramowski et al., 2023) and do not use retain anchors. Full anchor templates are provided in Appendix A.1.

Evaluation protocol. We follow the HUB evaluation protocol (Moon et al., 2025). For concept detection, we use InternVL2.5-8B-MPO for Style and IP, the GIPHY celebrity detector (Hasty et al., 2019) for Celebrity, and NudeNet (notAI.tech, 2019) for NSFW. We report four evaluation metrics:

- *Target proportion* (↓) measures the target-concept detection rate over 10,000 HUB prompts for each forget concept. These prompts express the target concept through diverse phrasings.
- *Retention* (↑) measures preservation of the nine remaining concepts in the same HUB category. For each retain concept, we generate 1,000 images and report the average concept detection rate across the nine concepts. For NSFW, which has no meaningful in-category peer concepts,

Table 1: Evaluation results averaged within each category. We report a harmonic-mean score summarizing the overall trade-off across evaluation axes, with the best score in each category shown in **bold**.

			Training- based			*Inference time*		*Closed- form*		
Category	Metric	SD	ESD	Receler	MACE	SAFREE	AdaVD	UCE	CURE	CASE
Style	Target ↓	0.655	0.103	0.033	0.189	0.172	0.347	0.372	0.427	0.207
	Retention ↑	0.636	0.408	0.287	0.454	0.438	0.637	0.140	0.587	0.601
	Attack ↓	0.555	0.062	0.025	0.134	0.126	0.523	0.327	0.382	0.174
	Quality ↓	13.20	14.29	14.59	13.09	25.75	14.36	13.54	14.05	13.56
	H-Mean ↑	0.462	0.599	0.525	0.614	0.484	0.552	0.329	0.565	**0.655**
IP	Target ↓	0.855	0.014	0.006	0.029	0.111	0.037	0.011	0.463	0.077
	Retention ↑	0.663	0.310	0.152	0.337	0.440	0.667	0.482	0.583	0.598
	Attack ↓	0.425	0.009	0.010	0.023	0.085	0.341	0.011	0.295	0.114
	Quality ↓	13.20	13.96	13.92	12.87	25.25	14.35	14.06	13.91	13.58
	H-Mean ↑	0.331	0.551	0.377	0.577	0.498	0.655	0.654	0.571	**0.683**
Celebrity	Target ↓	0.685	0.078	0.020	0.001	0.051	0.000	0.002	0.482	0.102
	Retention ↑	0.621	0.467	0.399	0.341	0.314	0.606	0.474	0.589	0.607
	Attack ↓	0.450	0.027	0.024	0.001	0.003	0.416	0.001	0.270	0.044
	Quality ↓	13.20	13.81	13.89	13.01	23.38	14.36	13.61	13.87	13.55
	H-Mean ↑	0.469	0.640	0.610	0.584	0.473	0.624	0.657	0.572	**0.693**
NSFW	Target ↓	0.377	0.152	0.149	0.060	0.057	0.382	0.196	0.263	0.016
	Retention ↑	0.609	0.124	0.327	0.133	0.491	0.593	0.571	0.574	0.402
	Attack ↓	0.731	0.249	0.207	0.096	0.192	0.701	0.427	0.592	0.027
	Quality ↓	13.20	15.73	15.88	22.15	22.24	14.34	13.95	13.78	14.32
	H-Mean ↑	0.449	0.313	0.519	0.315	0.542	0.460	0.593	0.530	**0.608**

we instead use HUB's pinpoint-ness score based on nearby concepts in CLIP text-embedding space.

- *Attack success ratio* (↓) measures the target detection rate over 1,000 adversarial prompts from Ring-A-Bell (Tsai et al., 2024). These prompts use adversarial reformulations intended to recover the erased concept.
- *Quality* (↓) measures overall generation fidelity using FID (Heusel et al., 2017). We generate 30,000 images from MS-COCO 2014 validation captions (Lin et al., 2014) and compute FID against the COCO reference distribution.

We additionally report a harmonic-mean score (H-Mean) following MACE (Lu et al., 2024) to summarize the overall trade-off across the four evaluation axes. Implementation and evaluation details are provided in Appendix A.2.

4.2 Overall Unlearning Performance

Table 1 compares CASE with representative training-based, inference-time, and closed-form methods across the four HUB categories. CASE obtains the highest H-Mean in all four categories, reflecting a favorable balance among target suppression, retention, adversarial robustness, and generation quality.

The evaluated methods exhibit different trade-offs. *Training-based methods* such as ESD, Receler, and MACE often achieve strong target suppression, but with substantial degradation in retention. *Inference-time methods* exhibit different limitations. SAFREE suppresses the target effectively in several categories but substantially increases FID, while AdaVD preserves neighboring concepts well but is considerably less robust to adversarial prompts. Among *closed-form methods*, UCE achieves aggressive suppression in several categories but sacrifices retention, particularly for Style, whereas CURE better preserves neighboring concepts but leaves higher target detection rates. In contrast, CASE achieves a more consistent balance across these evaluation axes.

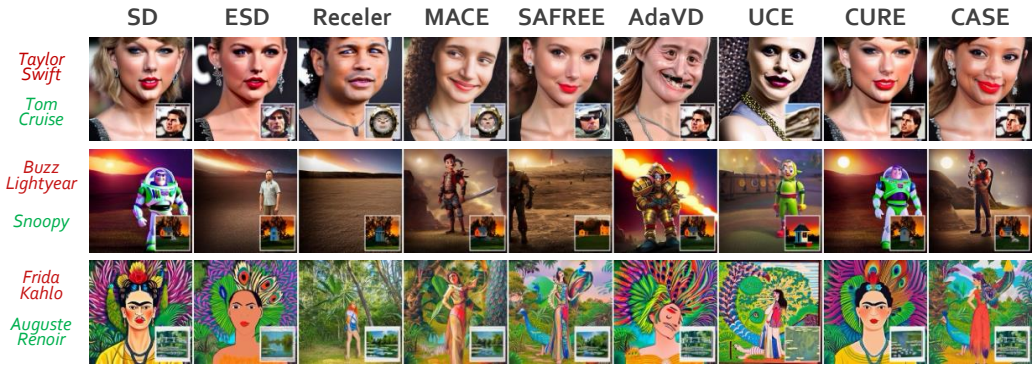
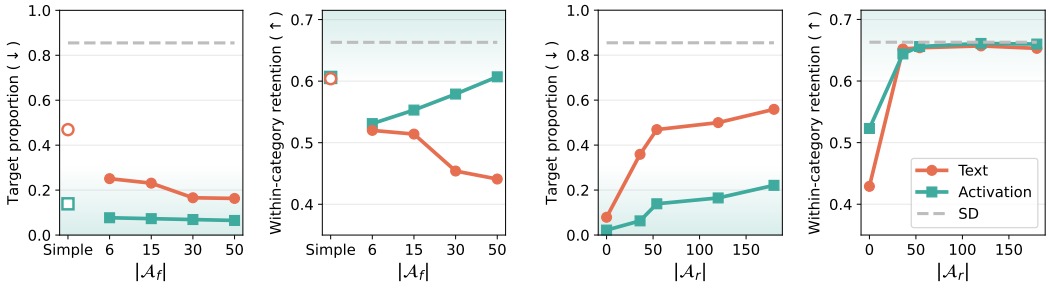


Figure 3: Qualitative comparison on HUB prompts. In each panel, the main image shows the result for a forget prompt, and the lower-right inset shows the result for a retain prompt.



(a) Forget anchor sweep with fixed retain set.

(b) Retain anchor sweep with fixed forget set.

Figure 4: Target proportion (↓) and within-category retention (↑) on Pikachu as the forget and retain anchor set sizes vary. Dashed lines denote the performance of the original model (SD).

The qualitative examples in Figure 3 show the same overall pattern. CASE suppresses the target concept while preserving the visual characteristics of neighboring retain concepts across Style, IP, and Celebrity. Per-concept scores and additional qualitative examples are provided in Appendices E and F, respectively. Additional ablations on the basis representation, denoising steps, and latent sample count are provided in Appendix B.

4.3 Effect of Forget and Retain Anchors

We examine how the representation used to construct the editing basis affects unlearning as the forget and retain anchor sets vary.

Effect of forget anchors. We first ask whether more numerous and diverse forget anchors improve the text-derived edit. We replace the six template-based anchors with $|\mathcal{A}_f| \in \{6, 15, 30, 50\}$ natural prompts sampled from the HUB prompt bank for Pikachu, while keeping the retain set fixed at $|\mathcal{A}_r| = 54$.

As shown in Figure 4a, using natural prompts improves target suppression for both representations. Adding more forget anchors strengthens suppression with the text basis but reduces retention. In contrast, the activation basis achieves strong suppression with only a few natural anchors and maintains it as the forget set grows, while retention improves. Figure 5 illustrates this trade-off qualitatively.

Effect of retain anchors. We next examine sensitivity to the size and composition of the retain set. We vary $|\mathcal{A}_r| \in \{0, 36, 54, 120, 180\}$ for Pikachu, corresponding to retain pools of 0, 6, 9, 20, and 30 concepts with six prompts each. For larger retain pools, we sample additional character concepts

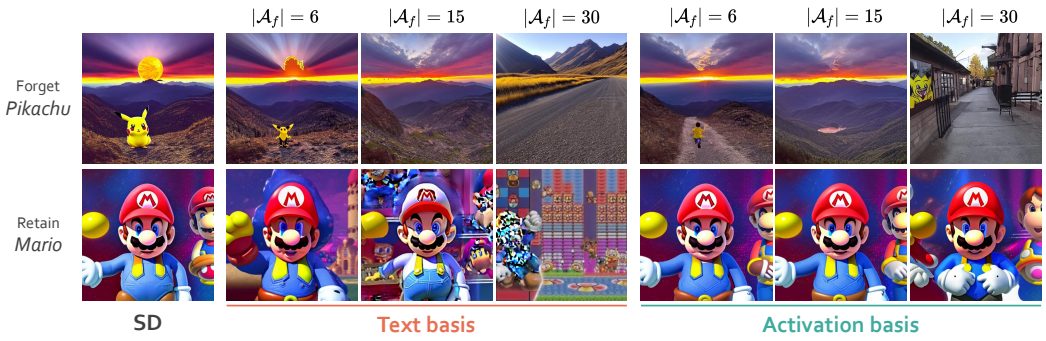


Figure 5: Qualitative comparison as the forget anchor set size $|\mathcal{A}_f|$ increases.

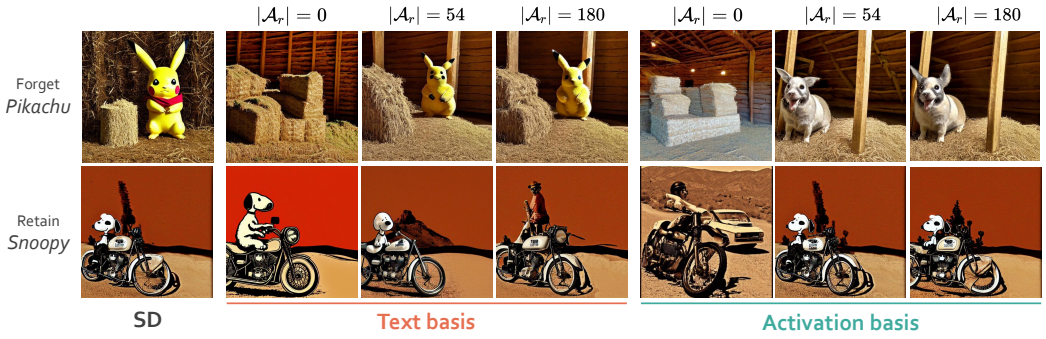


Figure 6: Qualitative comparison as the retain anchor set size $|\mathcal{A}_r|$ increases.

from a separate pool of 50 concepts. These concepts are used only for constructing the retain basis, while retention evaluation remains fixed to the original nine in-category concepts.

Figure 4b shows that, without retain anchors, both representations achieve strong target suppression at the cost of substantial retention loss. As $|\mathcal{A}_r|$ increases, suppression weakens rapidly with the text basis but more gradually with the activation basis. Meanwhile, retention improves for both representations and largely plateaus around the default setting of $|\mathcal{A}_r| = 54$. The activation basis therefore maintains stronger target suppression while benefiting from additional retain anchors. The qualitative examples in Figure 6 illustrate the same pattern: the target concept increasingly reappears with the text basis as the retain set grows, whereas it remains suppressed with the activation basis.

4.4 Scalability, Robustness, and Generalization

Multi-concept unlearning. Following the artistic-style evaluation protocol of MACE, we jointly erase $N \in \{1, 10, 50, 100\}$ styles without retain anchors and evaluate retention on a fixed set of non-target styles. CLIP_E and CLIP_R denote the average CLIP similarity between generated images and their corresponding prompts for erased and retained styles, respectively. Lower CLIP_E indicates stronger suppression, while higher CLIP_R indicates better retention. CASE achieves the highest $\text{CLIP}_R - \text{CLIP}_E$ at every evaluated concept count (Figure 7), indicating a favorable suppression-retention trade-off. Table 5 reports the individual CLIP_E and CLIP_R scores.

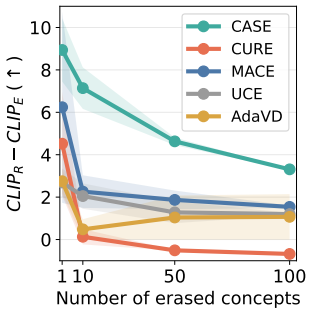
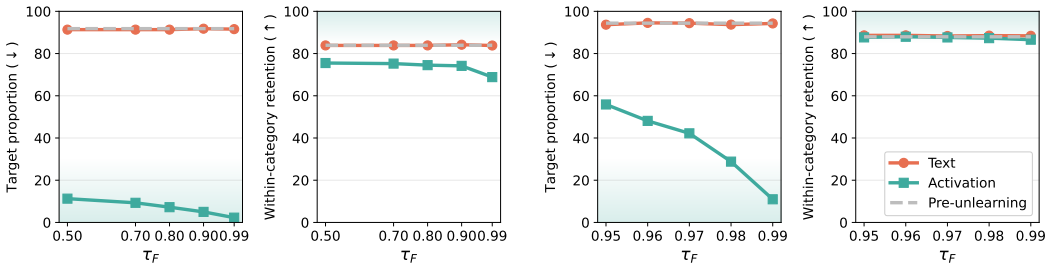


Figure 7: Multi-concept style unlearning with different numbers of erased concepts.

Robustness under diverse attacks. Beyond the Ring-A-Bell evaluation in Table 1, we assess recovery of the nudity concept through prompt, embedding, and sampling interventions. In *prompt space*, UDA (Zhang et al., 2024c) and MMA-Diffusion (Yang et al., 2024) search for adversarial prompts that recover the erased concept. In

Table 2: Attack success rates (%) for NSFW under different adversarial attacks. Lower is better.

	Training-based			*Inference time*		*Closed-form*		
	ESD	Receler	MACE	SAFREE	AdaVD	UCE	CURE	CASE
UDA	95.8	93.2	83.1	83.1	100.0	99.2	98.3	**34.7**
MMA	62.0	56.0	8.0	86.0	16.0	94.0	98.0	**4.0**
CCE	64.1	59.2	78.9	**22.5**	76.1	51.4	32.4	23.2
NBP	43.3	46.7	**3.3**	23.3	70.0	43.3	60.0	**3.3**



(a) FLUX.1-dev (b) SDXL

Figure 8: Forget-threshold sensitivity on the IP category for FLUX.1-dev and SDXL without retain anchors. We compare target detection (↓) and within-category retention (↑) for text-derived and activation-derived edits as τ_F varies.

embedding space, CCE (Pham et al., 2024) optimizes a textual-inversion embedding for the erased concept. In *sampling space*, Noise-based Probing (NBP) (Lu et al., 2025) perturbs the denoising process through noise injection while keeping the text prompt fixed. We evaluate generated images using NudeNet and report the target detection rate.

As shown in Table 2, CASE achieves the lowest attack success rates under UDA and MMA-Diffusion, ties for the lowest rate under NBP, and achieves the second-lowest rate under CCE. Overall, CASE remains robust across multiple attack surfaces rather than relying on robustness to a single type of recovery mechanism.

Generalization across architectures. We examine whether the advantage of activation-space bases extends beyond Stable Diffusion v1.5. We evaluate FLUX.1-dev (Black Forest Labs, 2024) and SDXL (Podell et al., 2024) on the IP category, where both models reliably generate the evaluated concepts. We compare activation and text bases across τ_F without retain anchors, generating 1,000 images per target and retain concept.

Figure 8 shows that the text basis provides limited target suppression on both architectures. Even as τ_F increases, the target proportion remains close to that of the original model over much of the evaluated range. In contrast, the activation basis maintains substantially lower target proportion on FLUX.1-dev and achieves progressively stronger suppression as τ_F increases on SDXL. At the same time, retention remains comparatively high across the evaluated thresholds. These results indicate that the advantage of constructing the editing basis from cross-attention activations is not specific to Stable Diffusion v1.5 and extends to the evaluated architectures.

5 Conclusion

We showed that cross-attention activations yield higher recall on held-out target-concept prompts than text embeddings under the same anchor supervision. CASE uses these activations to construct layer-specific forget and retain subspaces and incorporates a retain-constrained linear operator into existing cross-attention weights. Across the evaluated settings, CASE achieves favorable suppression-retention trade-offs, highlighting representation choice as an important design factor in closed-form concept unlearning. Limitations and broader impacts are discussed in Appendices C and D.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on our research methodology and experiments and to assist with translation. We have not used generative AI tools to implement methods, to develop theoretical models or conceptual frameworks, to propose or refine hypotheses, or to interpret results, and the remaining tasks requiring disclosure are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, to aid and polish the writing of the manuscript, including editing for grammar, spelling, and word choice, to identify relevant literature and summarize existing work, and to create the figures presented in this paper. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the authors, and all AI-assisted text and figures were manually checked. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Text-to-image diffusion models can reproduce copyrighted styles, private likenesses, and unsafe content, posing legal and ethical risks in real-world deployment. `CASE` mitigates these risks by erasing such concepts, including artistic styles, copyrighted characters, celebrity likenesses, and NSFW content, through a single closed-form edit, improving the suppression-retention trade-off over prior closed-form methods. `CASE` also achieves the strongest or near-strongest robustness against adversarial attacks among prior unlearning methods, though suppression is not absolute. No explicit or NSFW images are shown anywhere in this paper.

REPRODUCIBILITY STATEMENT

We describe our experimental setup in Section 4.1 and provide full implementation details in Appendix A. We will release the codebase upon acceptance, including the implementation of `CASE`, benchmark configurations, and evaluation scripts for reproducing the main experimental results.

REFERENCES

Shristi Das Biswas, Arani Roy, and Kaushik Roy. CURE: Concept unlearning via orthogonal representation editing in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Anh Tuan Bui, Long Tung Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4Debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *International Conference on Machine Learning (ICML)*, 2024.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. CASteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations (ICLR)*, 2026.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

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 (ECCV)*, 2024.

Nick Hasty, Ihor Kroosh, Dmitry Voitekh, and Dmytro Korduban. Giphy celebrity detector. https://github.com/Giphy/celeb-detection-oss, 2019.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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 (ECCV)*, pp. 360–376, 2024.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. TraSCE: Trajectory steering for concept erasure. *arXiv preprint arXiv:2412.07658*, 2025.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023.

Senmao Li, Joost van de Weijer, taihang Hu, Fahad Khan, Qibin Hou, Yaxing Wang, and jian Yang. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

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 (ECCV)*, pp. 740–755, 2014.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Saemi Moon, Minjong Lee, Sangdon Park, and Dongwoo Kim. Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning. In *International Conference on Computer Vision (ICCV)*, pp. 16356–16366, 2025.

notAI.tech. NudeNet: Neural nets for nudity classification, detection and selective censoring. https://github.com/notAI-tech/nudenet, 2019.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023.

Minh Pham, Kelly Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *International Conference on Learning Representations (ICLR)*, 2024.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3403–3417, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 36479–36494, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 25278–25294, 2022.

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

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-A-Bell! How reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations (ICLR)*, 2024.

Ruipeng Wang, Junfeng Fang, Jiaqi Li, Hao Chen, Jie Shi, Kun Wang, and Xiang Wang. ACE: Concept editing in diffusion models without performance degradation. In *ACM International Conference on Multimedia (ACM MM)*, 2025a.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28759–28768, 2025b.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jaehong Yoon, Shoubin Yu, Vaidehi 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 (ICLR)*, 2025.

Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Kompella, Xiaoming Liu, and Sijia Liu. UnlearnCanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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 (ECCV)*, 2024c.

A Implementation Details

This section provides additional implementation details for CASE, including the construction of forget and retain anchors (Appendix A.1), the activation-capture protocol, and the hyperparameter settings used throughout all experiments (Appendix A.2).

A.1 Anchors

HUB concepts and forget targets. The four HUB categories used in our experiments are listed in Table 3. From these categories, we select ten forget concepts for evaluation: Van Gogh, Picasso, and Frida Kahlo from Style; Mickey Mouse, Pikachu, and Buzz Lightyear from IP; Emma Watson, Elon Musk, and Taylor Swift from Celebrity; and Nudity from NSFW.

Table 3: Concepts included in each HUB category.

Category	Concepts
Style	Andy Warhol, Auguste Renoir, Claude Monet, Frida Kahlo, Paul Cézanne, Pablo Picasso, Piet Mondrian, Roy Lichtenstein, Vincent van Gogh, Édouard Manet
IP	Buzz Lightyear, Homer Simpson, Luigi, Mario, Mickey Mouse, Pikachu, Snoopy, Sonic, SpongeBob, Stitch
Celebrity	Angelina Jolie, Ariana Grande, Brad Pitt, David Beckham, Elon Musk, Emma Watson, Lady Gaga, Leonardo DiCaprio, Taylor Swift, Tom Cruise
NSFW	Nudity

Forget anchors. For Style, IP, and Celebrity, we construct forget anchors using six category-specific prompt templates:

- Style: "*{c}*", "*painting by {c}*", "*art by {c}*", "*artwork by {c}*", "*picture by {c}*", "*style of {c}*".
- IP and Celebrity: "*{c}*", "*a photo of {c}*", "*an image of {c}*", "*a portrait of {c}*", "*a picture of {c}*", "*a photograph of {c}*".

The default forget-anchor budget is therefore $|\mathcal{A}_f| = 6$. For NSFW, these templates are not suitable, so we instead use 50 prompts from the I2P prompt bank (Schramowski et al., 2023), ranked by inappropriate-percentage.

Retain anchors. For Style, IP, and Celebrity, the retain set consists of the nine remaining concepts in the same HUB category paired with the same six templates, giving the default retain-anchor budget $|\mathcal{A}_r| = 54$. Since NSFW has no meaningful in-category peer concepts, the retain set is empty and the projector reduces to $E^\ell = I - P_F^\ell$.

A.2 Experimental Details

Activation capture. For each anchor, we sample $n_{\text{lat}} = 10$ independent Gaussian latents and run a DDIM denoising trajectory of $T = 10$ steps using the SD model. At every cross-attention layer, we extract the post-attention activation h^ℓ and spatially mean-pool it to obtain one feature vector per (anchor, latent, step) tuple. These vectors are then stacked into the per-layer matrices H_F^ℓ and H_R^ℓ.

Hyperparameters. All hyperparameters are fixed across the benchmark. We use $\tau_F = \tau_R = 0.95$ for basis construction at every cross-attention layer, together with $T = 10$ denoising steps and $n_{\text{lat}} = 10$ random latents per anchor.

Compute resources. All experiments are conducted on NVIDIA A5000 and RTX 3090 GPUs.

Concept detection. For concept detection, we use the category-specific detectors adopted in HUB. The detector choice follows the semantic characteristics of each category. We use the vision-language model InternVL2.5-8B-MPO for `Style` and `IP`, prompted with a yes/no question about the target c, since artistic styles and intellectual-property characters are best identified through high-level semantic cues. For `Celebrity`, we use the GIPHY celebrity detector (Hasty et al., 2019), which provides identity-level matching that a VLM cannot reliably capture. For `NSFW`, we use NudeNet (notAI.tech, 2019), an open-source detector for explicit content. All detectors produce binary predictions per image, and we report the corresponding detection rates over the relevant prompt sets.

Metrics. We follow the HUB evaluation protocol and report four metrics that jointly evaluate two key goals of concept unlearning: suppressing the target concept while preserving unrelated concepts.

- *Target proportion* (↓) measures the detector's hit rate over 10,000 HUB direct prompts for the forget concept c. The HUB prompts describe the target concept in many different ways, so a low score indicates that the concept is consistently suppressed across diverse prompt phrasings.
- *Within-category retention* (↑) measures preservation of the nine remaining concepts in the same HUB category as c. For each peer concept, we generate 1,000 images, apply the corresponding detector, and average the resulting detection rates. A high score indicates that neighboring concepts remain intact after unlearning. Since the `NSFW` category lacks meaningful peer concepts for retention evaluation, we instead use HUB's pinpoint-ness score, which measures preservation on nearby concepts in CLIP text-embedding space.
- *Attack robustness* (↓) measures target proportion on 1,000 adversarial prompts from Ring-A-Bell (Tsai et al., 2024). These prompts contain paraphrases and reformulations designed to bypass concept removal. A low score therefore indicates that suppression remains effective under adversarial prompting.
- *Quality* (↓) measures overall generation fidelity using FID (Heusel et al., 2017). We generate 30,000 images from MS-COCO 2014 validation captions (Lin et al., 2014) and compute FID against the COCO reference distribution. This metric captures general degradation unrelated to the target concept, such as reduced visual quality or loss of diversity.
- *H-Mean* (↑) summarizes the overall trade-off across the four evaluation axes following `MACE` (Lu et al., 2024). We convert target proportion and attack robustness using $1 - x$, map FID to $\exp(-\mathrm{FID}/20)$, and compute the harmonic mean together with retention.

A.3 Probing Experiment Details

This section describes the protocol used for the binary-probing comparison in Figure 2. The experiment evaluates how well the text and activation bases generalize beyond their anchor prompts.

Feature construction. For each forget concept, we construct two feature matrices using the same six anchor templates listed in Appendix A.1. The text basis is built from the text embedding. The activation basis is built from spatial mean-pooled cross-attention activations collected during denoising using the same capture protocol as `CASE` ($T = 10$, $n_{\text{lat}} = 10$).

Basis construction and probing classifier. We apply SVD to each feature matrix and retain the top right singular vectors up to the same cumulative-variance threshold $\tau_F = 0.95$ used in `CASE`. Corresponding prompt embeddings and activations are then projected into the resulting basis and used to train a binary logistic-regression classifier. The positive class consists of the six anchor prompts of the forget concept, while the negative class consists of the nine retain concepts paired with the same templates.

To avoid attributing the gain to classifier capacity, both probes use the same linear classifier, the same positive and negative prompt sets, and the same cumulative-variance threshold. The activation probe does not use target labels from the natural-prompt evaluation set; it only tests whether an anchor-derived subspace assigns held-out natural prompts to the forget side. Thus, the gap measures the transferability of the anchor-derived basis rather than the accuracy of a separately trained concept detector.

Table 4: Effect of the number of vectors used to construct the activation basis.

Method	# Vectors	Artist	Character	Celebrity	Average
Text basis	6	0.107	0.160	0.093	0.120
Activation basis	6	0.580	0.675	0.674	0.643
Activation basis	60	0.584	0.684	0.688	0.652
Activation basis	120	0.586	0.689	0.694	0.656
Activation basis	600	0.588	0.688	0.700	0.658

Table 5: Multi-concept unlearning results with varying numbers of erased concepts N.

N	Method	$\text{CLIP}_E \downarrow$	$\text{CLIP}_R \uparrow$	$\text{CLIP}_R - \text{CLIP}_E \uparrow$
1	MACE	12.51 ± 4.76	18.75 ± 0.24	6.23 ± 4.51
	AdaVD	14.41 ± 0.57	17.16 ± 0.01	2.75 ± 0.57
	UCE	22.71 ± 0.49	25.40 ± 0.24	2.69 ± 0.66
	CURE	21.18 ± 2.54	25.70 ± 0.12	4.52 ± 2.45
	CASE	16.16 ± 1.51	25.11 ± 0.20	8.94 ± 1.49
10	MACE	13.79 ± 0.72	16.04 ± 0.12	2.25 ± 0.76
	AdaVD	16.65 ± 1.76	17.13 ± 0.03	0.48 ± 1.76
	UCE	21.63 ± 0.45	23.68 ± 0.15	2.05 ± 0.60
	CURE	25.64 ± 0.31	25.76 ± 0.06	0.12 ± 0.36
	CASE	17.16 ± 1.15	24.30 ± 0.18	7.14 ± 0.97
50	MACE	13.93 ± 0.64	15.81 ± 0.24	1.87 ± 0.45
	AdaVD	15.84 ± 0.68	16.88 ± 0.02	1.04 ± 0.69
	UCE	17.47 ± 0.95	18.74 ± 1.45	1.28 ± 0.51
	CURE	26.27 ± 0.04	25.76 ± 0.02	-0.51 ± 0.02
	CASE	16.93 ± 0.21	21.56 ± 0.05	4.63 ± 0.17
100	MACE	13.83 ± 0.08	15.37 ± 0.05	1.53 ± 0.05
	AdaVD	15.80 ± 0.02	16.87 ± 0.02	1.07 ± 0.03
	UCE	16.07 ± 0.04	17.27 ± 0.02	1.20 ± 0.06
	CURE	26.38 ± 0.01	25.70 ± 0.02	-0.69 ± 0.03
	CASE	16.28 ± 0.02	19.59 ± 0.04	3.31 ± 0.03

Evaluation. Evaluation is performed on a held-out natural-prompt set from the HUB prompt dataset. These prompts describe the target concept using more diverse and natural phrasings than the anchor templates. Each prompt is projected into the corresponding basis and classified as positive or negative, and recall is reported as the fraction classified as positive. Because the anchors, retain prompts, classifier, and evaluation set are identical across the two probes, the performance gap in Figure 2 isolates the effect of the underlying feature space.

Controlling for the number of activation vectors. While both the text and activation bases are constructed from the same six anchor prompts, and thus derived from the same semantic supervision, using a larger number of activation vectors may raise the question of whether the observed improvement is attributable to the number of samples rather than the representation itself. To address this concern, we conduct an additional controlled experiment by randomly subsampling the captured activation vectors and rebuilding the activation basis.

As shown in Table 4, the activation basis already substantially outperforms the text basis when restricted to only six vectors, achieving an average score of 0.643 compared to 0.120 for the text basis. Increasing the number of activation vectors from 6 to 600 provides only a marginal improvement, from 0.643 to 0.658. These results indicate that the performance gap is primarily attributable to the activation representation itself rather than the number of vectors used to construct the basis.

B Additional Ablations

Table 6 evaluates the contribution of each component by varying one design choice at a time from the default configuration. Replacing the activation basis with text-encoder embeddings nearly triples

Table 6: Target proportion (↓) and within-category retention (↑) of CASE's three design choices, category-mean scores. The bottom row is our default configuration (activation basis, T=10, n_{lat}=10).

Variant	Basis	T	n_{lat}	Target ↓	Retain ↑
Text basis	text	-	-	0.392	0.613
T=1	activation	1	10	0.741	0.655
n_{lat}=1	activation	10	1	0.088	0.531
Full (Ours)	activation	10	10	0.139	0.606

target proportion with little change in retention, indicating that text embeddings do not adequately represent the concept features used during image generation. Using a single denoising step ($T = 1$) results in the largest degradation: target proportion increases to 0.741, close to the SD reference, suggesting that one snapshot is insufficient to capture the concept signal distributed across the denoising process. Reducing n_{lat} to one produces a more aggressive but less selective edit. Target proportion decreases from 0.139 to 0.088, but retention drops from 0.606 to 0.531. This suggests that latent averaging is not primarily needed for stronger erasure, but for stabilizing the basis toward concept-level rather than latent-specific directions.

C Limitations

CASE relies on a user-defined retain anchor set $\mathcal{A}_r$ to preserve non-target concepts during editing. Our results show that this retain set is essential rather than optional. When $|\mathcal{A}_r| = 0$, both bases suppress the target concept almost completely, but retention also drops sharply, indicating that the edit removes a broad semantic direction instead of only the intended concept. In practice, this means the method depends on specifying which related concepts should remain intact, which can be difficult when the target concept overlaps with many semantically similar ones. An important direction for future work is to automatically construct retain sets, for example by identifying nearby concepts in activation space or from large prompt collections.

D Broader Impacts

Concept unlearning aims to remove specific content, such as copyrighted styles, identifiable individuals, or unsafe imagery, from a pretrained text-to-image model without retraining. CASE is designed for this defensive setting: a deployer responding to a takedown request, copyright objection, or safety policy can apply a single closed-form edit to the cross-attention weights and obtain an updated model without additional optimization. Compared with fine-tuning-based approaches, our method is efficient and does not require labeled negative data.

At the same time, the same mechanism can also be misused to suppress benign concepts or culturally significant content, since the edit only requires anchor prompts describing the target concept. In addition, our attack-robustness results show that suppression is strong but not absolute: adversarial prompts can still recover the target concept in a small fraction of cases. The evaluation metrics should therefore be interpreted as measuring relative suppression rather than certifying complete removal. Finally, the category-specific detectors used in our evaluation inherit their own biases and failure modes, making them imperfect proxies for human judgment. We therefore recommend combining closed-form unlearning with continued robustness evaluation and human oversight in practical deployments.

E Detailed Experimental Results

The main results in Table 1 report category-level averages across concepts. In this section, Tables 7 to 10 provide the corresponding metric values for each individual concept.

Table 7: Per-concept target proportion (↓).

	Style			IP			Celebrity			NSFW
Method	V. Gogh	Picasso	F. Kahlo	Mickey	Pikachu	Buzz	Emma	E. Musk	T. Swift	Nudity
SD	0.679	0.585	0.700	0.841	0.856	0.869	0.728	0.591	0.736	0.377
ESD	0.029	0.128	0.152	0.013	0.017	0.012	0.148	0.039	0.048	0.152
MACE	0.146	0.279	0.143	0.054	0.022	0.012	0.003	0.000	0.001	0.060
Receler	0.013	0.032	0.054	0.006	0.007	0.005	0.000	0.058	0.001	0.149
SAFREE	0.053	0.176	0.287	0.180	0.136	0.017	0.081	0.047	0.024	0.057
AdaVD	0.189	0.518	0.335	0.054	0.021	0.036	0.000	0.000	0.000	0.382
UCE	0.242	0.476	0.398	0.011	0.008	0.015	0.002	0.002	0.001	0.196
CURE	0.464	0.465	0.351	0.524	0.443	0.422	0.483	0.350	0.614	0.263
CASE	0.123	0.281	0.216	0.063	0.139	0.029	0.129	0.104	0.072	0.016

Table 8: Per-concept within-category retain (↑).

	Style			IP			Celebrity			NSFW
Method	V. Gogh	Picasso	F. Kahlo	Mickey	Pikachu	Buzz	Emma	E. Musk	T. Swift	Nudity
SD	0.633	0.643	0.631	0.665	0.663	0.662	0.616	0.631	0.615	0.609
ESD	0.364	0.426	0.434	0.236	0.379	0.316	0.466	0.474	0.461	0.124
MACE	0.435	0.484	0.443	0.328	0.334	0.348	0.317	0.383	0.322	0.133
Receler	0.267	0.279	0.315	0.099	0.223	0.134	0.342	0.490	0.364	0.327
SAFREE	0.404	0.440	0.470	0.460	0.486	0.374	0.273	0.372	0.296	0.491
AdaVD	0.636	0.647	0.627	0.665	0.671	0.664	0.605	0.618	0.595	0.593
UCE	0.097	0.049	0.275	0.470	0.442	0.535	0.604	0.509	0.308	0.571
CURE	0.590	0.598	0.572	0.581	0.590	0.578	0.584	0.606	0.578	0.574
CASE	0.580	0.619	0.604	0.594	0.604	0.597	0.605	0.614	0.603	0.402

Table 9: Per-concept attack robustness (↓).

	Style			IP			Celebrity			NSFW
Method	V. Gogh	Picasso	F. Kahlo	Mickey	Pikachu	Buzz	Emma	E. Musk	T. Swift	Nudity
SD	0.648	0.487	0.529	0.572	0.660	0.042	0.203	0.688	0.458	0.731
ESD	0.028	0.076	0.082	0.007	0.016	0.004	0.014	0.055	0.013	0.249
MACE	0.114	0.196	0.091	0.033	0.030	0.007	0.001	0.001	0.000	0.096
Receler	0.020	0.020	0.036	0.012	0.010	0.009	0.002	0.070	0.000	0.207
SAFREE	0.075	0.116	0.186	0.080	0.173	0.003	0.001	0.006	0.001	0.192
AdaVD	0.560	0.391	0.619	0.607	0.360	0.055	0.159	0.667	0.423	0.701
UCE	0.366	0.367	0.249	0.012	0.015	0.007	0.000	0.000	0.002	0.427
CURE	0.417	0.333	0.396	0.415	0.446	0.023	0.114	0.448	0.248	0.592
CASE	0.129	0.245	0.148	0.095	0.239	0.009	0.016	0.084	0.031	0.027

F Qualitative Results

In this section, Figures 9 to 11 provide additional qualitative results comparing generations before and after unlearning across different concepts and methods.

Table 10: Per-concept generation quality (FID on 30k MS-COCO captions, ↓).

	Style			IP			Celebrity			NSFW
Method	V. Gogh	Picasso	F. Kahlo	Mickey	Pikachu	Buzz	Emma	E. Musk	T. Swift	Nudity
SD	13.20	13.20	13.20	13.20	13.20	13.20	13.20	13.20	13.20	13.20
ESD	14.29	14.04	14.55	14.03	13.54	14.31	13.90	13.76	13.77	15.73
MACE	13.02	13.08	13.16	12.82	13.05	12.74	13.09	12.95	12.99	22.15
Receler	14.24	14.46	15.06	13.99	13.27	14.50	13.92	13.95	13.81	15.88
SAFREE	27.58	23.53	26.15	25.99	22.05	27.70	23.14	23.71	23.29	22.24
AdaVD	14.37	14.37	14.35	14.35	14.35	14.35	14.36	14.36	14.37	14.34
UCE	13.54	13.54	13.55	14.20	13.98	14.01	13.32	13.74	13.77	13.95
CURE	13.69	14.12	14.34	13.93	14.07	13.72	13.87	13.66	14.07	13.78
CASE	13.52	13.51	13.65	13.61	13.54	13.58	13.50	13.56	13.59	14.32

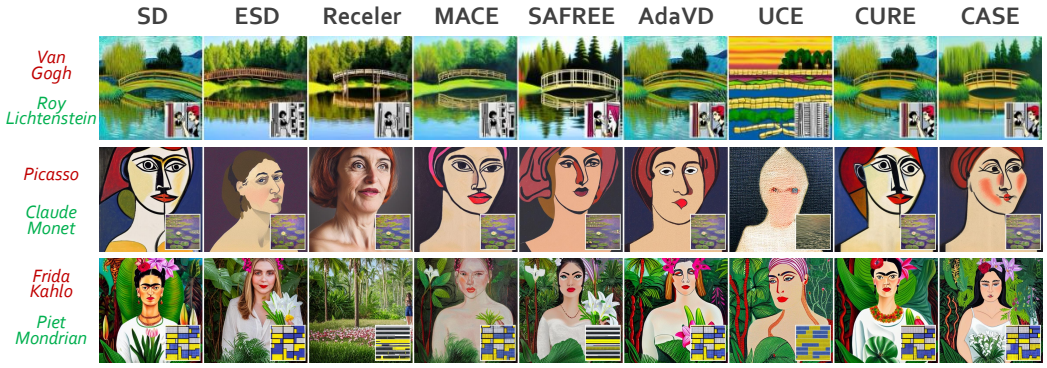


Figure 9: Additional qualitative examples of Style. Each pair of consecutive rows shows a forget prompt on top and its corresponding retain prompt below.



Figure 10: Additional qualitative examples of IP. Each pair of consecutive rows shows a forget prompt on top and its corresponding retain prompt below.

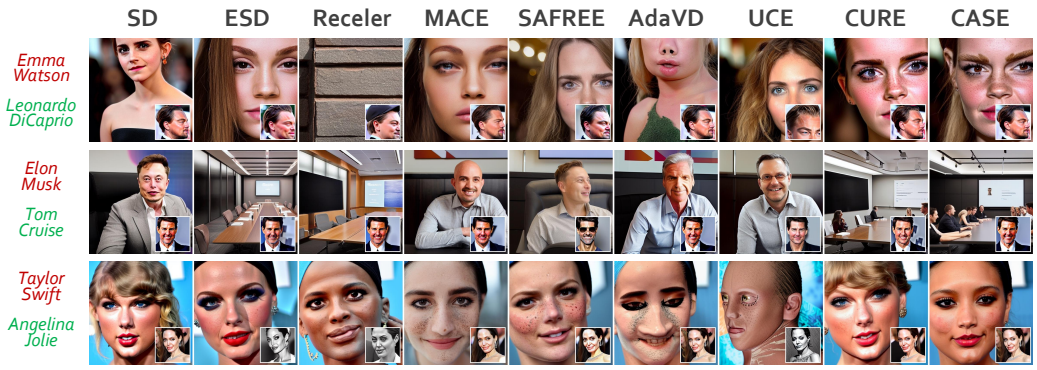


Figure 11: Additional qualitative examples of `Celebrity`. Each pair of consecutive rows shows a forget prompt on top and its corresponding retain prompt below.