Concept Unlearning via Cross-Attention Activation Projection for Diffusion Models

Anonymous Authors[1]

Abstract

Concept unlearning aims to erase a target concept from a pretrained text-to-image diffusion model without retraining. Closed-form methods are attractive in this setting because they apply a single deterministic edit to the cross-attention weights and add no inference-time cost. Existing closed-form methods, however, represent the target concept through the text encoder's response to a few short anchor prompts that name it, and paraphrased prompts that evoke the concept without naming it consistently bypass the edit. We argue that the target should instead be represented in the cross-attention activation space. Text embeddings describe the user's prompt, while cross-attention activations describe what the model is about to render, and the latter generalize to paraphrase the anchor templates do not cover. Building on this observation, we propose PURE (Projection in U-Net Rendering for Erasure), a closed-form method that builds the forget and retain bases from per-layer cross-attention activations captured along a short denoising trajectory and applies a single linear projector to the cross-attention key and value weights. On a recent holistic concept-unlearning benchmark covering ten concepts across artistic style, intellectual property, celebrity, and NSFW categories, PURE significantly reduces target leakage under paraphrased and adversarial prompts while preserving retain concepts close to the unedited model, yielding the best overall forget-retain trade-off among evaluated methods.

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 also reproduce content a deployed system should not regenerate, such as copyrighted artistic styles, the likeness of public figures, and unsafe imagery (Schuhmann et al., 2022; Qu et al., 2023; Kumari et al., 2023). Concept unlearning addresses this gap by editing a pretrained model to suppress target concepts without retraining from scratch (Gandikota et al., 2023; Lu et al., 2024; Biswas et al., 2025; Wang et al., 2025).

An effective unlearning method must satisfy two requirements at the same time. First, the target concept must no longer appear in generations, even under prompt variations, paraphrases, and adversarial inputs. Second, every other concept the model represents must remain intact, since the vast majority of prompts a deployed model receives are unrelated to the target concept. We refer to the two requirements as *forget* and *retain*, and a useful method is one that meets both without trading one against the other.

Recent closed-form methods (Gandikota et al., 2024; Biswas et al., 2025) address this trade-off in a single deterministic update without fine-tuning. UCE redirects the forget concept to a general concept while protecting a retain set. CURE replaces this redirection with a spectral construction over text-encoder embeddings of anchor prompts, then projects the resulting subspace out of the cross-attention key and value projections. Despite their differences, these methods share one structural choice: the forget basis is built from the text encoder's response to short anchor prompts such as "*Van Gogh*" or "*a painting by Van Gogh*". Closed-form erasure is only as robust as the subspace it removes. If a paraphrased prompt expresses the target concept outside the anchor-derived basis, the edit has no mechanism to detect or suppress it at inference time.

We hypothesize that constructing the basis from cross-attention activations alleviates this limitation. To compare the two feature spaces, we use the same anchor prompts to build two forget bases: one from text-encoder embeddings and the other from cross-attention activations collected during denoising. For each basis, we apply SVD to the anchor features and train a binary linear classifier using the forget concept as positives and nearby retain concepts as nega-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

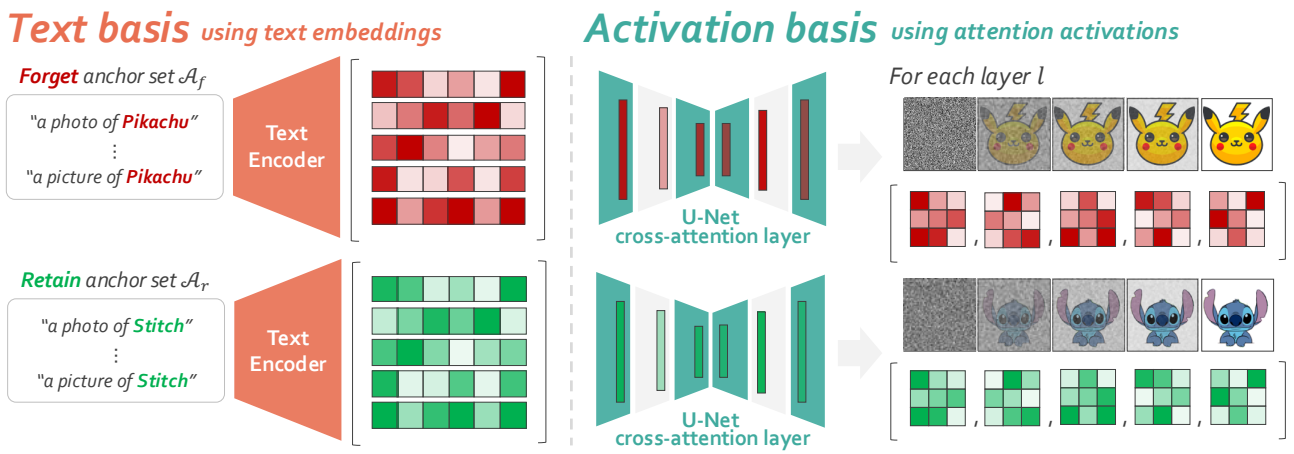


Figure 1. Comparison between text and activation bases. Prior closed-form methods build a shared text-space basis from anchor prompt embeddings. PURE instead captures layer-specific cross-attention activations during denoising and constructs an activation-space basis for each U-Net cross-attention layer.

tives. We then evaluate recall on natural prompts with more diverse phrasings. The text basis achieves substantially lower recall, whereas the activation basis improves recall by roughly fivefold (Figure 2). Text embeddings encode the prompt before denoising, while cross-attention activations reflect how the denoiser uses that prompt under the current latent state and timestep. We therefore expect activation features to better capture the concept evidence actually used during image generation.

This motivates *Projection in U-Net Rendering for Erasure* (PURE), a closed-form unlearning method that builds the forget and retain bases from cross-attention activations rather than text-encoder embeddings (Figure 1). We capture cross-attention activations as the U-Net traces a small number of anchor prompts through a short denoising trajectory, perform a per-layer SVD to obtain forget and retain subspaces, and apply a single linear edit to the cross-attention key and value projections. PURE inherits the projection-and-cancellation form of CURE and changes only the basis source and, as a forced consequence, the multiplication side.

Across ten concepts from four categories in the Holistic Unlearning Benchmark (HUB) (Moon et al., 2025), we compare PURE with five representative unlearning baselines. Among closed-form methods, PURE better balances forgetting and retention. It reduces the target leakage observed in CURE while avoiding the retention loss caused by more aggressive text-space edits such as UCE. Furthermore, PURE achieves the highest within-category retention in Style, IP, and Celebrity categories, and obtains the best harmonic-mean summary of target proportion, retention, attack robustness, and generation quality across all four evaluation categories (Table 1).

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 training (Orgad et al., 2023; Gong et al., 2024; Wang et al., 2025; Gaintseva et al., 2026). 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 re-weighting, and projects the resulting subspace out of the same weights. Both build the forget basis in text-embedding space. Our method inherits CURE's projection-and-cancellation structure but moves the SVD from text-embedding space to per-layer cross-attention activation space.

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., 2024a; 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 thread evalu-

ates how reliably an erased concept stays erased. Adversarial prompts can recover the target concept from models that pass standard erasure checks. Ring-A-Bell (Tsai et al., 2024) optimizes a prompt embedding to align with the target's CLIP representation, while UnlearnDiffAtk (Zhang et al., 2024c) and P4D (Chin et al., 2024) adapt internal red-teaming attacks to the same goal. Several benchmarks have been proposed for evaluating concept unlearning (Moon et al., 2025; Zhang et al., 2024b; Schramowski et al., 2023), and we follow the HUB protocol throughout Section 4, because it unifies the evaluation axes of prior benchmarks.

3. Closed-Form Edit in Activation Space

In this section, we introduce PURE, a closed-form concept unlearning method that builds the forget direction from per-layer cross-attention activations rather than text-encoder embeddings. We first introduce the cross-attention notation used throughout the method in Section 3.1, then present a probing experiment motivating the activation basis in Section 3.2, and finally describe the closed-form edit applied at every cross-attention layer in Section 3.3.

3.1. Setup

A pretrained text-to-image diffusion model passes the text condition through cross-attention layers in its U-Net denoiser. Concept unlearning edits these layers. At layer ℓ, image features form queries Q^ℓ, and a text embedding $e \in \mathbb{R}^{d_e}$ is projected to keys and values by $W_K^\ell, W_V^\ell \in \mathbb{R}^{d^\ell \times d_e}$. The post-attention activation at a single query position is

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

For a fixed text input, h^ℓ varies across the denoising trajectory because Q^ℓ depends on the noisy latent at each step. A text-encoder embedding and induced K and V, in contrast, is a deterministic function of the input prompt alone.

The probing experiment in Section 3.2 and the closed-form edit in Section 3.3 share the same two prompt sets. The *forget anchor set* $\mathcal{A}_f$ contains n_f short phrasings of concepts to be removed. The *retain anchor set* $\mathcal{A}_r$ contains n_r phrasings of concepts to be preserved. Prompt construction is described in Appendix A.1.

3.2. Probing the Forget Basis

Existing closed-form methods such as UCE and CURE derive the forget direction from text-encoder embeddings of $\mathcal{A}_f$ and reuse it at every cross-attention layer and denoising step. We ask whether building the same direction from cross-attention activations of the same anchors yields a basis that generalizes better to prompts the anchor templates do not literally cover. The question is operational: the closed-form edit can only suppress components represented in the constructed basis. A paraphrase that is weakly captured by the basis is therefore less likely to be affected by the edit.

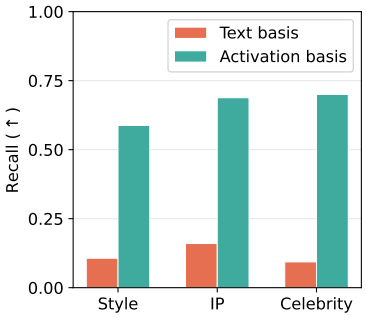


Figure 2. Binary-probing recall on natural prompts across categories (↑).

For each forget concept we construct two candidate bases that differ only in the features fed to SVD. The text basis uses the text-encoder embedding of each anchor in $\mathcal{A}_f$; the activation basis uses the spatial mean-pooled cross-attention activation. In both cases we apply SVD to the resulting feature matrix, keep the top right singular vectors up to threshold τ_F as the forget basis V_F, and train a binary logistic classifier in the V_F subspace with $\mathcal{A}_f$ as the positive class and the in-category retain pool as the negative class. The forget anchors, retain pool, cumulative-variance cutoff, and classifier are identical across the two probes, so the basis source is the only variable. We evaluate on a held-out set of natural prompts that describe the concept in longer, more varied form than the anchor templates, and report recall, the fraction classified as positive.

Figure 2 reports recall across the three category-level concept groups. The activation basis recalls natural prompts roughly five times more often than the text basis. Because the two probes share the same classifier, cutoff, and prompt sets, the gap reflects differences in the underlying representation rather than classifier capacity. Text embeddings encode the prompt before denoising, whereas cross-attention activations reflect how the denoiser uses the prompt under the current latent and timestep. The higher recall therefore suggests that activation-space bases capture concept evidence that is directly involved in generation. Full details are in Appendix A.3.

3.3. PURE: Projection in U-Net Rendering for Erasure

The PURE edit runs in three stages: capture activations from the model, build forget and retain subspaces by per-layer SVD, and apply one linear update to W_K^ℓ and W_V^ℓ. PURE preserves the practical advantages of closed-form editing: no gradient fine-tuning, no auxiliary loss, and no additional

Algorithm 1 Closed-form concept unlearning in cross-attention activation space

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
```

inference-time cost after the edit is applied.

For each anchor $p \in \mathcal{A}_f$, we run the diffusion sampler with n_{lat} random latents over T denoising steps. At every layer ℓ and every step, we read the post-attention activation from Equation (1). The per-step output stacks one h^ℓ vector per spatial position of the noisy latent into a tensor of shape $S^\ell \times d^\ell$, where S^ℓ is the spatial extent of the latent at layer ℓ. We mean-pool along the spatial axis, so each anchor, latent, and step contributes one row of length d^ℓ to a per-layer matrix H_F^ℓ. We repeat the protocol on $\mathcal{A}_r$ to obtain H_R^ℓ.

We compute the SVD of H_F^ℓ and collect its top right singular vectors up to a cumulative-variance threshold τ_F as the columns of V_F^ℓ. The forget projector is $P_F^\ell = V_F^\ell (V_F^\ell)^\top$. The retain projector $P_R^\ell = V_R^\ell (V_R^\ell)^\top$ is built identically from H_R^ℓ at threshold τ_R.

We left-multiply W_K^ℓ and W_V^ℓ by $E^\ell = \boldsymbol{I} - P_F^\ell (\boldsymbol{I} - P_R^\ell)$:

$$W_K^\ell \leftarrow E^\ell W_K^\ell, \qquad W_V^\ell \leftarrow E^\ell W_V^\ell. \tag{2}$$

Reading E^ℓ from right to left as it acts on an input x, the factor $\boldsymbol{I} - P_R^\ell$ removes any retain-aligned component of x, so P_F^ℓ acts only on the retain-orthogonal part. Subtracting from the identity then leaves any retain-aligned input unchanged: if $P_R^\ell x = x$ then $(\boldsymbol{I} - P_R^\ell)\, x = 0$ and $E^\ell x = x$, so retain inputs pass through the edit at the linear K, V level. `UCE` and `CURE` apply the same projection-and-cancellation template by right-multiplication in the text-encoder space; the activation-space projector has shape $d^\ell \times d^\ell$ and can multiply $W_K^\ell, W_V^\ell \in \mathbb{R}^{d^\ell \times d_e}$ only from the left. The shapes of W_K^ℓ and W_V^ℓ are unchanged, so the edit adds no runtime cost at inference.

`PURE` inherits two structural choices from `CURE`: the projection-and-cancellation form of the edit operator, and the cross-attention key and value projections as the target weights. `PURE` departs in two places. The forget basis is built from per-layer cross-attention activations rather than text-encoder embeddings, and the edit acts on the key and value weights by left-multiplication rather than right-multiplication. An activation-space basis lives in the post-projection space of dimension d^ℓ, which forces both the multiplication side and a per-layer edit. Algorithm 1 summarizes the full procedure.

4. Experiments

In this section, we study two questions. First, how effective is `PURE` with existing concept-unlearning methods? Second, how do anchor construction and activation-capture hyperparameters affect the forget-and-retain trade-off? Section 4.2 compares `PURE` against representative baselines, while Section 4.3 analyzes the forget anchor set, retain anchor set, and activation-capture hyperparameters.

4.1. Experimental Setting

We use Stable Diffusion v1.5 (Rombach et al., 2022) as the base model. We compare against representative training-based and closed-form unlearning methods: `ESD` (Gandikota et al., 2023), `MACE` (Lu et al., 2024), `RECELER` (Huang et al., 2024), `UCE` (Gandikota et al., 2024), and `CURE` (Biswas et al., 2025). For `PURE`, we use $T = 10$ capture steps, $n_{\text{lat}} = 10$ independent denoising runs per anchor, $\tau_F = \tau_R = 0.95$, and edit all $L = 16$ cross-attention layers. Additional details are provided in Appendix A.

Forget concepts. We choose target concepts from HUB (Moon et al., 2025), a benchmark designed for concept unlearning. HUB contains four categories: `Style`, `IP`, `Celebrity`, and `NSFW`. The `Style`, `IP`, and `Celebrity` categories each contain ten concepts, while `NSFW` contains three. From these categories, we select ten forget concepts for evaluation: `Style` (Van Gogh, Picasso, Frida Kahlo), `IP` (Mickey Mouse, Pikachu, Buzz Lightyear), `Celebrity` (Emma Watson, Elon Musk, Taylor Swift), and `NSFW` (Nudity).

Anchor set. For `Style`, `IP`, and `Celebrity`, we construct the forget anchor set $\mathcal{A}_f$ using six category-specific templates with the target concept name (e.g., "*a painting in the style of c*"). The retain anchor set $\mathcal{A}_r$ uses the same templates with the remaining nine concepts in the same category, resulting in $|\mathcal{A}_r| = 54$ prompts. For `NSFW`, these templates are not suitable because prompts such as "*a photo of Nudity*" do not reliably describe unsafe content. Instead, we select the top 50 prompts from the I2P dataset (Schramowski et al., 2023), ranked by inappropriate-percentage. Since `NSFW` has

Table 1. Evaluation results averaged within each category. We additionally 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			*Closed-form*		
Category	Metric	SD	ESD	MACE	RECELER	UCE	CURE	PURE
Style	Target ↓	0.655	0.103	0.189	0.033	0.372	0.427	0.207
	Retention ↑	0.636	0.408	0.454	0.287	0.140	0.587	0.601
	Attack ↓	0.555	0.062	0.134	0.025	0.327	0.382	0.174
	Quality ↓	13.20	14.29	13.09	14.59	13.54	14.05	13.56
	H-Mean ↑	0.462	0.599	0.614	0.525	0.328	0.565	**0.655**
IP	Target ↓	0.855	0.014	0.029	0.006	0.011	0.463	0.077
	Retention ↑	0.663	0.310	0.337	0.152	0.482	0.583	0.598
	Attack ↓	0.425	0.009	0.023	0.010	0.011	0.295	0.114
	Quality ↓	13.20	13.96	12.87	13.92	14.06	13.91	13.58
	H-Mean ↑	0.331	0.551	0.578	0.377	0.654	0.571	**0.683**
Celebrity	Target ↓	0.685	0.078	0.001	0.020	0.002	0.482	0.102
	Retention ↑	0.621	0.467	0.341	0.399	0.474	0.589	0.607
	Attack ↓	0.450	0.027	0.001	0.024	0.001	0.270	0.044
	Quality ↓	13.20	13.81	13.01	13.89	13.61	13.87	13.55
	H-Mean ↑	0.469	0.640	0.584	0.610	0.657	0.572	**0.693**
NSFW	Target ↓	0.515	0.222	0.399	0.163	0.514	0.512	0.352
	Retention ↑	0.609	0.124	0.133	0.327	0.571	0.574	0.402
	Attack ↓	0.623	0.217	0.221	0.206	0.629	0.646	0.335
	Quality ↓	13.20	15.73	22.15	15.88	13.95	13.78	14.32
	H-Mean ↑	0.482	0.312	0.296	0.518	0.470	0.465	**0.528**

no in-category peer concepts, the retain set is empty.

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 the CLIP-based Q16 classifier (Schramowski et al., 2022), trained to detect inappropriate 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.

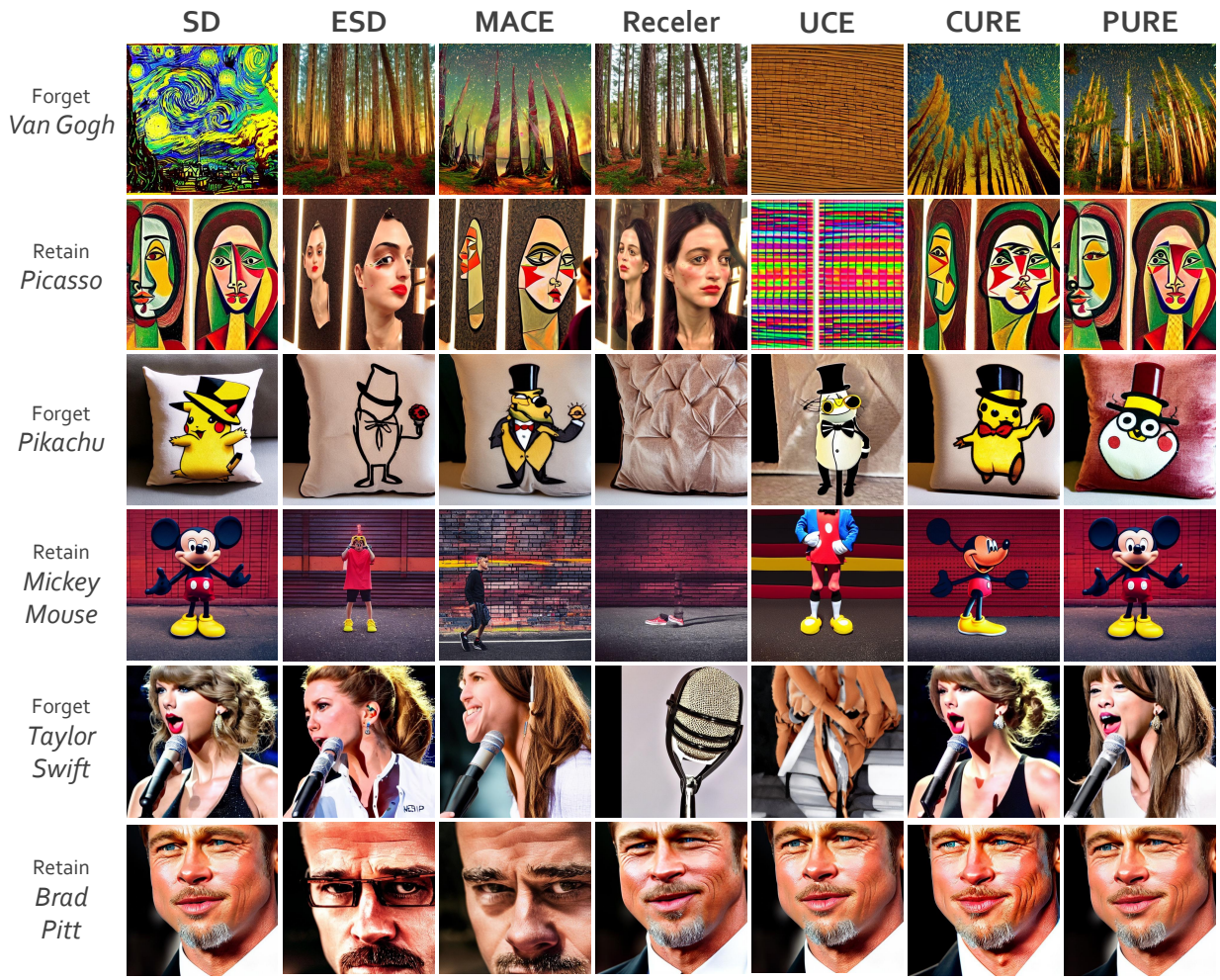


Figure 3. Qualitative comparison on forget and retain prompts from HUB. Training-based methods often damage neighboring concepts while suppressing the target concept, whereas prior closed-form methods preserve neighboring concepts but leave noticeable target leakage. PURE achieves stronger target suppression while preserving retain-image quality across categories.

4.2. Results

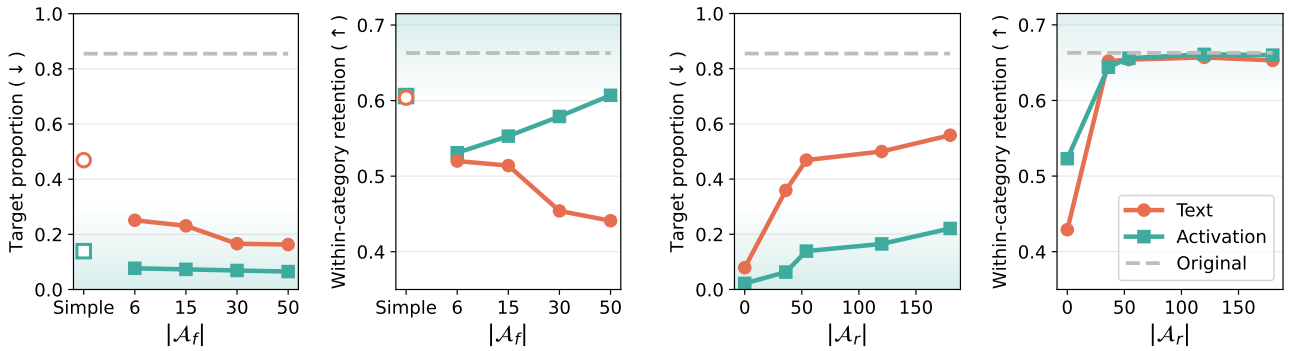
Table 1 shows per-category averages over the four evaluation metrics. Compared with CURE, PURE achieves lower target proportion across all categories. It also obtains the highest within-category retention in Style, IP, and Celebrity. Together, these results show that PURE improves the suppression-retention trade-off among closed-form methods. The qualitative examples in Figure 3 show the same trend, with reduced target generations and retain generations that remain close to the original model outputs.

Existing closed-form methods exhibit opposite failure modes. CURE preserves neighboring concepts but leaks target concepts, whereas UCE can erase aggressively in some categories but sacrifices retention, especially in Style. By contrast, training-based methods (ESD, MACE, and RECELER) achieve stronger target removal, but their fine-tuning process also damages nearby concepts, leading to substantially lower retention.

To summarize the overall trade-off across evaluation axes, we additionally report a harmonic-mean score (Lu et al., 2024) combining all four metrics. The target proportion and attack robustness metrics are converted using $1 - x$, while FID is mapped to $\exp(-\mathrm{FID}/20)$ to obtain a bounded quality score before computing the harmonic mean. Under this summary measure, our method achieves the strongest overall performance across all categories.

4.3. Ablations

Forget anchor set size. A natural question is whether the text basis improves when provided with a larger and more diverse set of forget anchors. To test this, 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 pool fixed at $|\mathcal{A}_r| = 54$.



(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 SD reference.

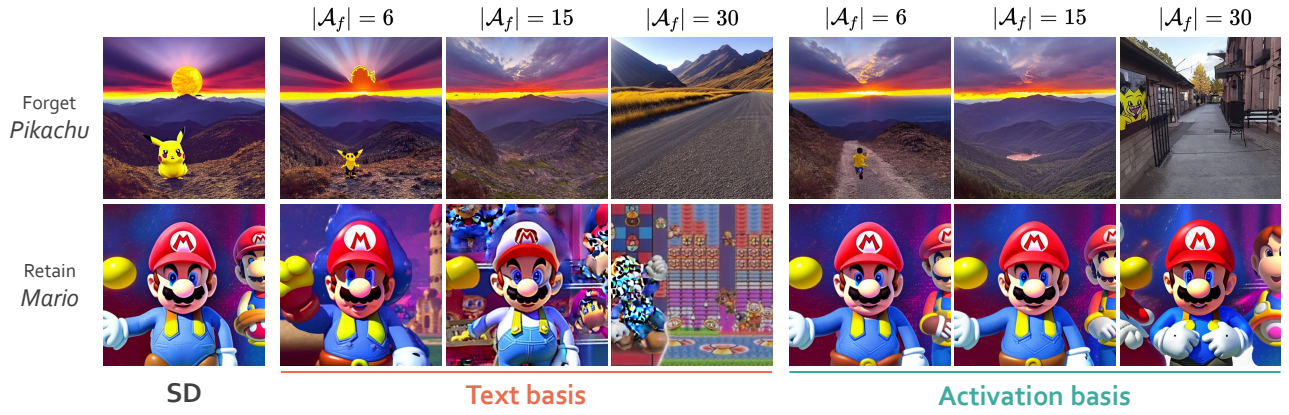


Figure 5. Qualitative comparison as the forget anchor set size $|\mathcal{A}_f|$ increases. Forget prompt: "*Pikachu runs up a mountain with the sun setting behind it.*" Retain prompt: "*Mario standing in a Mushroom Kingdom street.*"

Figure 4a shows several consistent patterns. Natural prompts provide stronger forget anchors than the original templates. Even at $|\mathcal{A}_f| = 6$, replacing template prompts with natural prompts substantially reduces target proportion for both bases, with only a modest reduction in retention.

As the forget set grows, the text basis exhibits a clear trade-off between suppression and retention. Increasing $|\mathcal{A}_f|$ from 6 to 50 lowers target proportion from 0.251 to 0.163, but retention also drops from 0.520 to 0.441. This suggests that adding more text-embedding anchors improves suppression at the cost of greater interference with neighboring concepts. The activation basis behaves differently. Target proportion is already strong at $|\mathcal{A}_f| = 6$ and changes only marginally as more anchors are added, while retention improves from 0.531 to 0.607, approaching the SD reference value.

One possible explanation is that additional prompts refine the activation space estimate of the forget concept, while in text space they may also introduce prompt-level semantic variation that overlaps with retain concepts. The same pattern is also visible qualitatively in Figure 5. As the forget set becomes larger, the text basis increasingly damages the retain concept, while the activation basis better preserves retain generations.

Retain anchor set size. A practical method should remain stable as the retain anchor set $\mathcal{A}_r$ changes, since users cannot enumerate every concept that should be preserved. To study this, we sweep $|\mathcal{A}_r| \in \{0, 36, 54, 120, 180\}$ on Pikachu, corresponding to retain pools of {0, 6, 9, 20, 30} concepts with six prompts each.

For larger retain pools, we manually construct a separate set of 50 character concepts and randomly sample retain concepts from it. These retain concepts are used only during basis construction. The retention evaluation itself remains fixed to the original nine in-category concepts throughout the sweep.

Figure 4b shows a clear difference between the two bases. When the retain set is removed entirely ($|\mathcal{A}_r| = 0$), both bases suppress the target strongly, but retention drops far below the SD reference. This confirms that retain anchors

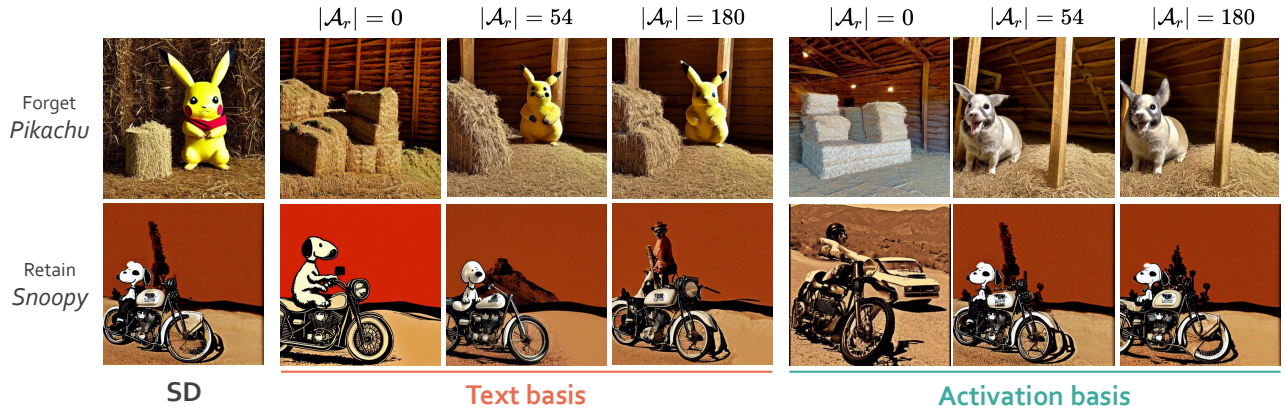


Figure 6. Qualitative comparison as the retain anchor set size $|\mathcal{A}_r|$ increases. Forget prompt: "*Pikachu sitting on a pile of hay in a rustic barn.*" Retain prompt: "*Snoopy sitting on a vintage motorcycle in a sunny desert landscape.*"

Table 2. Target proportion (↓) and within-category retention (↑) of PURE'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

are important for preserving neighboring concepts. As $|\mathcal{A}_r|$ increases, the text basis quickly loses suppression strength. Target proportion rises from 0.079 to 0.469 at $|\mathcal{A}_r| = 54$, and further increases to 0.559 at $|\mathcal{A}_r| = 180$. In contrast, the activation basis degrades much more slowly, reaching 0.221 at $|\mathcal{A}_r| = 180$.

Retention improves as more retain anchors are added, but largely saturates once $|\mathcal{A}_r| \geq 54$. Beyond this point, larger retain sets provide little additional benefit. Across all settings, the activation basis degrades far more slowly than the text basis. The qualitative results in Figure 6 make this difference visually clear. As the retain set increases, the forget concept increasingly reappears under the text basis, whereas the activation basis maintains stronger target suppression.

Design choices. Table 2 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 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 capture 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.

5. Conclusion

We revisit closed-form concept unlearning by changing how the forget direction is constructed. Instead of using text-encoder embeddings as in prior closed-form methods, PURE builds the forget basis from cross-attention activations collected during denoising. A binary probing experiment shows that the activation basis captures natural prompts much more reliably than the text basis. On HUB, which covers ten concepts across four categories, PURE achieves the highest within-category retention in three categories while maintaining target suppression. Overall, this leads to a substantially better suppression-retention trade-off than prior unlearning methods.

Limitations PURE 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 that 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 the activation space or from large prompt collections.

References

Biswas, S. D., Roy, A., and Roy, K. CURE: Concept unlearning via orthogonal representation editing in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Bui, A. T., Vuong, L. T., Doan, K., Le, T., Montague, P., Abraham, T., and Phung, D. Erasing undesirable concepts in diffusion models with adversarial preservation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Chin, Z.-Y., Jiang, C.-M., Huang, C.-C., Chen, P.-Y., and Chiu, W.-C. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *International Conference on Machine Learning (ICML)*, 2024.

Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Gaintseva, T., Oncescu, A.-M., Ma, C., Liu, Z., Benning, M., Slabaugh, G., Deng, J., and Elezi, I. CASteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations (ICLR)*, 2026.

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

Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., and Bau, D. Unified concept editing in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5111–5120, 2024.

Gong, C., Chen, K., Wei, Z., Chen, J., and Jiang, Y.-G. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

Hasty, N., Kroosh, I., Voitekh, D., and Korduban, D. Giphy celebrity detector. https://github.com/Giphy/celeb-detection-oss, 2019.

Heng, A. and Soh, H. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 17170–17194, 2023.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Huang, C.-P., Chang, K.-P., Tsai, C.-T., Lai, Y.-H., Yang, F.-E., and Wang, Y.-C. F. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision (ECCV)*, pp. 360–376, 2024.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Lu, S., Wang, Z., Li, L., Liu, Y., and Kong, A. W. MACE: Mass concept erasure in diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6430–6440, 2024.

Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., and Ding, G. 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.

Moon, S., Lee, M., Park, S., and Kim, D. 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.

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

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., and Zhang, Y. 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.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 36479–36494, 2022.

Schramowski, P., Tauchmann, C., and Kersting, K. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1350–1361, 2022.

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. 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.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. 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.

Srivatsan, K., Shamshad, F., Naseer, M., Patel, V. M., and Nandakumar, K. 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.

Tsai, Y.-L., Hsu, C.-Y., Xie, C., Lin, C.-H., Chen, J. Y., Li, B., Chen, P.-Y., Yu, C.-M., and Huang, C.-Y. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations (ICLR)*, 2024.

Wang, R., Fang, J., Li, J., Chen, H., Shi, J., Wang, K., and Wang, X. ACE: Concept editing in diffusion models without performance degradation. In *ACM International Conference on Multimedia (ACM MM)*, pp. 10525–10534, 2025.

Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., and Liu, S. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 36748–36776, 2024a.

Zhang, Y., Fan, C., Zhang, Y., Yao, Y., Jia, J., Liu, J., Zhang, G., Liu, G., Kompella, R., Liu, X., and Liu, S. UnlearnCanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 96387–96423, 2024b.

Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., and 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 (ECCV)*, 2024c.

A. Implementation Details

This section provides additional implementation details for PURE, 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, Disturbing, Violent

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 from Equation (1) 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. A single-concept edit, including activation capture, per-layer SVD, and the linear update, completes in approximately one minute on a single GPU.

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 PURE ($T = 10$, $n_{\text{lat}} = 10$).

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 PURE. 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.

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.

B. Reproducibility

We will release the codebase upon acceptance, including the implementation of PURE, benchmark configurations, and evaluation scripts for reproducing the main experimental results.

C. 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. PURE 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.

D. Detailed Experimental Results

The main results in Table 1 report category-level averages across concepts. In this section, Tables 4 to 7 provide the corresponding metric values for each individual concept.

Table 4. 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	0.515
ESD	0.029	0.128	0.152	0.013	0.017	0.012	0.148	0.039	0.048	0.222
MACE	0.146	0.279	0.143	0.054	0.022	0.012	0.003	0.000	0.001	0.399
RECELER	0.013	0.032	0.054	0.006	0.007	0.005	0.000	0.058	0.001	0.163
UCE	0.242	0.476	0.398	0.011	0.008	0.015	0.002	0.002	0.001	0.514
CURE	0.464	0.465	0.351	0.524	0.443	0.422	0.483	0.350	0.614	0.512
PURE	0.123	0.281	0.216	0.063	0.139	0.029	0.129	0.104	0.072	0.352

Table 5. 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
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
PURE	0.580	0.619	0.604	0.594	0.604	0.597	0.605	0.614	0.603	0.402

Table 6. 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	0.623
ESD	0.028	0.076	0.082	0.007	0.016	0.004	0.014	0.055	0.013	0.217
MACE	0.114	0.196	0.091	0.033	0.030	0.007	0.001	0.001	0.000	0.221
RECELER	0.020	0.020	0.036	0.012	0.010	0.009	0.002	0.070	0.000	0.206
UCE	0.366	0.367	0.249	0.012	0.015	0.007	0.000	0.000	0.002	0.629
CURE	0.417	0.333	0.396	0.415	0.446	0.023	0.114	0.448	0.248	0.646
PURE	0.129	0.245	0.148	0.095	0.239	0.009	0.016	0.084	0.031	0.335

E. Qualitative Results

In this section, Figures 7 to 9 provide additional qualitative results comparing generations before and after unlearning across different concepts and methods.

Table 7. 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
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
PURE	13.52	13.51	13.65	13.61	13.54	13.58	13.50	13.56	13.59	14.32

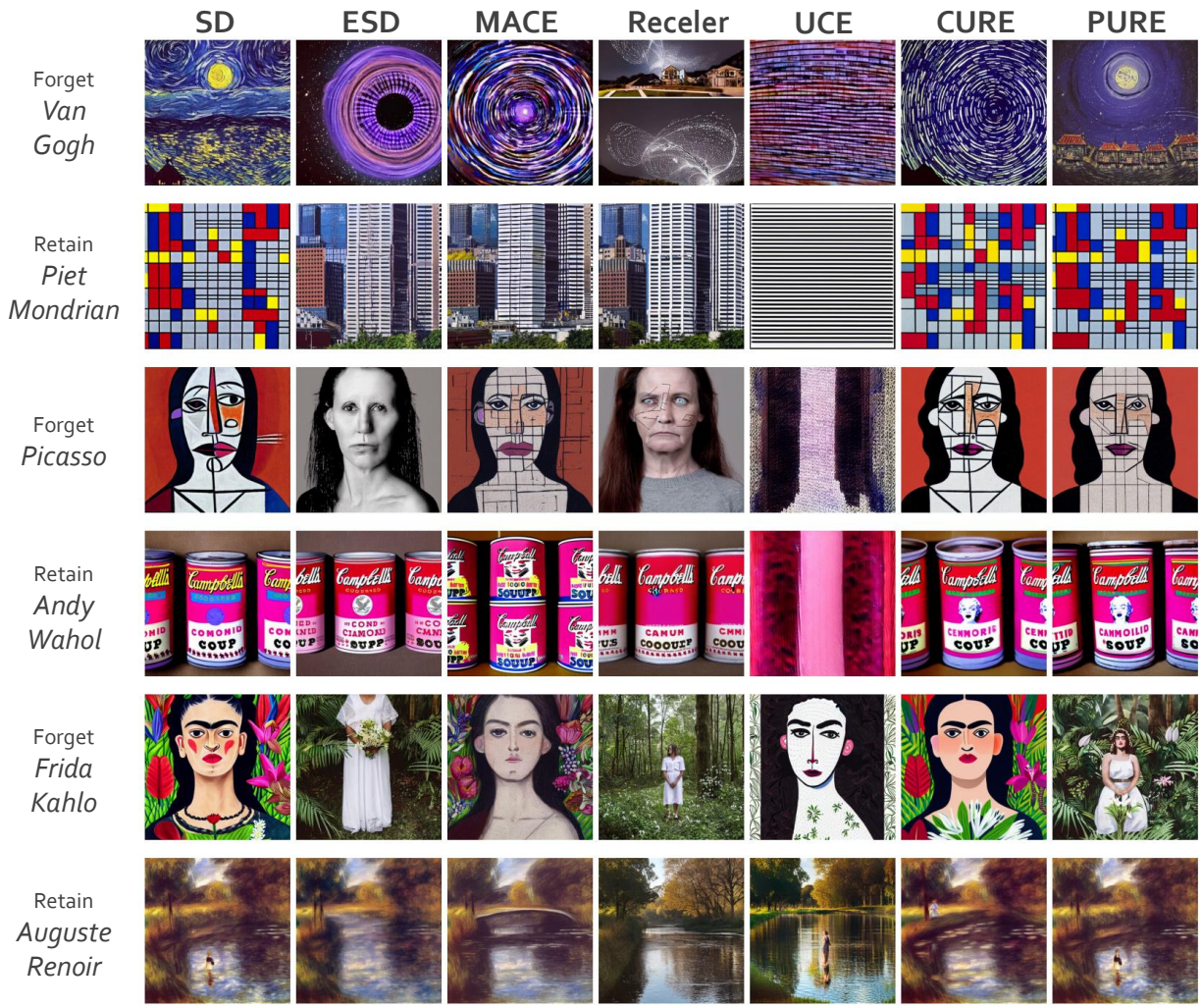


Figure 7. Additional qualitative examples of `Style`.

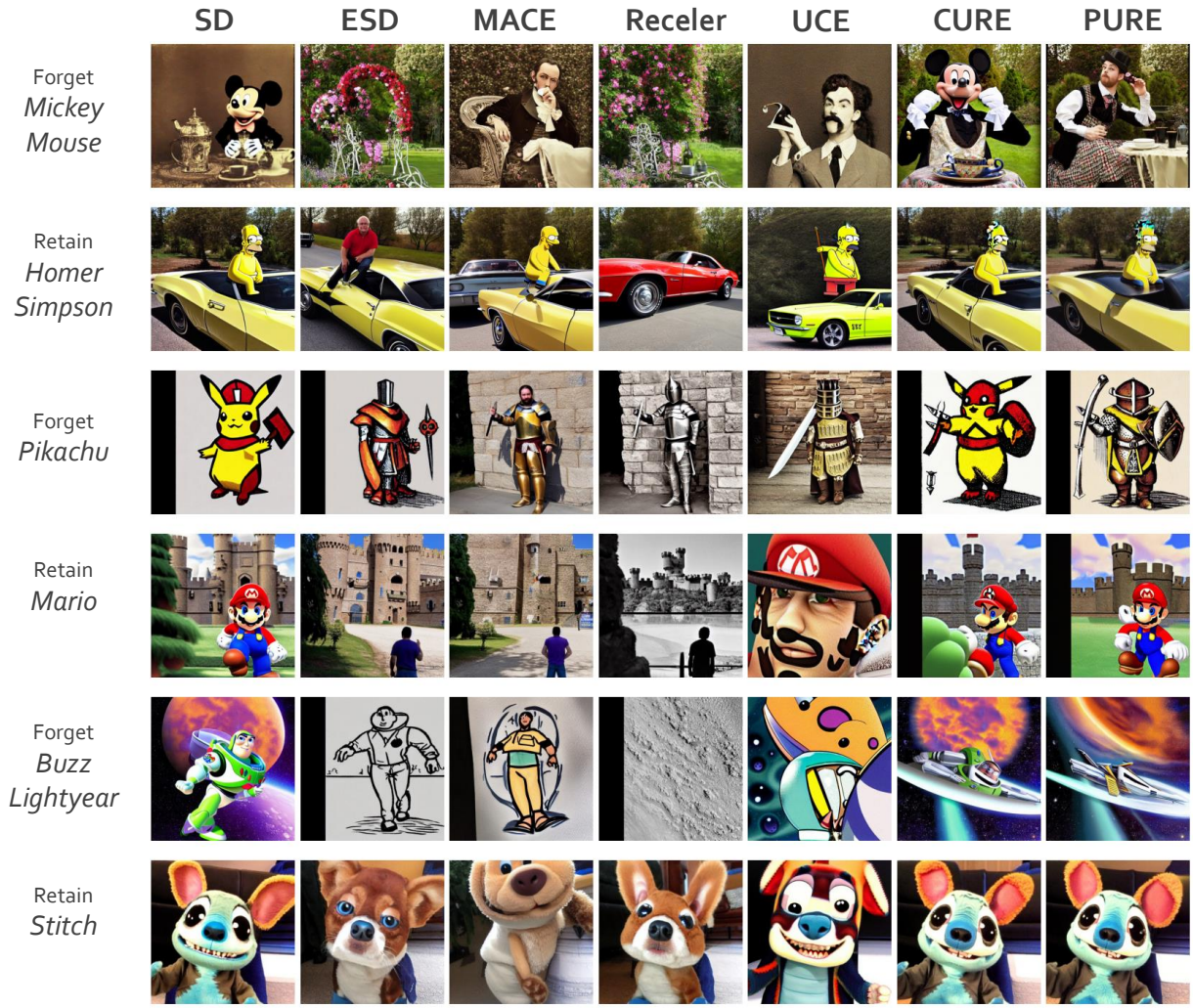


Figure 8. Additional qualitative examples of IP.

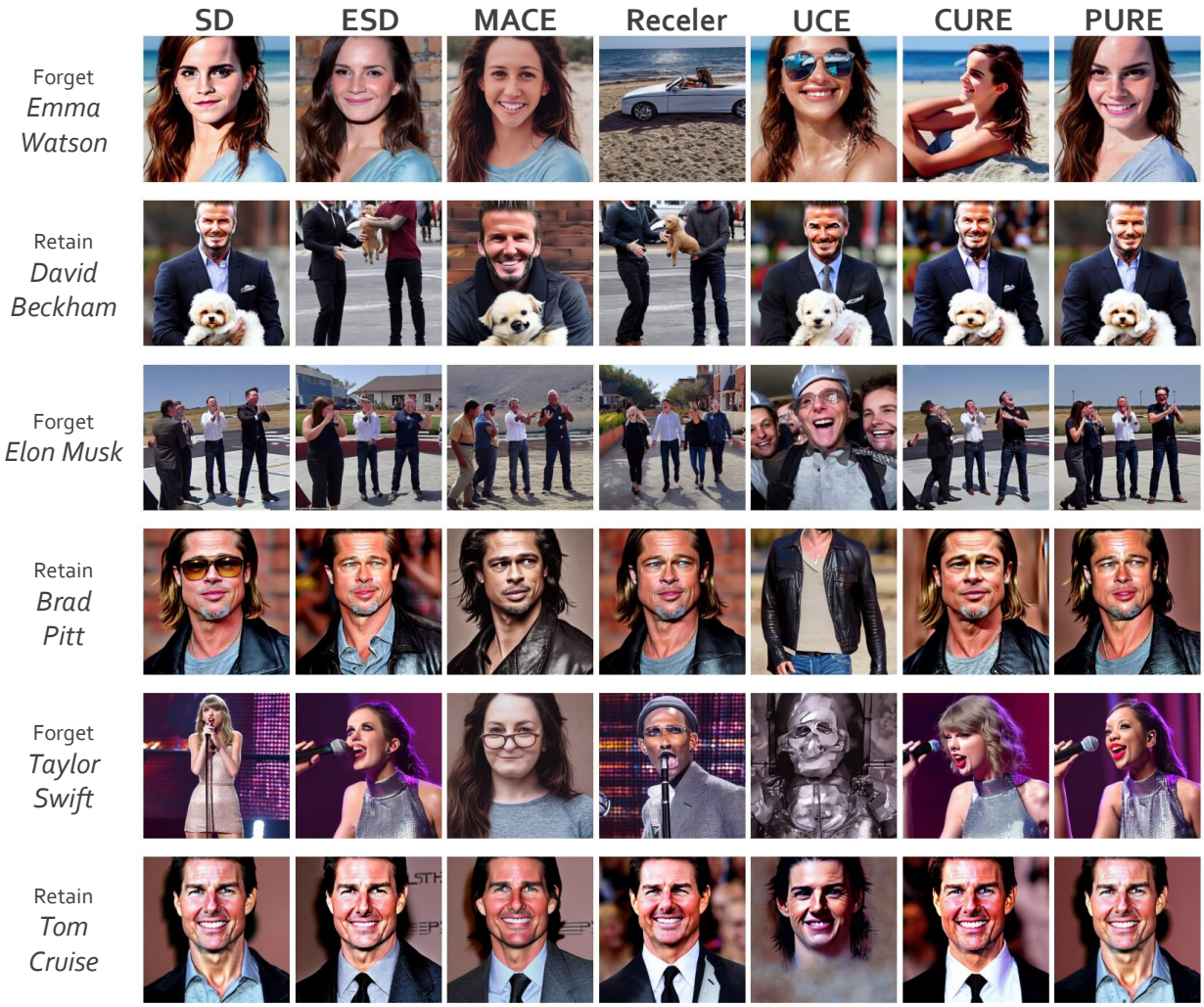


Figure 9. Additional qualitative examples of `Celebrity`.