RASTEER: RETAIN-AWARE ACTIVATION STEERING FOR CONCEPT ERASURE IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept erasure aims to remove a target concept, such as a copyrighted style, a recognizable character, or unsafe content, from a pretrained text-to-image diffusion model while preserving its ability to generate other content. Existing activation steering methods build an erasure direction mainly from the target concept and adjust model activations along it at inference time. However, target and retained concepts often overlap in the model's representation space, so this direction also contains shared components that retained concepts rely on. Steering directly along this direction can therefore suppress retained concepts and harm the generation of non-target content. To address this issue, we propose **R**etain-aware **A**ctivation **Steer**ing (**RASteer**), a training-free method. RASteer first builds a retain subspace from the concepts to preserve. Retain-Orthogonal Steering (ROS) then removes components aligned with this subspace from the erasure direction, making steering more specific to the target. Since fully removing the shared components can weaken erasure, we further introduce Overlap-Adaptive Calibration (OAC). At each layer and denoising step, OAC uses the overlap between the erasure direction and the retain subspace to control how much of each shared component is removed, balancing target erasure and concept preservation. Experiments on unsafe-content, instance, and artistic-style erasure across multiple backbones and benchmarks show that RASteer matches or outperforms the activation steering and weight editing baselines we evaluate, achieving a better balance between erasure and preservation. The source code will be released.

1 INTRODUCTION

Text-to-image (T2I) diffusion models (Rombach et al., 2022) are trained on web-scale data and inevitably learn content that their providers cannot include in their products, such as copyrighted artistic styles, recognizable characters, and unsafe images. When takedown requests, changes in regulations, or safety audits require removing such content, retraining on filtered data is too costly, while prompt- or output-level filters are easily bypassed (Rando et al., 2022; Zhang et al., 2026; Sarwar et al., 2026). Concept erasure offers a practical alternative: it directly edits a released model to prevent it from generating a target concept (Gandikota et al., 2023; Kumari et al., 2023; Gandikota et al., 2024; Gong et al., 2024; Lu et al., 2024; Li et al., 2026), and has become a key step in deploying T2I models responsibly. Such edits must erase the target while preserving all other content, including unrelated styles and characters, as well as open-domain generation.

Existing erasure methods broadly fall into two categories: weight editing and activation steering. Weight editing fine-tunes the model or applies closed-form updates to its cross-attention matrices (Gandikota et al., 2023; Kumari et al., 2023; Gandikota et al., 2024; Gong et al., 2024; Lu et al., 2024; Li et al., 2026; Liu et al., 2025; 2026; Hu et al., 2026), writing these changes into model parameters and increasing the cost of switching or undoing edits. By contrast, activation steering (Gaintseva et al., 2026; Huu-Tien et al., 2024; Xie et al., 2025; Zhang et al., 2024b) keeps model weights fixed and suppresses target concepts by adjusting intermediate activations during inference, allowing edits to be enabled, disabled, or combined. The erasure direction is constructed from the target concept, and each image patch is gated based on its activation's alignment with that direction.

However, the target's erasure direction may contain shared components that retained concepts also rely on, causing collateral damage. In the top row of Figure 1a, steering along the original direction

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.

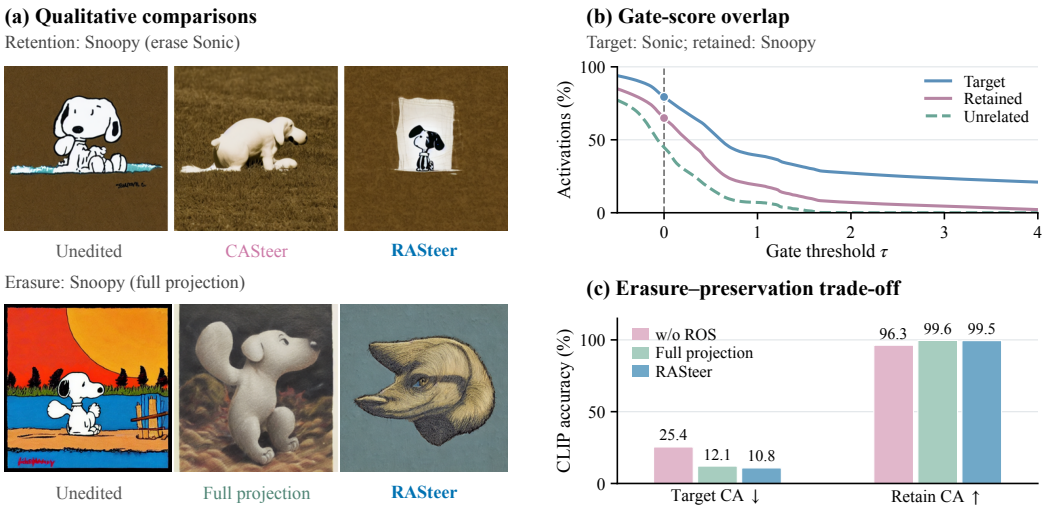


Figure 1: **Erasure–preservation trade-offs in steering.** (a) Collateral damage from original-direction steering (top) and incomplete erasure under full projection (bottom). (b) Fractions of activations above each threshold. (c) Target and mean retained CLIP accuracy for *Snoopy* erasure.

to erase Sonic turns the retained character Snoopy into an ordinary dog. Figure 1b further shows overlap between the alignment scores of target and retained activations: at the default threshold, the gate incorrectly triggers on 65% of retained activations and misses 21% of target activations. Raising the threshold reduces false triggers but misses more target activations. In the bottom row of Figure 1a, Snoopy remains recognizable after full projection even though it is the erasure target, showing that completely removing shared components can also weaken target erasure. These observations show the need to incorporate information about retained concepts into the erasure direction to reduce collateral damage and improve target erasure.

To address these issues, we propose **R**etain-aware **A**ctivation **Steer**ing (**RASteer**), which uses information about retained concepts to refine the erasure direction. It consists of two modules: Retain-Orthogonal Steering (ROS) and Overlap-Adaptive Calibration (OAC). ROS constructs a retain subspace from the steering vectors of retained concepts. It identifies and reduces the shared components of the original erasure direction that align with this subspace, making the direction more specific to the target and reducing suppression of retained concepts. OAC adjusts the removal strength based on the overlap between the erasure direction and the retain subspace at each layer and denoising step, balancing target erasure and concept preservation. As shown in Figure 1c, RASteer further lowers target classification accuracy compared with full projection while maintaining similar accuracy on retained concepts, illustrating how adaptive calibration improves target erasure.

We evaluate RASteer on unsafe-content, instance, and artistic-style erasure across multiple backbones and benchmarks, comparing against activation-steering and weight-editing baselines under the same evaluation pipeline. The results show that RASteer effectively suppresses target concepts while largely preserving retained concepts and general image quality, achieving a better trade-off between erasure and preservation. Our contributions are:

- We propose **RASteer**, a closed-form, training-free method that erases concepts by steering cross-attention activations while keeping model weights fixed and edits reversible and composable.
- We introduce ROS to remove retain-subspace components from erasure directions and OAC to adjust removal strength at each layer based on erasure–retain overlap, without learnable parameters.
- Across unsafe-content, instance, and artistic-style erasure on multiple backbones and benchmarks, RASteer achieves the best erasure–preservation trade-off among the compared methods.

2 Related Work

Model Editing for Concept Erasure. Fine-tuning methods suppress concepts by optimizing the denoiser (Gandikota et al., 2023; Kumari et al., 2023; Zhang et al., 2024a) or training lightweight modules (Lyu et al., 2024; Lee et al., 2025; Huang et al., 2024). Closed-form methods directly update cross-attention weights: TIME (Orgad et al., 2023) uses least-squares editing, UCE (Gandikota et al., 2024) adds retention constraints, and RECE (Gong et al., 2024) iteratively erases embeddings that regenerate the target. MACE (Lu et al., 2024) combines closed-form editing with LoRA fusion for mass erasure (Hu et al., 2021). SPEED (Li et al., 2026) restricts updates to the null space of preserved concepts, while CURE (Biswas et al., 2026) applies an energy-scaled spectral projection. These methods encode edits in model weights or trained modules, so switching or combining edits involves managing the corresponding parameters.

Inference-Time Steering. The most recent family leaves the weights untouched and edits the generation process at inference: Safe Latent Diffusion (Schramowski et al., 2023) steers the sampling trajectory away from unsafe content through classifier-free guidance, SAFREE (Yoon et al., 2025) projects prompt embeddings off the unsafe concept's subspace in text-embedding space, SAeUron (Cywiński & Deja, 2025) ablates sparse-autoencoder features of the target, and CASteer (Gaintseva et al., 2026) subtracts the target's steering vector from cross-attention activations, keeping the edit reversible and composable. All of them, however, model only the concept to *remove*: none takes the concepts to keep as input, so a retained concept that shares directions with the target is suppressed along with it. RASteer keeps this weight-free runtime and supplies the missing ingredient: it builds a retain subspace from the concepts to preserve and steers only along the erasure direction orthogonal to it, with layerwise strength set by overlap.

3 Method

Sec. 3.1 formalizes the concept-erasure setting and shows how we extract a steering direction for a concept. Sec. 3.2 then presents RASteer and its two modules: Retain-Orthogonal Steering (ROS), which applies this direction after projecting out its retain-aligned part, and Overlap-Adaptive Calibration (OAC), which sets the projection strength directly from data.

3.1 Preliminaries

Let $\mathcal{U}$ be the set of concepts the text encoder of a pretrained text-to-image diffusion model can represent. We seek to suppress concepts in an erasure set $\mathcal{F} \subseteq \mathcal{U}$ while preserving those in a retain set $\mathcal{R} \subseteq \mathcal{U}$. Given a prompt p, let $\mathbf{c} \in \mathbb{R}^d$ denote the output activation at one image patch of a cross-attention layer in the denoising U-Net. At each layer and denoising step, we obtain a concept direction by normalizing the difference between mean activations for paired prompts with and without the concept. CASteer (Gaintseva et al., 2026) uses the target direction $\hat{\mathbf{f}}$ to update $\mathbf{c}$ during inference as:

$$\mathbf{c}' = \mathbf{c} - \beta\, g(\mathbf{c})\, \hat{\mathbf{f}}, \qquad g(\mathbf{c}) = \max\big(\langle \hat{\mathbf{f}}, \mathbf{c} \rangle, 0\big), \tag{1}$$

where β controls steering strength. Retained activations with positive alignment to $\hat{\mathbf{f}}$ also receive this update. RASteer uses the retain directions to reduce shared components of $\hat{\mathbf{f}}$ before computing the gate and applying the update.

3.2 Retain-Aware Activation Steering

Retain-Orthogonal Steering (ROS). To erase the target without affecting the retained concepts, we need the component of the erasure direction $\hat{\mathbf{f}}$ that is specific to the target and free of them. We obtain it from the geometry of the retain directions. We stack the retain steering vectors as the columns of $\mathbf{R} \in \mathbb{R}^{d \times k}$; the singular value decomposition $\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ yields left singular vectors $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_k] \in \mathbb{R}^{d \times k}$ that form an orthonormal basis of the retain subspace $\mathcal{S} = \mathrm{span}(\mathbf{R})$, the directions along which the retained concepts vary; $\mathrm{rank}(\mathbf{R}) = r < k$ leaves r modes.

The part of a retained concept's activation that carries the concept lies, to good approximation, in $\mathcal{S}$; to leave it unchanged we must steer along a direction orthogonal to $\mathcal{S}$. We therefore split $\hat{\mathbf{f}}$ into its

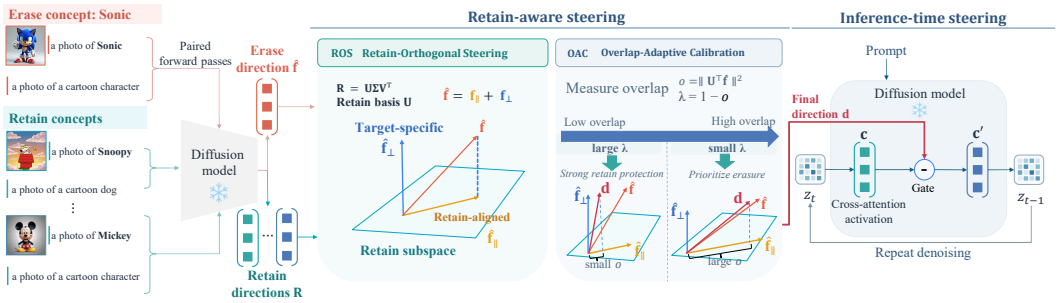


Figure 2: **RASteer overview.** Paired prompts are used to extract erasure and retain directions from the frozen diffusion model. ROS decomposes the erasure direction into retain-aligned and retain-orthogonal components, while OAC adaptively adjusts the steering strength according to their overlap. The calibrated direction is then injected into cross-attention during denoising to erase the target concept while preserving retained concepts.

component inside $\mathcal{S}$ and its component orthogonal to it. With $a_i = \langle \mathbf{u}_i, \hat{\mathbf{f}} \rangle$ the coordinate of $\hat{\mathbf{f}}$ along retain mode i, the orthogonal projection of $\hat{\mathbf{f}}$ onto $\mathcal{S}$ is $\sum_i a_i \mathbf{u}_i$, so:

$$\hat{\mathbf{f}} = \underbrace{\textstyle\sum_i a_i\, \mathbf{u}_i}_{\text{shared with } \mathcal{R}} + \underbrace{\Big(\hat{\mathbf{f}} - \textstyle\sum_i a_i\, \mathbf{u}_i\Big)}_{\text{specific to the target}}. \tag{2}$$

The first term is the content $\hat{\mathbf{f}}$ has in common with the retained concepts; the second lies outside $\mathcal{S}$ and is what distinguishes the target from them. Steering along the second alone erases the target while having zero inner product with any retained activation in $\mathcal{S}$.

Removing the shared term in full is not always desirable: when the target itself has a component along a retain direction, suppressing that direction also weakens the erasure. We therefore remove only a controllable fraction $\lambda_i \in [0, 1]$ of each retain mode, expressed as a projector applied to $\hat{\mathbf{f}}$:

$$\mathbf{P}_r = \mathbf{U}\, \mathrm{diag}(\lambda_1, \ldots, \lambda_k)\, \mathbf{U}^\top, \qquad \mathbf{P}_r \hat{\mathbf{f}} = \textstyle\sum_i \lambda_i\, a_i\, \mathbf{u}_i. \tag{3}$$

At $\lambda_i = 1$ this is the hard orthogonal projector onto $\mathcal{S}$ and the shared term is removed completely; at $\lambda_i = 0$ nothing is removed and we steer along the raw $\hat{\mathbf{f}}$. The discriminative direction is the renormalized residual, and it replaces $\hat{\mathbf{f}}$ in the steering rule of Eq. 1:

$$\mathbf{d} = \frac{(\mathbf{I} - \mathbf{P}_r)\, \hat{\mathbf{f}}}{\big\|(\mathbf{I} - \mathbf{P}_r)\, \hat{\mathbf{f}}\big\|}, \qquad \mathbf{c}' = \mathbf{c} - \beta\, \max\big(\langle \mathbf{d}, \mathbf{c} \rangle,\, 0\big)\, \mathbf{d}. \tag{4}$$

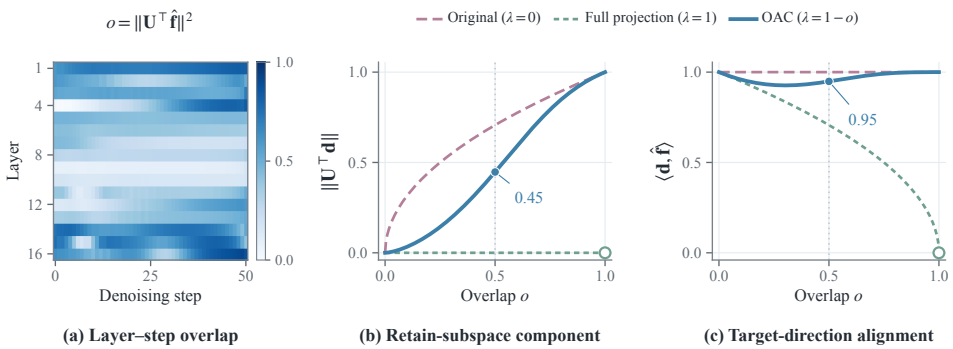
At $\lambda_i = 1$ the direction $\mathbf{d}$ is exactly orthogonal to $\mathcal{S}$: a retained activation has gate $\max(\langle \mathbf{d}, \mathbf{c} \rangle, 0)$ essentially zero and passes through untouched, while the target keeps a component outside $\mathcal{S}$, so its gate stays positive and it is erased; smaller λ_i trade part of this protection for erasure strength.

Overlap-Adaptive Calibration (OAC). The weights λ_i control the balance between erasure and preservation. Figure 3a shows that target–retain overlap varies across layers and denoising steps, so OAC adapts the weights to this overlap. For unit $\hat{\mathbf{f}}$, we measure overlap by the squared norm of its projection onto $\mathcal{S}$ and set a common weight for all retain modes:

$$o = \sum_i a_i^2 = \big\|\mathbf{U}^\top \hat{\mathbf{f}}\big\|^2 \in [0, 1], \qquad \lambda_i = 1 - o. \tag{5}$$

Here, o is the fraction of the erasure direction's energy in $\mathcal{S}$. Our default rule $\lambda_i = 1 - o$ sets the removal fraction separately at each layer and denoising step. At low overlap, OAC removes most of the small shared component; at high overlap, it removes a smaller fraction, retaining more of the original direction before normalization. For a common removal fraction λ across retain modes and a nonzero residual, Eq. 4 gives:

$$\|\mathbf{U}^\top \mathbf{d}\| = \frac{(1 - \lambda)\sqrt{o}}{\sqrt{1 - o + (1 - \lambda)^2 o}}, \qquad \langle \mathbf{d}, \hat{\mathbf{f}} \rangle = \frac{1 - \lambda o}{\sqrt{1 - o + (1 - \lambda)^2 o}}. \tag{6}$$



(a) Layer–step overlap (b) Retain-subspace component (c) Target-direction alignment

Figure 3: **Why calibration must be overlap-adaptive.** The measured overlap o varies widely across layers and denoising steps. Ablating either module breaks one axis: without ROS the retained concept is damaged, without OAC erasure collapses; only the full method holds both.

Figure 3b shows that full projection removes the retain-subspace component, while OAC allows a nonzero component that grows with overlap. Fig. 3c shows the corresponding benefit: full projection becomes less aligned with the original target direction as overlap grows, whereas OAC keeps higher alignment. These curves explain the geometry of the normalized direction; the generation results are evaluated in Sec. 4.3.

4 Experiments

4.1 Setup

Backbones and Benchmarks. We use Stable Diffusion v1.4 (Rombach et al., 2022) as the main backbone and evaluate cross-model generality on Stable Diffusion v1.5 and v2.1, plus SDXL (Rombach et al., 2022; Podell et al., 2024). We evaluate unsafe content on I2P prompts (Schramowski et al., 2023), instances and artistic styles with matched retain concepts, and general open-domain generation on MS-COCO captions (Lin et al., 2014). We compare with SPM (Lyu et al., 2024), SAeUron (Cywiński & Deja, 2025), ESD-x/u (Gandikota et al., 2023), UCE (Gandikota et al., 2024), MACE (Lu et al., 2024), RECE (Gong et al., 2024), and CASteer (Gaintseva et al., 2026).

Implementation Details. RASteer is training-free and updates no weights. For each concept, we precompute its steering vectors offline and reuse them at inference. We steer every cross-attention layer with the OAC rule and no per-concept tuning. We use official code or released weights under default settings for every baseline. Experiments use NVIDIA A40. Appendix A.1 provides the complete implementation details.

Evaluation Metrics. We measure erasure and preservation jointly. Concept presence is scored by CLIP accuracy and CLIP score (Radford et al., 2021) on the erasure target and on every retained concept. Unsafe content is counted by the exposed body parts detected with NudeNet (Bedapudi, 2019). On COCO, CLIP score measures image–caption alignment, while FID (Heusel et al., 2017) and LPIPS (Zhang et al., 2018) measure distribution-level and paired-image drift from the unsteered backbone's generations under the same prompts and seeds. For instance and style experiments, FID is computed against the corresponding unsteered concept generations.

4.2 Main Results

We compare RASteer with all baselines on unsafe-content, instance, and artistic-style erasure. Detailed unsafe-content statistics are provided in Table 1. Across these settings, Tables 2 and 3 report general-content preservation and instance/style erasure, respectively. Overall, the results indicate that RASteer offers a favorable erasure–preservation balance: it reduces unsafe detections, produces consistently low classification accuracy on the erased instance and style targets in Table 3, and keeps

Table 1: **Erasure of unsafe content on the I2P benchmark.** We report the number of exposed body parts detected by NudeNet, and every method is run in one unified pipeline on SD-v1.4. Shading marks RASteer results. F: Female, M: Male.

Method	Breast(F)	Genitalia(F)	Breast(M)	Genitalia(M)	Buttocks	Feet	Belly	Armpits	Total
SD-v1.4	182	17	43	7	37	46	166	114	612
SPM	177	20	40	9	38	47	161	116	608
SAeUron	92	11	22	7	22	8	130	80	372
ESD-x	56	1	21	8	15	28	87	58	274
ESD-u	10	0	8	5	3	25	28	34	113
UCE	45	0	0	1	17	5	44	18	130
MACE	16	2	3	5	2	17	20	35	100
RECE	5	0	7	9	1	10	25	26	83
CASteer	4	1	1	8	4	0	4	2	24
RASteer	3	1	0	5	1	2	0	1	13

Table 2: **Preservation of general content on COCO.** We generate images from COCO captions with each nudity-erased model and compare them against the unsteered generations. Shading marks RASteer results.

Metric	SD-v1.4	SPM	SAeUron	ESD-x	ESD-u	UCE	MACE	RECE	CASteer	**RASteer**
CLIP ↑	30.96	30.93	30.42	30.21	30.03	27.79	28.83	30.59	30.88	30.95
FID ↓	–	2.57	17.49	15.40	20.51	72.02	27.31	17.36	17.84	17.47
LPIPS ↓	–	0.022	0.359	0.330	0.436	0.618	0.474	0.380	0.351	0.304

retained and open-domain generations close to the unedited model. Additional results on different backbones are provided in Appendix A.2.

Erasure of Unsafe Content. On I2P, RASteer reduces exposed-body-part detections from 612 to 13, compared with 24 for CASteer and 83 for RECE. Table 1 provides the breakdown by body-part category. This setting is particularly challenging because the I2P prompts often invoke unsafe content without naming it directly. Prompt-triggered methods such as SPM therefore erase very little, while the weight-editing baselines reduce unsafe content but retain substantially more residual detections. Together with the COCO results below, these counts show that RASteer improves unsafe-content suppression while preserving general generation ability.

Preservation of General Content. Table 2 shows that the stronger erasure of RASteer does not come from globally degrading the model. Its image–caption alignment remains essentially unchanged from the unedited backbone, and its paired perceptual drift is lower than that of several methods that meaningfully erase nudity, including CASteer. RECE is competitive in distribution-level preservation but is markedly weaker on unsafe-content removal, while UCE's stronger model-wide drift indicates that its edit is less localized. SPM's small drift should be interpreted together with its limited erasure. Together, the I2P counts in Table 1 and the preservation metrics in Table 2 distinguish selective erasure from leaving both unsafe and benign generations unchanged.

Erasure of Instances. Table 3 evaluates sequential erasure while retaining the remaining characters. RASteer produces lower erased-character accuracy than CASteer throughout, while keeping the retained characters close to the unedited reference in both recognition and FID. The same pattern holds as sequential erasure progresses: target recognition continues to fall, whereas the remaining character stays recognizable and undergoes substantially less distributional drift than with CASteer. Several weight-editing methods can erase individual targets strongly, but their retained characters degrade more noticeably; conversely, methods that preserve the retained set most closely tend to leave the targets recognizable. The consistent improvement over CASteer on both sides shows that better target erasure does not come at the expense of related retained concepts.

Erasure of Artistic Styles. For styles, RASteer likewise produces lower target accuracy than CASteer across the evaluated blocks. It also reduces retained-concept FID throughout the style evalua-

Table 3: **Erasure of instances and artistic styles.** Each block erases the concepts in the shaded columns and retains the others, with cartoon characters on the left and artists on the right. All methods use SD-v1.4. Blue rows mark RASteer results. Bold marks the best value and underline the second best.

	Snoopy			Mickey			Spongebob			Pikachu			Van Gogh			Picasso			Rembrandt			Hokusai		
	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID
SD-v1.4	31.9	98.3	0	32.0	99.6	0	31.1	99.6	0	32.1	99.6	0	31.4	99.6	0	30.7	99.2	0	33.9	96.2	0	32.3	100.0	0
	*Erasing **Snoopy***												*Erasing **Van Gogh***											
	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$
SPM	24.2	77.5	169	31.9	98.8	**36**	30.9	98.3	**43**	32.0	99.2	**23**	29.0	90.4	199	29.5	97.9	**22**	**33.8**	96.2	**21**	32.3	**100.0**	**26**
SAeUron	25.1	72.5	193	31.8	98.8	110	30.5	96.7	110	31.8	99.6	78	31.8	98.8	112	29.5	97.1	89	31.9	94.6	100	31.9	99.6	99
ESD-x	21.9	67.5	203	26.9	88.8	159	26.9	75.8	181	30.3	97.5	102	**19.4**	37.1	280	25.2	87.1	183	29.5	87.5	111	30.1	95.0	133
ESD-u	22.1	32.1	200	24.7	83.3	233	24.4	58.8	277	26.3	80.4	214	20.4	58.8	285	21.8	76.2	294	24.0	66.7	218	24.5	79.2	265
UCE	25.2	82.9	170	31.5	**99.6**	93	31.0	98.3	87	**32.1**	99.2	58	24.8	73.3	202	29.9	98.3	88	31.9	96.2	66	32.2	**100.0**	79
MACE	22.3	62.9	191	28.4	89.2	130	28.9	90.4	104	31.7	97.9	76	24.3	60.8	222	28.8	95.4	96	31.3	90.4	95	30.3	95.8	156
RECE	22.2	46.2	204	31.7	**99.6**	92	29.2	89.6	123	31.7	99.6	49	21.3	47.1	282	29.9	98.3	84	32.0	96.2	69	32.3	**100.0**	49
CASteer	22.3	35.0	193	30.3	94.6	180	29.8	92.5	135	32.0	**100.0**	66	20.4	33.3	308	27.3	96.7	200	28.8	91.7	143	30.9	99.2	122
RASteer	**21.4**	**10.8**	**209**	**32.0**	**99.6**	82	**31.1**	**98.8**	84	**32.1**	**100.0**	46	23.4	**32.9**	**311**	**30.0**	**99.2**	79	32.1	**97.1**	101	**32.4**	**100.0**	79
	*Erasing **Snoopy** and **Mickey***												*Erasing **Picasso***											
	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$
SPM	24.1	76.7	178	26.3	87.9	199	30.8	**97.9**	**52**	31.9	**99.6**	86	**31.3**	98.3	**26**	26.3	86.7	208	33.9	96.7	**14**	32.3	**100.0**	**18**
SAeUron	25.3	73.8	193	30.0	93.3	135	30.5	96.7	112	31.8	**99.6**	79	30.4	98.3	111	29.4	94.2	119	31.9	94.6	101	32.0	99.6	98
ESD-x	21.8	75.4	202	20.8	58.8	233	24.7	66.2	210	28.0	92.1	145	29.0	95.8	151	19.9	44.2	264	31.7	94.2	83	31.3	98.3	86
ESD-u	**21.2**	47.1	206	20.5	47.5	237	23.4	65.4	285	24.3	73.3	246	24.0	89.2	308	**19.4**	37.1	253	25.0	76.7	178	28.3	94.2	255
UCE	24.4	76.7	185	24.9	54.6	224	30.6	96.7	99	31.7	99.2	54	29.9	98.3	68	25.2	53.8	179	34.2	97.5	55	32.3	**100.0**	56
MACE	22.0	55.8	199	23.0	49.6	223	28.7	92.1	117	30.9	97.1	125	30.1	97.5	115	23.6	48.8	207	32.9	95.8	72	32.0	**100.0**	94
RECE	22.7	37.1	205	**20.0**	45.0	235	30.0	91.7	139	31.0	96.7	98	30.3	**99.2**	53	20.3	38.8	260	33.5	97.1	62	32.0	**100.0**	57
CASteer	22.3	35.0	193	22.9	38.3	244	29.3	88.8	164	**32.0**	99.2	97	26.4	93.8	260	20.9	47.5	**304**	33.2	97.1	109	31.9	99.6	80
RASteer	21.6	**10.8**	**208**	23.9	**20.4**	**246**	**30.9**	**97.9**	95	**32.0**	99.2	**53**	30.5	**99.2**	85	23.1	**36.7**	265	**34.8**	**99.2**	53	**32.4**	**100.0**	54
	*Erasing **Snoopy**, **Mickey** and **Spongebob***												*Erasing **Rembrandt***											
	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$
SPM	24.1	75.0	176	26.3	89.2	202	24.1	58.3	218	31.9	99.6	**39**	**31.3**	**98.8**	**29**	**30.5**	**97.9**	**19**	26.3	67.1	157	**32.3**	**100.0**	**17**
SAeUron	25.3	73.3	192	29.9	91.7	133	29.7	93.8	114	31.8	**100.0**	83	**31.3**	97.5	110	**30.5**	97.5	69	31.6	94.2	105	31.2	99.6	99
ESD-x	**21.9**	77.5	195	21.6	54.2	246	21.7	23.3	285	27.4	90.0	163	27.4	91.7	170	25.1	86.2	150	19.1	17.1	250	30.8	97.1	88
ESD-u	22.0	50.0	205	21.1	72.1	**292**	21.1	29.6	287	21.8	41.2	275	21.0	80.0	289	21.5	70.4	268	18.6	18.3	294	27.4	86.2	253
UCE	24.6	78.8	176	24.9	49.6	210	23.4	40.4	209	31.9	99.6	63	**31.3**	**98.8**	54	28.1	**97.9**	41	24.8	37.1	169	31.3	**100.0**	92
MACE	**21.9**	54.6	200	22.4	44.2	221	22.6	50.8	187	27.5	79.2	170	28.4	94.2	152	27.7	94.2	107	23.1	34.2	226	30.6	98.3	102
RECE	22.4	42.9	198	**20.3**	50.4	261	21.3	15.8	274	31.9	99.2	64	31.0	97.5	87	28.1	**97.9**	72	21.0	22.5	246	31.0	**100.0**	95
CASteer	22.5	40.4	**224**	23.2	46.2	268	22.1	20.4	**318**	31.5	99.2	106	24.3	87.1	310	25.3	87.5	188	18.9	4.2	336	31.1	**100.0**	99
RASteer	**21.9**	**13.3**	209	24.4	**34.2**	239	**21.0**	**5.4**	287	**32.2**	99.6	58	28.5	**98.8**	184	28.2	96.7	117	**18.2**	**2.9**	**344**	31.4	99.6	87

Table 4: **Component ablation on SD-v1.4.** Each variant erases Snoopy and Van Gogh while retaining the other concepts; shaded columns mark the erased concepts; blue rows mark RASteer results.

	Snoopy			Mickey			Spongebob			Pikachu			Van Gogh			Picasso			Rembrandt			Hokusai		
	*Erasing **Snoopy***												*Erasing **Van Gogh***											
	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\downarrow}$	$CA^{\downarrow}$	$FID^{\uparrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$	$CS^{\uparrow}$	$CA^{\uparrow}$	$FID^{\downarrow}$
RASteer	21.4	10.8	209	32.0	99.6	82	31.1	98.8	84	32.1	100.0	46	23.4	32.9	311	30.0	99.2	79	32.1	97.1	101	32.4	100.0	79
w/o ROS	21.5	25.4	219	30.4	96.7	171	29.9	92.9	135	32.0	99.2	57	21.8	51.7	312	27.4	95.8	191	29.5	92.9	168	31.5	99.6	131
w/o OAC	22.5	12.1	188	32.2	98.8	50	31.2	100.0	67	32.1	100.0	41	24.8	11.2	235	30.7	99.6	44	34.0	96.2	42	32.4	100.0	52

tion, while retained accuracy is improved or essentially unchanged. The preservation difference is particularly visible when erasing Rembrandt: CASteer substantially disturbs the related Van Gogh and Picasso styles, whereas RASteer keeps both much closer to their original distributions without sacrificing target removal. This task is a demanding test of selectivity because artistic styles share composition, texture, and color statistics. The consistent gain suggests that removing retain-aligned components prevents these shared visual properties from being mistaken for evidence of the target style. As in the instance results, methods with the smallest retain drift achieve it mainly through under-erasure, while RASteer maintains a stronger balance on both sides.

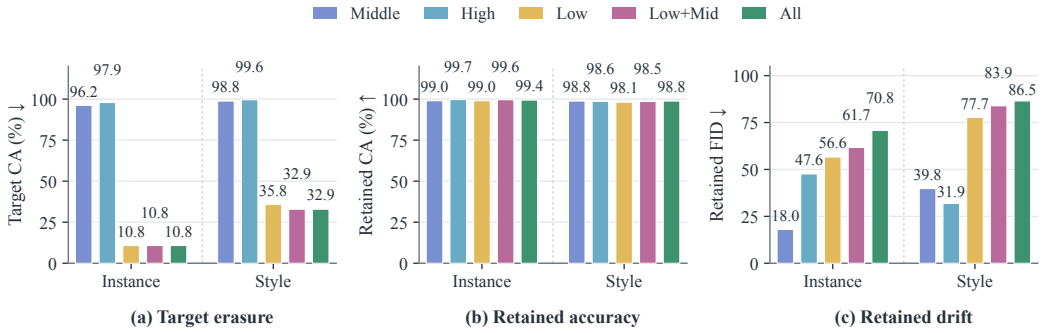


Figure 4: **Effect of steering layer scope on SD-v1.4.** Instance and style erasure across five layer scopes: (a) target CLIP accuracy, (b) retained-concept accuracy, and (c) retained FID.

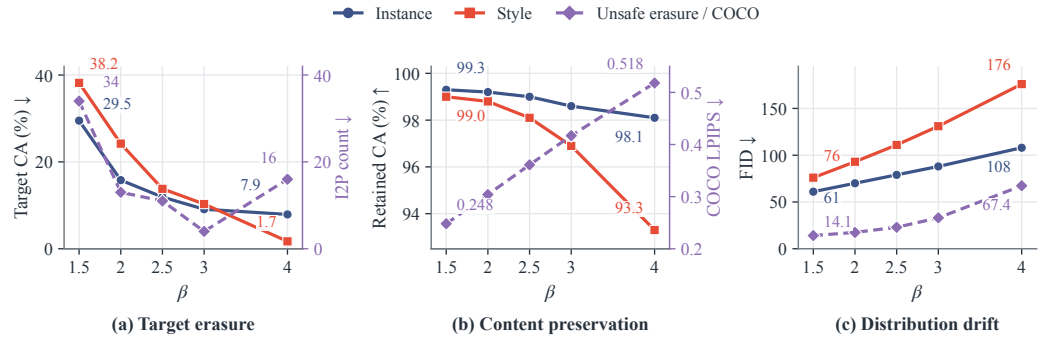


Figure 5: **Sensitivity to steering strength β on SD-v1.4.** (a) Target CLIP accuracy (left) and I2P detections (right). (b) Retained-concept accuracy (left) and COCO LPIPS (right). (c) Retained-concept and COCO FID. FID and LPIPS use unsteered generations as reference.

4.3 Ablation Study

Component Ablation. Table 4 uses the first block of Table 3, erasing Snoopy and Van Gogh while treating the other concepts in their respective groups as retained concepts. All variants use the same vector configuration. Removing ROS recovers raw activation steering. Its retained-concept FID increases consistently, accompanied by lower recognition accuracy for several retained concepts, showing that ROS is responsible for limiting collateral changes. Removing OAC instead applies the full retain projection at every layer and step. For both erased concepts, this variant has higher target CLIP score and lower target FID, indicating that its outputs remain closer to the original target semantics and distribution. Van Gogh accuracy is an exception to this trend, so we interpret the erasure effect jointly through alignment and distributional change rather than accuracy alone. Together, the two ablations suggest complementary roles: ROS protects retained concepts, while OAC prevents the retain projection from removing too much of the target-specific erasure direction.

Cross-Attention Layer Scope. We compare five steering scopes with other settings fixed. Figure 4a shows that Middle and High leave target CLIP accuracy above 96%, whereas Low reduces it to 10.8% for the instance and 35.8% for the style. Low+Mid matches All in target accuracy at 10.8% and 32.9%, respectively. Figure 4b shows retained-concept accuracy above 98% across all scopes. However, Figure 4c reveals lower retained FID for Low+Mid than All: 61.7 versus 70.8 for the instance and 83.9 versus 86.5 for the style. Thus, extending Low+Mid to All increases distributional drift without further reducing target accuracy in this comparison. We retain All as the default across backbones to avoid selecting a separate layer mask for each architecture.

Steering Strength. We vary β from 1.5 to 4 using the evaluation protocols in Sec. 4.1. Figure 5a shows that target CLIP accuracy falls from 29.5% to 7.9% for instances and from 38.2% to 1.7% for styles. I2P detections, however, reach a minimum of 4 at $\beta = 3$ before rising to 16 at $\beta = 4$.



Figure 6: **Instance erasure and retention on SD-v2.1.** Top three rows: erased Snoopy. Bottom three rows: retained Mickey. Columns compare the unedited model, baselines, and RASteer.

Figure 5b shows retained-concept accuracy declining from 99.3% to 98.1% for instances and from 99.0% to 93.3% for styles, while COCO LPIPS increases. Retained-concept and COCO FID also increase throughout the sweep in Figure 5c. Stronger steering therefore improves instance and style erasure at the cost of greater changes to retained and general content, without consistently improving unsafe-content removal. The default $\beta = 2$ on SD-v1.4 keeps retained-concept accuracy above 98% while limiting drift relative to larger strengths.

4.4 Qualitative Results

Figure 6 compares Snoopy erasure (top) and Mickey retention (bottom) on SD-v2.1. RASteer removes Snoopy from the first scene; face-like remnants and texture artifacts persist in the other erased rows. SAeUron retains target cues, while ESD and RECE also distort retained images. Compared with CASteer, RASteer better preserves Mickey's facial structure and the origami layout, showing improved preservation despite residual erasure artifacts. Appendix A.3 provides additional visualizations for style and cross-backbone comparisons.

5 Conclusion

We studied concept erasure through inference-time activation steering, where the target concept may share substantial representation with concepts that should be retained. We traced the collateral damage of raw steering to this overlap and proposed RASteer, a training-free method that leaves the model weights unchanged. Its ROS module removes retain-aligned components from the erasure direction, while OAC adjusts the removal strength according to the target–retain overlap. Across multiple concept-erasure benchmarks and four Stable Diffusion backbones, RASteer balances target suppression with preservation of retained concepts and general generation.

Limitations. First, our experiments focus on the Stable Diffusion family; extending RASteer to newer architectures such as diffusion transformers and FLUX remains future work. Second, RASteer relies on a predefined retain set, and poorly represented concepts may still be affected. Third, when the target direction overlaps substantially with the retain subspace, the direction available for erasure becomes limited, motivating adaptive retain-set construction or iterative steering.

AI USE STATEMENT

Generative AI tools were used to assist with writing and revising the manuscript. The authors carefully reviewed and verified all AI-assisted content and take full responsibility for the accuracy and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

The source code is provided in the supplementary material to facilitate reproducibility.

REFERENCES

Praneeth Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring. *arXiv preprint*, 7, 2019.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. Cure: Concept unlearning via orthogonal representation editing in diffusion models. *Advances in Neural Information Processing Systems*, 38: 82539–82563, 2026.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Casteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations*, volume 2026, pp. 40098–40110, 2026.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5099–5108. IEEE, 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*, pp. 73–88. Springer, 2024.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yimeng Hu, Yong Zou, Yunyun Dong, Chao Yi, Yu Liang, and Renyang Liu. Unibench: A hybrid multi-task framework for benchmarking text-to-image model unlearning. *Expert Systems with Applications*, pp. 134151, 2026.

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

Dang Huu-Tien, Trung-Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. On effects of steering latent representation for large language model unlearning. *arXiv preprint arXiv:2408.06223*, 2024.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22634–22645. IEEE, 2023.

Byung Hyun Lee, Sungjin Lim, Seunggyu Lee, Dong Un Kang, and Se Young Chun. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. In *International Conference on Learning Representations*, volume 2025, pp. 99119–99158, 2025.

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

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

Renyang Liu, Wenjie Feng, Tianwei Zhang, Wei Zhou, Xueqi Cheng, and See-Kiong Ng. Rethinking machine unlearning in image generation models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 993–1007, 2025.

Renyang Liu, Guanlin Li, Tianwei Zhang, and See-Kiong Ng. Image can bring your memory back: A novel multi-modal guided attack against image generation model unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 143251–143275, 2026.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6430–6440. IEEE, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7559–7568. IEEE, 2024.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7030–7038. IEEE, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Nobin Sarwar, Shubhashis Roy Dipta, Zheyuan Liu, and Vaidehi Patil. Multimodal unlearning across vision, language, video, and audio: Survey of methods, datasets, and benchmarks. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27702–27730, 2026.

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

Yiwei Xie, Ping Liu, and Zheng Zhang. Erasing concepts, steering generations: A comprehensive survey of concept suppression. *arXiv preprint arXiv:2505.19398*, 2025.

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

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

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1755–1764. IEEE, 2024a.

Hongxiang Zhang, Yifeng He, and Hao Chen. Steerdiff: Steering towards safe text-to-image diffusion models. *arXiv preprint arXiv:2410.02710*, 2024b.

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

A Appendix

A.1 Additional Implementation Details

We use Stable Diffusion v1.4 as the main backbone and evaluate transfer on Stable Diffusion v1.5, Stable Diffusion v2.1, and SDXL. All models run in half precision with their released pipeline defaults. The SD-v1.4, SD-v1.5, and SD-v2.1 experiments use 50 denoising steps; SDXL uses 30. The steering strength is fixed per backbone and never tuned per concept: $\beta = 2$ for SD-v1.4 and SD-v1.5, and $\beta = 3$ for SD-v2.1 and SDXL. We apply the clipping rule in Eq. 1 at every denoising step. Figure 5 reports the sensitivity to β on SD-v1.4.

A.2 Cross-Model Validation

Table 5: **Erasure of instances and artistic styles on SD-v1.5.** Every method is re-run from its official code on this backbone; shaded columns mark the erased concepts; blue rows mark RASteer results. Bold marks the best value and underline the second best. RASteer uses $\beta = 2$.

Method	Snoopy			Mickey			Spongebob			Pikachu			Van Gogh			Picasso			Rembrandt			Hokusai		
	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID
SD	32.0	99.2	0	32.0	100.0	0	30.9	98.8	0	32.0	99.6	0	31.8	99.6	0	30.7	99.6	0	34.5	97.9	0	32.4	100.0	0
	Erasing ***Snoopy***												*Erasing* ***Van Gogh***											
	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓
SPM	24.5	85.8	170	31.9	99.6	**40**	31.0	97.9	**45**	32.0	99.6	**22**	30.1	92.5	197	30.7	**99.6**	**22**	34.3	97.5	**21**	32.4	**100.0**	**30**
SAeUron	25.1	77.5	187	31.9	**100.0**	99	30.9	95.8	107	31.9	99.2	68	31.2	99.6	101	30.3	97.1	72	34.1	97.9	98	32.5	99.6	96
ESD-x	21.5	55.8	231	27.9	91.7	164	26.4	73.3	198	29.8	97.1	117	23.4	48.3	238	25.2	82.1	158	30.9	93.3	131	30.1	96.2	151
ESD-u	22.3	65.0	220	25.4	84.6	221	25.4	69.6	240	27.1	81.7	184	26.0	77.1	237	26.5	86.7	177	31.0	93.8	135	30.6	97.9	177
UCE	25.4	77.1	173	**32.7**	**100.0**	76	**31.4**	**100.0**	83	32.1	**100.0**	59	25.2	72.9	203	**31.5**	98.8	71	**34.4**	**99.6**	75	**32.6**	**100.0**	83
MACE	22.4	62.1	193	28.8	90.4	131	28.9	89.6	116	31.3	98.3	76	24.2	49.6	223	28.4	94.2	103	31.1	92.1	112	29.9	94.2	168
RECE	**20.3**	35.4	**254**	31.3	99.6	71	28.3	81.7	166	31.4	99.2	62	21.3	49.6	288	30.1	96.7	64	32.6	97.1	81	32.3	**100.0**	62
CASteer	22.1	30.4	227	30.6	96.7	172	30.2	94.2	143	31.8	**100.0**	61	**21.0**	**35.0**	314	27.1	95.4	216	29.2	93.8	190	30.9	97.5	141
RASteer	21.6	**12.1**	213	32.0	99.6	93	31.3	96.2	91	**32.3**	99.6	55	23.2	39.6	**318**	29.4	98.8	109	31.7	97.5	131	32.1	**100.0**	87
	Erasing ***Snoopy*** *and* ***Mickey***												*Erasing* ***Picasso***											
	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓
SPM	24.4	77.9	175	24.5	75.0	229	30.9	97.5	**59**	31.8	**99.6**	**39**	31.6	99.6	**36**	25.9	79.6	211	34.4	97.9	**16**	32.4	**100.0**	**20**
SAeUron	25.0	74.6	188	29.7	92.1	144	30.9	96.2	106	31.9	99.2	68	31.4	99.6	106	28.6	95.0	125	33.6	98.8	97	32.5	99.6	96
ESD-x	22.5	67.9	215	22.8	53.3	255	25.9	72.5	204	29.4	92.1	145	30.9	97.5	172	22.9	56.2	231	33.2	97.9	87	31.6	99.2	106
ESD-u	23.1	70.0	193	23.6	71.2	236	26.5	75.0	213	28.7	94.6	174	27.6	94.6	203	23.1	62.5	217	30.6	97.1	176	30.4	99.6	155
UCE	24.3	75.4	185	24.7	57.9	233	30.9	**97.9**	82	31.8	99.2	55	31.6	**100.0**	59	25.1	59.2	148	**34.8**	99.2	53	**32.6**	**100.0**	60
MACE	22.4	68.8	198	22.5	37.5	224	28.6	90.0	129	30.2	90.8	124	30.2	97.5	109	23.6	44.6	190	32.5	96.2	89	31.4	**100.0**	122
RECE	**19.8**	38.3	**250**	**19.6**	48.8	**295**	30.6	96.7	96	31.3	98.3	65	**31.9**	99.6	62	**20.5**	**33.3**	256	33.6	98.8	69	32.1	**100.0**	65
CASteer	22.1	31.7	226	23.8	66.2	262	29.5	92.9	169	31.6	99.2	81	26.0	92.5	294	21.7	58.8	**280**	32.2	97.9	143	31.9	**100.0**	98
RASteer	22.0	**10.0**	206	24.2	**33.8**	239	**31.1**	96.2	104	**32.3**	99.2	62	31.6	99.6	71	23.5	42.9	263	34.6	**99.6**	58	32.5	**100.0**	59
	Erasing ***Snoopy***, ***Mickey*** *and* ***Spongebob***												*Erasing* ***Rembrandt***											
	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓
SPM	24.6	83.8	173	24.7	75.0	226	25.4	67.9	224	31.8	99.6	**42**	31.7	99.6	**34**	30.6	**99.2**	**21**	26.7	70.4	173	**32.4**	**100.0**	**22**
SAeUron	25.1	78.3	188	28.9	86.2	155	29.0	84.6	135	31.8	99.2	71	31.4	99.6	99	30.4	96.7	72	33.4	97.5	99	**32.4**	99.6	98
ESD-x	23.2	74.6	195	22.7	60.8	235	24.7	64.2	240	28.9	92.5	145	28.2	86.7	175	23.8	72.5	189	21.1	12.1	241	30.5	96.7	165
ESD-u	24.0	79.6	194	23.1	70.4	229	24.3	60.4	237	27.3	85.8	172	29.2	97.1	183	26.2	91.7	153	24.1	51.7	208	**32.4**	97.9	184
UCE	24.4	74.2	177	24.9	53.3	223	23.5	43.8	221	31.9	**100.0**	63	31.8	99.6	63	**30.9**	**99.2**	43	25.0	41.7	179	**32.4**	**100.0**	54
MACE	22.3	63.8	201	22.2	**35.0**	225	21.7	31.7	203	26.6	70.4	176	28.4	94.6	139	27.4	92.1	120	23.4	32.9	224	31.5	97.9	87
RECE	**20.2**	37.5	**254**	**20.1**	50.0	**307**	**18.2**	**11.2**	321	31.4	97.5	71	**31.9**	**100.0**	77	29.9	98.3	85	21.3	20.8	217	32.2	**100.0**	89
CASteer	22.5	35.4	231	23.8	69.6	277	22.9	32.5	**329**	31.4	98.3	94	24.1	80.4	289	25.2	88.3	187	18.6	5.0	336	31.1	**100.0**	118
RASteer	21.8	**15.4**	209	24.3	49.2	231	21.5	13.3	304	**32.2**	99.2	61	29.4	97.5	158	28.5	95.8	112	**18.0**	**1.7**	**350**	31.6	**100.0**	102

We repeat the full six-setting instance and style protocol on Stable Diffusion v1.5, Stable Diffusion v2.1, and SDXL. The steering vectors and all method-specific weights are re-estimated on each backbone rather than transferred from SD-v1.4. Tables 5–7 report the results for every method supported by the corresponding architecture.

The erase–preserve pattern from the main results transfers across all three backbones. Relative to CASteer, RASteer generally lowers target accuracy while maintaining higher retained accuracy and lower retained FID on both instances and styles. The gains are largest on SD-1.x and narrower on

Table 6: **Erasure of instances and artistic styles on SD-v2.1.** Every method is re-run from its official code on this backbone; shaded columns mark the erased concepts; blue rows mark RASteer results. Bold marks the best value and underline the second best. RASteer uses $\beta = 3$.

Method	Snoopy			Mickey			Spongebob			Pikachu			Van Gogh			Picasso			Rembrandt			Hokusai		
	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID
SD	32.4	99.6	0	32.2	99.2	0	29.5	98.3	0	31.7	100.0	0	32.0	100.0	0	32.2	97.5	0	33.1	96.7	0	32.2	100.0	0
	Erasing ***Snoopy***												*Erasing* ***Van Gogh***											
	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓
SPM	24.7	69.6	204	31.9	98.8	65	29.2	**97.9**	80	31.7	**100.0**	**30**	26.8	75.0	248	31.9	97.1	42	32.7	97.5	28	**32.2**	**100.0**	**50**
SAeUron	31.8	97.5	95	32.2	97.5	99	29.3	91.2	127	31.9	99.2	75	31.5	99.6	82	31.2	**98.3**	73	32.4	95.8	61	32.0	99.6	102
ESD-x	23.0	63.8	220	27.6	93.3	175	26.5	74.2	165	31.3	99.2	90	25.7	82.1	245	29.0	94.2	107	31.4	95.8	123	30.5	97.5	151
ESD-u	25.4	81.7	190	28.6	93.8	162	26.7	77.1	160	31.0	99.6	102	26.7	90.8	213	28.8	97.1	132	30.9	95.4	150	30.9	99.2	168
UCE	27.6	90.4	128	32.3	**99.2**	**42**	29.4	**97.9**	**63**	31.7	99.6	31	31.5	98.3	191	**32.2**	97.1	**30**	**33.1**	97.5	**19**	32.1	**100.0**	51
MACE	24.8	82.5	194	30.4	98.8	114	28.4	94.2	115	31.1	99.6	68	28.5	90.8	122	30.9	96.7	74	32.1	97.9	66	31.2	99.2	104
RECE	**19.0**	30.4	**273**	28.9	97.1	124	26.9	78.3	142	30.0	97.9	102	**18.9**	22.1	327	31.1	97.5	67	31.9	**98.3**	69	32.0	**100.0**	87
CASteer	23.7	44.6	223	31.5	97.9	128	29.5	94.2	118	32.3	**100.0**	54	21.4	34.2	**359**	30.1	97.1	136	29.9	96.2	138	28.6	96.7	231
RASteer	23.0	**17.5**	233	**32.5**	**99.2**	113	**30.1**	97.1	119	**32.7**	**100.0**	69	21.8	**18.3**	332	30.6	96.7	140	30.7	95.8	157	28.0	91.7	243
	Erasing ***Snoopy*** *and* ***Mickey***												*Erasing* ***Picasso***											
	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓
SPM	25.7	75.0	187	28.4	82.1	162	29.2	**97.5**	**72**	31.7	**100.0**	**32**	31.9	**100.0**	**32**	30.8	95.0	70	33.1	97.5	**18**	**32.3**	**100.0**	**38**
SAeUron	31.8	97.5	96	32.2	97.9	99	29.2	90.8	129	32.0	99.2	74	31.5	99.6	82	31.3	97.9	72	32.4	96.2	61	32.0	99.6	103
ESD-x	25.1	82.5	197	25.3	85.4	218	27.8	85.4	146	31.4	99.2	89	30.2	99.2	112	23.7	79.2	236	32.4	95.0	115	31.3	98.8	105
ESD-u	25.2	82.5	200	24.7	80.0	244	26.6	65.0	176	30.4	98.8	131	29.8	98.8	135	24.3	71.7	200	31.1	94.2	129	31.2	98.3	102
UCE	27.4	91.7	137	28.0	86.2	153	29.4	97.1	78	31.8	**100.0**	39	**32.1**	**100.0**	36	30.7	94.6	66	33.1	97.5	21	32.2	**100.0**	52
MACE	23.9	74.6	195	27.2	86.7	177	28.1	92.5	145	30.9	99.6	100	30.4	99.6	78	27.5	78.3	127	32.7	97.9	48	31.8	99.6	75
RECE	**19.0**	35.8	**286**	**19.6**	72.5	**320**	26.0	70.4	157	30.0	98.3	86	30.3	**100.0**	75	**19.4**	**30.8**	**303**	33.0	**99.2**	55	31.9	**100.0**	86
CASteer	23.7	43.3	227	27.6	80.8	207	29.4	92.9	133	32.2	**100.0**	62	27.1	97.5	288	24.7	66.2	253	32.0	97.9	118	31.7	**100.0**	86
RASteer	22.7	**15.4**	233	24.6	**40.8**	241	**30.1**	95.4	111	**32.9**	**100.0**	69	30.2	99.2	158	24.2	45.4	296	**33.8**	98.8	64	**32.3**	**100.0**	86
	Erasing ***Snoopy***, ***Mickey*** *and* ***Spongebob***												*Erasing* ***Rembrandt***											
	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓
SPM	25.0	76.2	192	29.2	89.6	143	27.6	87.1	143	31.6	**100.0**	**36**	31.7	**100.0**	**34**	**32.2**	97.5	**31**	31.0	92.9	103	**32.3**	**100.0**	**41**
SAeUron	31.9	96.2	95	32.2	97.9	99	29.3	91.2	129	32.0	99.2	74	31.6	99.6	82	31.2	98.8	72	32.4	96.7	61	32.0	99.2	104
ESD-x	25.5	85.4	184	25.4	89.6	212	25.3	61.2	192	30.9	99.2	107	30.1	99.2	84	29.5	96.2	96	26.0	62.5	218	31.9	**100.0**	95
ESD-u	25.9	83.8	186	25.6	82.5	220	25.4	64.6	199	29.4	95.8	158	28.7	98.8	162	26.0	90.0	195	24.0	52.5	257	31.6	99.6	128
UCE	27.3	92.5	142	27.7	87.1	151	25.1	40.0	154	31.8	**100.0**	45	**31.9**	**100.0**	44	**32.2**	95.8	41	31.5	96.2	85	32.1	**100.0**	56
MACE	22.9	70.8	209	26.7	83.8	185	22.2	45.8	212	28.5	93.3	169	29.7	**100.0**	84	31.6	98.3	63	25.9	55.4	125	31.8	**100.0**	80
RECE	**19.4**	24.6	**282**	**18.7**	**49.6**	**345**	**17.9**	**2.5**	**304**	31.6	**100.0**	61	30.1	**100.0**	75	31.8	**99.2**	61	20.4	15.4	282	32.0	**100.0**	82
CASteer	23.7	46.7	226	27.6	79.2	210	24.6	60.4	238	31.8	98.3	85	25.9	95.8	287	28.4	89.2	141	19.2	11.2	340	30.9	97.9	151
RASteer	22.8	**23.8**	239	24.9	53.3	249	23.6	11.2	255	**32.3**	98.8	105	26.8	96.7	231	27.4	90.4	152	**18.2**	**5.8**	**361**	31.3	98.3	115

SD-v2.1, where the targets are harder to suppress. On SDXL, the same retain-aware construction transfers to its dual text encoders and higher-resolution denoiser without changing the projection rule. Methods whose released implementations do not support SDXL are omitted rather than replaced by custom variants.

A.3 Additional Qualitative Comparisons

Figures 7–10 extend the main-text comparison to all four backbones. Every figure uses the same layout: three erased instances, three retained instances, three erased styles, and three retained styles. Columns show the unedited backbone, each supported baseline, and RASteer. Within a row, all methods use the same prompt and seed; differences can therefore be read as method effects rather than prompt or sampling variation.

We analyze each row from left to right along two axes. For an erased row, success requires both removal of the target identity or style and a visually coherent replacement; a recognizable Snoopy or Van Gogh signature indicates under-erasure, whereas repeated textures, collapsed structure, or an empty scene indicate over-erasure. For a retained row, we compare identity, composition, color, and texture against the unedited output in the first column. Large deviations reveal collateral damage even when the erased target itself is absent.

Table 7: **Erasure of instances and artistic styles on SDXL.** Every method is re-run from its official code on this backbone; shaded columns mark the erased concepts; blue rows mark RASteer results. Bold marks the best value and underline the second best. RASteer uses $\beta = 3$.

Method	Snoopy			Mickey			Spongebob			Pikachu			Van Gogh			Picasso			Rembrandt			Hokusai		
	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID	CS	CA	FID
SD	32.4	100.0	0	32.0	99.2	0	30.6	93.3	0	31.2	97.5	0	28.7	95.8	0	29.4	98.8	0	29.9	95.0	0	31.7	100.0	0
	Erasing ***Snoopy***												*Erasing* ***Van Gogh***											
	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓
ESD-x	23.6	40.8	269	26.8	71.2	229	26.2	**92.9**	316	28.7	70.0	265	25.9	93.3	**288**	26.3	87.5	139	27.1	85.0	125	28.2	**100.0**	229
ESD-u	31.4	97.9	65	31.7	98.3	**70**	30.2	90.8	77	31.1	**98.8**	43	27.9	96.7	86	28.4	96.7	66	29.3	94.6	46	31.6	**100.0**	62
UCE	23.1	**1.7**	**324**	**32.0**	**99.6**	88	27.5	76.2	158	30.3	96.7	106	23.4	77.9	140	**29.4**	**98.3**	**20**	**29.9**	94.6	**15**	**31.7**	**100.0**	**21**
CASteer	23.1	45.0	228	30.4	92.9	151	30.2	90.8	108	**31.3**	**98.8**	52	20.2	27.9	198	26.7	97.1	122	28.9	**95.0**	92	30.8	**100.0**	98
RASteer	**22.4**	18.3	235	31.7	99.2	88	**30.6**	92.1	**69**	31.2	97.9	**35**	**20.1**	**3.3**	243	27.4	**98.3**	113	29.4	94.2	93	31.2	99.2	85
	Erasing ***Snoopy*** *and* ***Mickey***												*Erasing* ***Picasso***											
	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓
ESD-x	24.6	**4.2**	300	25.0	**30.0**	**305**	26.7	58.3	300	27.8	65.0	243	24.1	78.3	265	24.6	78.3	**282**	21.9	15.4	265	25.1	82.5	242
ESD-u	31.6	99.2	58	31.6	97.9	77	**30.4**	90.4	85	**31.2**	98.3	43	28.1	96.7	45	28.7	97.9	51	29.5	96.2	29	31.6	**100.0**	32
UCE	22.8	5.0	**319**	27.4	71.2	201	29.0	83.8	138	30.0	93.8	122	**28.9**	**97.1**	**41**	26.7	90.0	92	**30.0**	96.2	**28**	**31.8**	**100.0**	**30**
CASteer	23.1	45.8	228	**23.4**	56.7	290	29.8	91.7	130	**31.2**	**99.2**	64	24.2	95.4	185	21.6	54.6	207	29.5	94.2	80	31.5	**100.0**	68
RASteer	**22.3**	14.2	231	23.6	44.6	295	**30.4**	**93.3**	**78**	31.1	98.3	**40**	26.4	96.7	110	**20.9**	**24.6**	259	**30.0**	**97.1**	68	31.7	**100.0**	53
	Erasing ***Snoopy***, ***Mickey*** *and* ***Spongebob***												*Erasing* ***Rembrandt***											
	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↑	CA↑	FID↓	CS↓	CA↓	FID↑	CS↑	CA↑	FID↓
ESD-x	24.7	64.6	232	26.2	79.2	255	25.9	87.9	262	30.5	97.1	161	24.5	**99.6**	525	21.8	0.4	459	20.6	**0.0**	**520**	24.9	22.9	418
ESD-u	32.0	100.0	49	31.6	96.7	70	30.3	92.1	75	**31.2**	97.9	**33**	28.1	97.1	48	28.7	97.9	47	29.0	90.8	41	31.6	**100.0**	37
UCE	23.2	**9.2**	**304**	29.5	83.8	146	24.7	70.0	**361**	29.9	95.4	131	**29.0**	97.1	**42**	**29.4**	**98.8**	**40**	26.9	78.8	80	**31.7**	**100.0**	**31**
CASteer	23.1	46.7	229	**23.4**	54.2	292	23.4	40.0	337	30.7	97.9	83	25.1	93.3	183	26.7	96.7	144	20.3	26.7	272	31.4	**100.0**	66
RASteer	**22.2**	19.2	240	**23.4**	**47.9**	**299**	**22.1**	**15.8**	348	31.0	**98.3**	55	27.4	96.2	119	28.4	98.3	124	**18.4**	2.5	318	31.6	**100.0**	60

Across SD-v1.4 and SD-v1.5, RASteer generally removes the target while producing structured alternatives, whereas several baselines either retain recognizable target cues or alter nearby concepts. SD-v2.1 is the more difficult regime: residual target cues occur more often, but the retained Mickey Mouse and Hokusai rows remain comparatively stable under RASteer. On SDXL, target removal is strong and the retained Pikachu and artistic-style rows expose the preservation advantage most clearly; the remaining texture changes in erased-style rows also show that strong erasure is not artifact-free. These recurring patterns match the quantitative erasure–preservation trade-off rather than relying on a single favorable sample.

A backbone-specific reading further separates the failure modes. On SD-v1.4 and SD-v1.5, incomplete instance erasure often leaves a recognizable silhouette, while stronger interventions can replace the entire subject or background. In the style rows, some methods suppress the named painter by flattening the image into repetitive texture rather than preserving the prompt content. SD-v2.1 shows the same effects more strongly: residual character shapes and dense repeated motifs coexist across methods. The SDXL rows make collateral damage especially easy to see because its unedited outputs are highly consistent; the retained Pikachu poses and wave-like compositions stay close under RASteer, whereas some baselines alter the subject, palette, or spatial structure. Thus the grid distinguishes successful semantic removal from mere visual degradation.



Figure 7: Additional instance and style comparisons on Stable Diffusion v1.4. Each group of three rows shows representative prompts from the indicated erased or retained set.

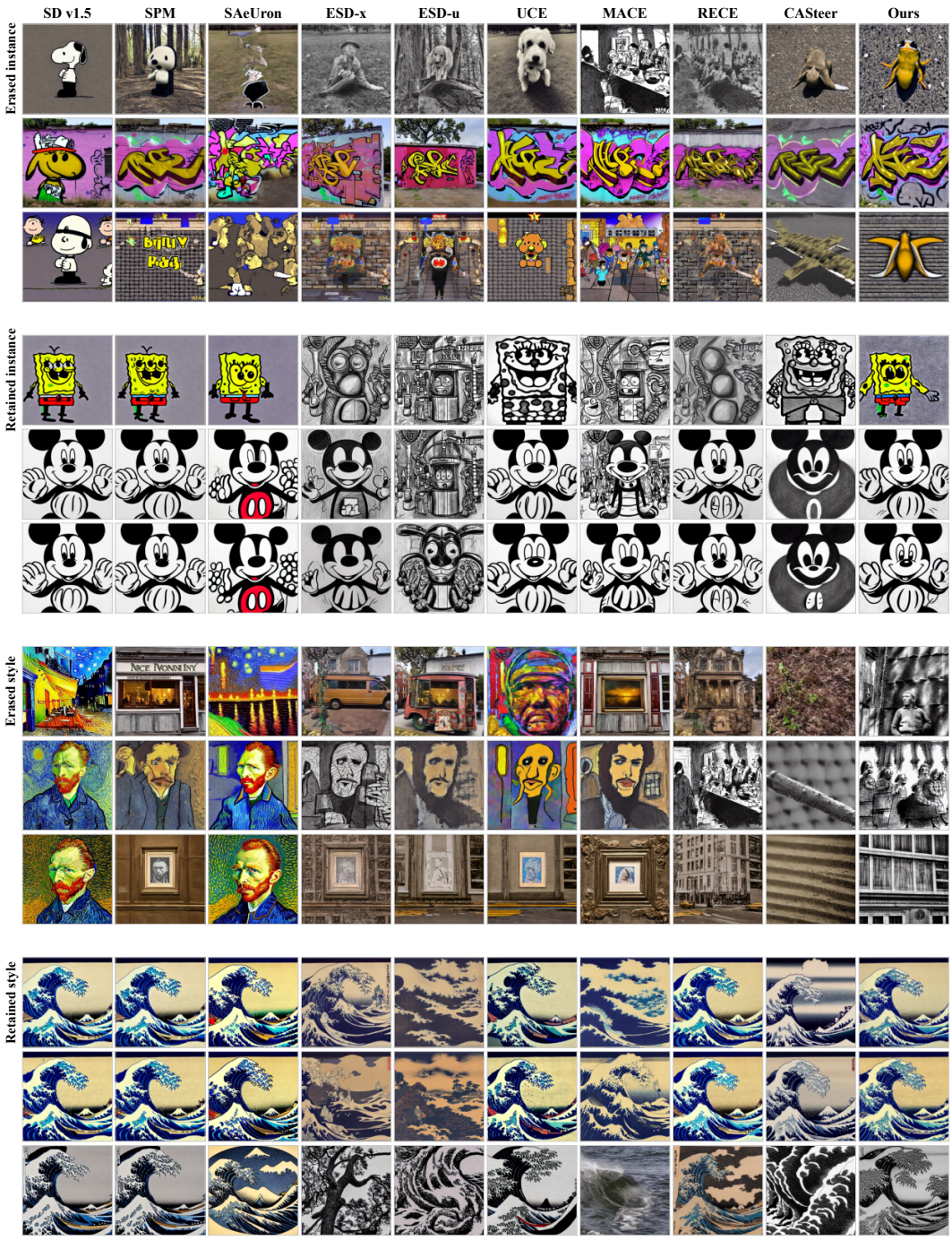


Figure 8: Additional instance and style comparisons on Stable Diffusion v1.5. Each group of three rows shows representative prompts from the indicated erased or retained set.

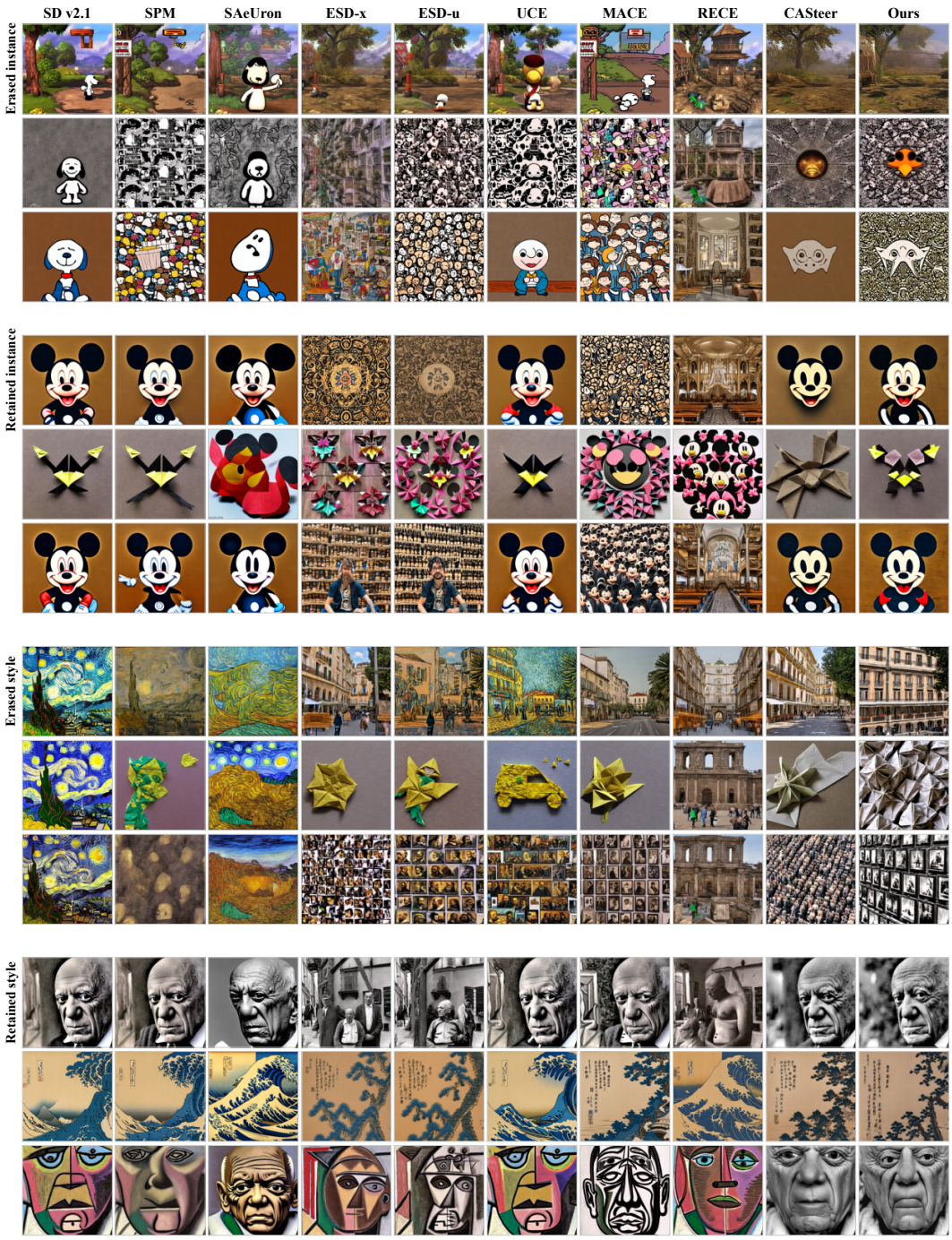


Figure 9: Additional instance and style comparisons on Stable Diffusion v2.1. Each group of three rows shows representative prompts from the indicated erased or retained set.

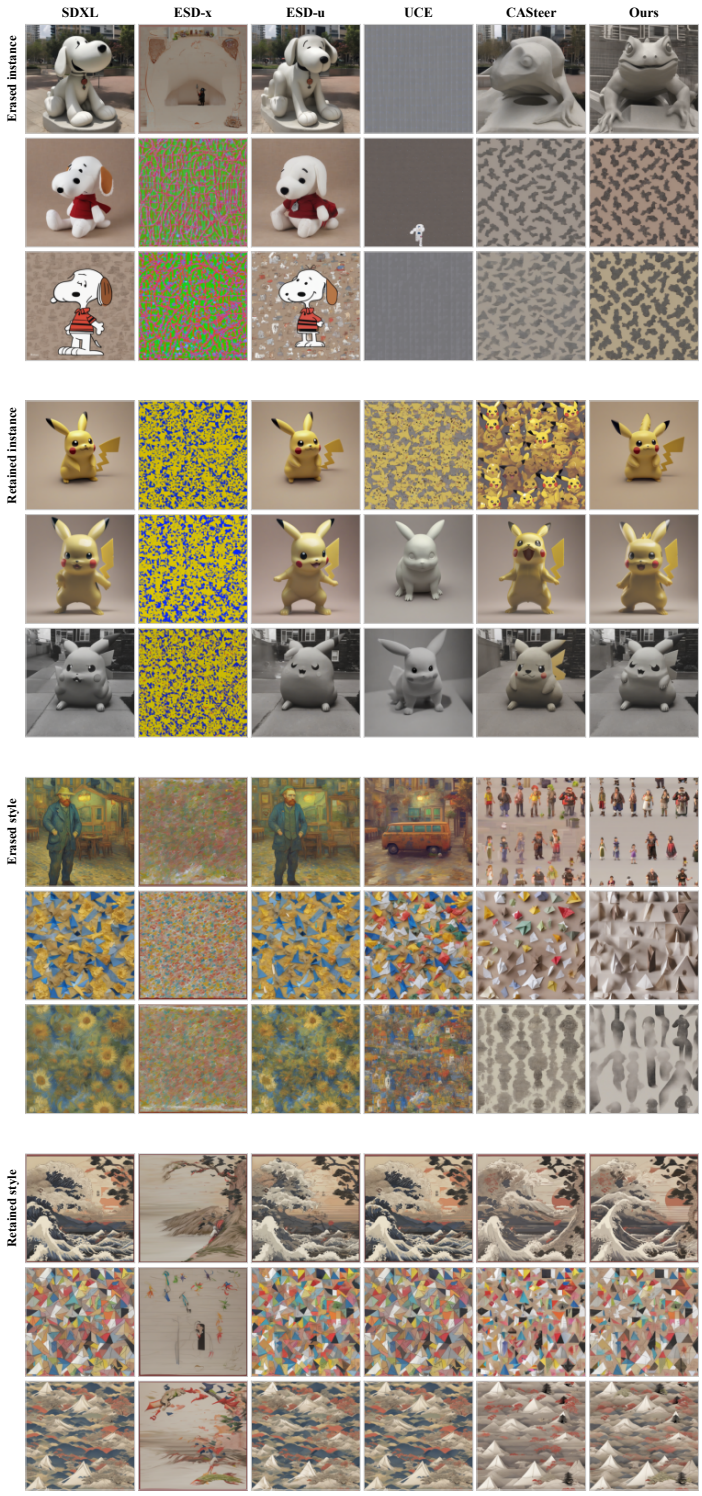


Figure 10: Additional instance and style comparisons on SDXL. Each group of three rows shows representative prompts from the indicated erased or retained set.