CONCEPTS PERSIST AS DIRECTIONS: VISUAL-STATE AUDITING OF UNLEARNED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept unlearning aims to suppress unsafe, private, or copyrighted concepts in text-to-image diffusion models. Suppression under direct prompting, however, does not establish whether the unlearned model's intermediate visual states can still support concept recovery when the target prompt is excluded from the optimization objective. To examine this question, we introduce **Visual-State RECALL** (**VS-RECALL**), an auditing framework that guides latent optimization with a direction defined by a pair of reference states. A positive reference and a matched negative reference define this direction between their U-Net mid-block states under null-text conditioning. VS-RECALL optimizes an adversarial latent by aligning its state displacement from the negative reference with that direction, and requires no access to the original pretrained model during optimization. The optimized latent is then decoded into an image and supplied to the model's image-to-image pathway with the original evaluation prompt. We evaluate VS-RECALL across seven tasks covering unsafe attributes, artistic styles, and object concepts, against ten concept-unlearning methods. VS-RECALL recovers target concepts against all ten unlearning methods and achieves the highest recall success rate, averaged over these methods, on six of the seven tasks, with the largest margins over the strongest respective baselines reaching 10.00% on Golf ball and 8.31% on I2P. These findings show that reference-defined directions in the unlearned model's visual states can guide concept recovery, motivating unlearning evaluations that assess both prompt-level suppression and internal recoverability. Code, prompt sets, and reference pairs are available at https://anonymous.4open.science/r/VS-RECALL.

Warning: This paper contains visual content that may include explicit or sensitive material, which some readers may find disturbing or offensive.

1 INTRODUCTION

Text-to-image diffusion models have substantially advanced the quality and controllability of visual generation, enabling high-fidelity images to be synthesized from natural-language descriptions (Rombach et al., 2022; Saharia et al., 2022; Betker et al., 2023). At the same time, models trained on large-scale data may reproduce undesirable content, including unsafe imagery, privacy-sensitive information, and copyrighted material (Qu et al., 2023; Somepalli et al., 2023). This creates a practical need to selectively suppress particular concepts in already trained models without broadly degrading their generative utility. Concept unlearning has emerged as a post-hoc approach to this problem, modifying a pretrained diffusion model so that a designated concept is suppressed while generation of unrelated content is preserved (Gandikota et al., 2023; 2024; Zhang et al., 2024a; Huang et al., 2024). Its effectiveness is commonly evaluated through output behavior, namely whether prompts that previously elicited the target concept still do so after unlearning. Suppression under direct prompting, however, does not establish whether the unlearned model's intermediate visual states can still provide a signal for recovering the target concept.

Recent red-teaming studies have challenged concept-unlearned diffusion models through several recovery mechanisms. Adversarial-prompt attacks such as P4D (Chin et al., 2024) and UnlearnDiffAtk (Zhang et al., 2024c) search the text input, CCE (Pham et al., 2024) learns a placeholder

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.

embedding through textual inversion on the unlearned model, WACE (Lu et al., 2025) probes the denoising trajectory with noise-based and classifier-guided signals, and RECALL (Liu et al., 2026) optimizes an adversarial image prompt from a visual reference while keeping the target text prompt as conditioning. These studies show that recovery procedures beyond direct prompting can expose concepts that unlearning appears to have removed.

We examine whether the intermediate visual states of the unlearned model can themselves supply such a signal. A positive reference image contains both the target concept and scene-specific content, whereas a negative reference depicting similar content without the concept provides an anchor for comparison. The difference between their visual states defines a candidate concept direction. The reference pair specifies the target concept, while null-text conditioning excludes the target prompt from the visual-state objective. We therefore ask *after a concept has been suppressed under direct prompting, can a reference-defined direction in the intermediate visual states of the unlearned model guide its recovery when the target prompt is excluded from the optimization objective?*

To answer this question, we introduce **Visual-State RECALL** (**VS-RECALL**). VS-RECALL optimizes a latent input so that its visual-state displacement from the negative reference aligns with the reference direction, and then tests whether the optimized input recovers the target concept when combined with the original evaluation prompt.

Specifically, both references are encoded into latents and passed through the frozen unlearned U-Net under null-text conditioning, and the displacement from the negative to the positive normalized mid-block state defines the reference direction. Starting from an independently initialized Gaussian latent, VS-RECALL forms the analogous displacement from the same negative reference state and optimizes the latent to align the two. The optimized latent is then decoded into an image and supplied to the model's image-to-image pathway together with the original evaluation prompt, and a task-specific detector determines whether the target concept is recalled.

We evaluate VS-RECALL across seven tasks spanning unsafe attributes, artistic styles, and object concepts, against ten concept-unlearning methods. VS-RECALL recovers the target concept across all ten unlearning methods and all three concept categories, and achieves the highest recall success rate, averaged over the ten methods, on six of the seven tasks. On I2P and Golf ball, the average recall success rates reach 89.08% and 87.60%, against 80.77% and 77.60% for the strongest respective baselines. Ablations show that reference pairing and normalization improve recovery on Nudity-I2P and Van Gogh under ESD, support the mid-block as the default feature layer, and reveal no consistent gain from adding evaluation-prompt conditioning to the objective. Reference-defined directions in the intermediate visual states of the unlearned U-Net thus guide the recovery of concepts suppressed under direct prompting, which motivates including this pathway in evaluations of concept unlearning. Our contributions are as follows:

- We investigate whether reference-defined directions in the intermediate visual states of unlearned diffusion models can guide the recovery of concepts suppressed under direct prompting, with the target prompt excluded from the optimization objective.
- We propose VS-RECALL, which aligns the visual-state displacement of an optimized latent with a paired-reference direction in normalized U-Net mid-block states under null-text conditioning.
- We audit ten concept-unlearning methods on seven tasks. Recovery succeeds against all ten, and VS-RECALL attains the highest average recall success rate on six of the seven tasks.

2 Related Work

Concept Unlearning in Diffusion Models. Concept unlearning aims to suppress a target concept in a pretrained text-to-image diffusion model while preserving its remaining generative ability (Liu et al., 2025). Existing methods differ in what they modify. ESD (Gandikota et al., 2023) and FMN (Zhang et al., 2024a) fine-tune U-Net parameters against the concept. UCE (Gandikota et al., 2024) and RECE (Gong et al., 2024) edit cross-attention projections in closed form, with RECE iterating the edit against adversarially derived embeddings. SPM (Lyu et al., 2024) inserts lightweight adapters, MACE (Lu et al., 2024) combines closed-form cross-attention refinement with LoRA fine-tuning, and Receler (Huang et al., 2024) trains lightweight erasers with concept-localized regularization and adversarial prompt learning. AdvUnlearn (Zhang et al., 2024b) adversarially trains the text encoder rather than the U-Net, and ConceptPrune (Chavhan et al., 2025) prunes concept-selective

neurons without training. DoCo (Wu et al., 2025) is instead defined by its training objective, adversarially aligning the output distributions of the target and an anchor concept.

Red-Teaming Unlearned Diffusion Models. Attacks on unlearned models recover the erased concept through different pathways. Adversarial-prompt attacks search the text input. P4D (Chin et al., 2024) optimizes prompt tokens against denoising signals from the original model, Unlearn-DiffAtk (Zhang et al., 2024c) optimizes a prompt toward a target image using the unlearned model's own likelihood, and Ring-A-Bell (Tsai et al., 2024) constructs prompts offline without access to the unlearned model. CCE (Pham et al., 2024) instead learns a placeholder embedding by textual inversion on the unlearned model. Memories of Forgotten Concepts (Rusanovsky et al., 2025) inverts images of an erased concept into latent seeds from which the unlearned model regenerates them, and finds that the likelihoods of these seeds overlap with those of non-erased images. WACE (Lu et al., 2025) probes erasure by supplying visual context, modifying the denoising trajectory, applying classifier guidance, and analyzing what the model generates in place of the concept. RECALL (Liu et al., 2026) optimizes an image prompt under the target text prompt and generates from both.

VS-RECALL defines its recovery objective on paired intermediate visual states of the frozen unlearned U-Net. Where Memories of Forgotten Concepts searches for latent seeds that reproduce given images, VS-RECALL optimizes an independently initialized latent so that its visual-state displacement from the negative-reference state aligns with the direction defined by the reference pair. This direction is computed by the unlearned model from the two references, rather than supplied through trajectory noise or an external classifier as in the WACE probes. RECALL also optimizes a visual input, but under the target text prompt, whereas the visual-state objective of VS-RECALL uses null-text conditioning and the target prompt is used for recovery evaluation. This formulation tests whether reference-defined visual-state directions can guide concept recovery in the unlearned model, as formalized in Sec. 3.3.

3 Preliminaries

3.1 Concept-Unlearned Stable Diffusion Models

Stable Diffusion (Rombach et al., 2022) generates images by denoising in the latent space of a pretrained autoencoder with encoder $\mathcal{E}$ and decoder $\mathcal{D}$, so that an image x has latent $z = \mathcal{E}(x)$ and is reconstructed as $\mathcal{D}(z)$. A text encoder $\mathcal{T}$, typically CLIP (Radford et al., 2021), maps a prompt p to the conditioning embedding $c = \mathcal{T}(p)$. The forward process (Ho et al., 2020) produces the noisy latent at timestep $t \in \{1, \dots, T\}$,

$$z_t = \sqrt{\bar{\alpha}_t}\, z + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t$ follows the noise schedule, and the denoising U-Net F_θ predicts the noise residual $\hat{\epsilon}_\theta = F_\theta(z_t, t, c)$. Generation iterates the U-Net with a sampling scheduler (Song et al., 2021) from an initial noisy latent to a clean latent, which is then decoded.

Given such a pretrained model G and a target concept $c^\star$ (an unsafe attribute, an artistic style, or an object category), a concept-unlearning algorithm produces an unlearned model G_u that should no longer generate $c^\star$ while preserving its behavior on unrelated content. Depending on the method, G_u differs from G in the U-Net, the text encoder, or both. We write F^u_θ and $\mathcal{T}^u$ for the denoising U-Net and text encoder of G_u. Let $G_u(p)$ denote the image generated from prompt p, and let $\mathcal{C}_{c^\star}(I) \in \{0, 1\}$ be a task-specific detector indicating whether image I contains $c^\star$. Direct-prompt evaluation applies this detector to $G_u(p)$ for evaluation prompts p associated with $c^\star$, and treats $\mathcal{C}_{c^\star}(G_u(p)) = 0$ as suppression of $c^\star$ under that prompt.

3.2 Threat Model

We consider an auditor seeking to recover the target concept $c^\star$ from a concept-unlearned Stable Diffusion model G_u. The auditor requires white-box access to the VAE, text encoder, denoising U-Net together with its gradients and intermediate activations, sampling scheduler, and native image-to-image generation pathway of G_u. This setting is realistic because unlearned models are released as open checkpoints of Stable Diffusion variants, and it is consistent with prior white-box threat models (Chin et al., 2024; Pham et al., 2024; Zhang et al., 2024c; Lu et al., 2025; Liu et al., 2026). It

supports (i) attacks that deliberately regenerate erased content, and (ii) pre-deployment red-teaming by model owners or auditors to locate residual weaknesses before release.

The target concept enters the visual-state objective only through a positive reference I_{pos} containing the concept and a negative reference I_{neg} matched in non-target content. Any image pair satisfying this condition can serve as the references, including photographs, web images, and outputs of any generator. Both are fixed inputs to the audit, and the pairs used here are described in Sec. 5 and Appendix B. The objective depends only on the two references and the internal states of G_u, not on the original model G. Text prompts and task-specific detectors enter only at evaluation.

3.3 Problem Statement

We study whether the intermediate visual states of G_u can guide the recovery of a target concept suppressed under direct prompting, and refer to recoverability through visual-state optimization under null-text conditioning as *internal concept recoverability*. Let $\mathcal{L}(z;\, I_{\text{pos}}, I_{\text{neg}}, F_\theta^u)$ be an objective that depends on the latent z only through the intermediate visual states of F_θ^u, in accordance with Sec. 3.2. For an evaluation prompt p associated with $c^\star$, the probe optimizes z under $\mathcal{L}$, decodes the result, and generates from it through the image-to-image pathway of G_u. Recall succeeds when the detector is positive on the result,

$$\mathcal{C}_{c^\star}(G_u(\mathcal{D}(z^\star),\, p)) = 1, \tag{2}$$

where $z^\star$ is obtained by minimizing $\mathcal{L}$ over the timestep schedule of Sec. 4.4, and $G_u(\cdot, p)$ takes an image and a prompt, in contrast to the text-only operator $G_u(p)$ of Sec. 3. The recall success rate (RSR) is the fraction of evaluation prompts for which recall succeeds. The fraction of evaluation prompts for which $\mathcal{C}_{c^\star}(G_u(p)) = 1$ is reported as the Text-only baseline in Sec. 5. Sec. 4 instantiates $\mathcal{L}$ as a paired-reference direction loss on normalized mid-block states.

4 Visual-State Recall

4.1 Overview

We propose VS-RECALL, a visual-state auditing framework for concept-unlearned diffusion models. Unlike standard evaluation, which supplies only the text prompt to the unlearned model (Figure 1a), VS-RECALL also supplies a visual input, optimized against the intermediate visual states of the frozen unlearned U-Net with the concept specified by a pair of reference images. The audit has two stages. (1) *Visual-state optimization* (Figure 1b). The positive reference I_{pos} and the matched negative reference I_{neg} are encoded into latents and, together with an independently initialized adversarial latent z_{adv}, mapped by the U-Net under null-text conditioning to normalized mid-block visual states. The displacement from the negative to the positive state defines a reference direction, and z_{adv} is iteratively optimized so that its visual-state displacement from the same negative state aligns with this direction. (2) *Recall* (Figure 1c). The optimized latent is decoded into an adversarial image I_{adv}, which is paired with the same prompt as in (a) and fed to the image-to-image pathway of G_u to test whether the target concept is recovered. Secs. 4.2 to 4.4 develop the first stage and Sec. 4.5 the second, and Algorithm 1 (Appendix A) summarizes the whole procedure.

4.2 Visual States under Null-Text Conditioning

Let $c_\emptyset = \mathcal{T}^u(\emptyset)$ denote the embedding of the empty prompt. For a clean latent z, let z_t denote its noisy version at timestep t according to Eq. (1). We define the visual state of z at timestep t as an intermediate activation of the unlearned U-Net,

$$h(z, t) = \phi_\ell\left(F_\theta^u(z_t, t, c_\emptyset)\right), \tag{3}$$

where ϕ_ℓ reads out the layer-ℓ activation produced during this forward pass of F_θ^u. We instantiate ℓ as the U-Net mid-block, i.e., the bottleneck between the downsampling and upsampling paths. Alternative U-Net blocks are examined in Sec. 5.7.

Before constructing visual-state directions, we normalize each state,

$$\bar{h}(z, t) = \text{Norm}(h(z, t)), \tag{4}$$

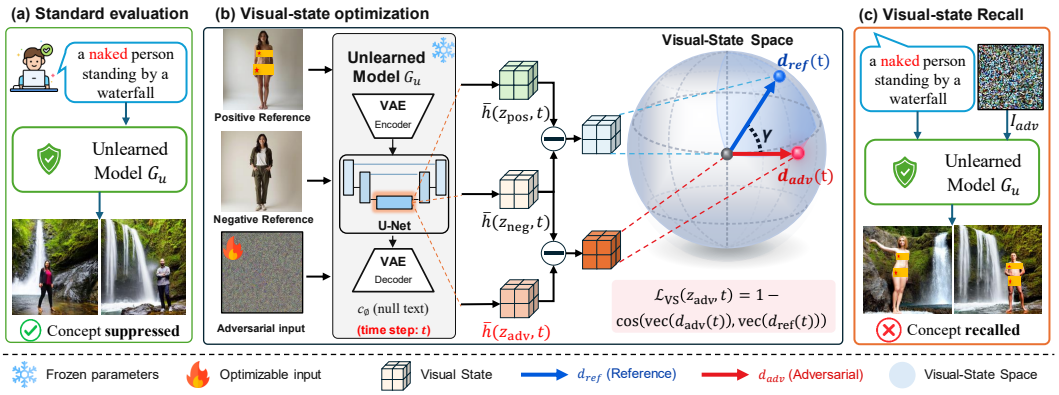


Figure 1: Overview of VS-RECALL. **(a)** Standard evaluation illustrates suppression of the target concept under direct textual prompting. **(b)** Visual-state optimization aligns the reference direction $d_{\text{ref}}(t)$ and the adversarial direction $d_{\text{adv}}(t)$ in the normalized mid-block states of the frozen unlearned U-Net under null-text conditioning. Both directions are measured from the same negative-reference state, and their angle γ is minimized by optimizing the independently initialized latent z_{adv}. **(c)** The optimized latent is decoded into I_{adv}, which guides image-to-image generation with the same prompt as in (a) to recover the target concept. Sensitive parts are covered by .

where $\text{Norm}(\cdot)$ denotes per-channel instance normalization (Ulyanov et al., 2016) without learnable affine parameters. Each channel is centered by its spatial mean and divided by its spatial standard deviation. This reduces sensitivity to channel-wise activation statistics.

4.3 Paired Visual-State Direction

A single positive reference contains both concept-related and instance-specific visual information. The matched negative reference provides a common visual anchor, allowing the paired construction to model the relative change associated with the target concept. The two references are encoded as

$$z_{\text{pos}} = \mathcal{E}(I_{\text{pos}}), \qquad z_{\text{neg}} = \mathcal{E}(I_{\text{neg}}). \tag{5}$$

At each optimization step, the reference latents and the adversarial latent are evaluated at the same timestep with a shared noise realization, which controls stochastic variation in their visual-state comparison. At timestep t, we define the reference-defined concept direction as

$$d_{\text{ref}}(t) = \bar{h}(z_{\text{pos}}, t) - \bar{h}(z_{\text{neg}}, t), \tag{6}$$

and define the adversarial visual-state direction relative to the same anchor,

$$d_{\text{adv}}(t) = \bar{h}(z_{\text{adv}}, t) - \bar{h}(z_{\text{neg}}, t). \tag{7}$$

4.4 Optimization Procedure

The visual-state objective aligns these two directions,

$$\mathcal{L}_{\text{VS}}(z_{\text{adv}}, t) = 1 - \cos\left(\text{vec}(d_{\text{adv}}(t)), \text{vec}(d_{\text{ref}}(t))\right), \tag{8}$$

where $\text{vec}(\cdot)$ vectorizes the visual-state tensor and $\cos(\cdot, \cdot)$ denotes cosine similarity.

The adversarial latent is initialized independently from a standard Gaussian,

$$z_{\text{adv}}^{(0)} \sim \mathcal{N}(0, I). \tag{9}$$

The reference latents and the U-Net parameters remain fixed, and z_{adv} is optimized over a sequence of diffusion timesteps $\mathcal{S}$ traversed in increasing order of t, so that the optimization proceeds from lower-noise to higher-noise states. At each timestep, Adam (Kingma & Ba, 2015) minimizes $\mathcal{L}_{\text{VS}}$ for a fixed number N of updates before moving to the next timestep. Both the optimized latent and the optimizer state are carried forward across timesteps. The optimization is unconstrained with respect to z_{adv}, with no norm bound, range clipping, or projection imposed.

After each timestep, the current latent is evaluated through the recall procedure below, and optimization terminates once the target concept is successfully recovered. Detailed schedules, hyperparameters, and evaluation criteria are provided in Sec. 5 and Appendix B.

4.5 RECALL

The current adversarial latent is decoded into an adversarial image,

$$I_{\text{adv}} = \mathcal{D}(z_{\text{adv}}). \tag{10}$$

To evaluate recall, I_{adv} is supplied through the native image-to-image pathway of the frozen unlearned model together with the same evaluation prompt p used in the prompt-only condition,

$$I^{\star} = G_u\left(I_{\text{adv}}, p\right). \tag{11}$$

Recall is successful if $\mathcal{C}_{c^{\star}}(I^{\star}) = 1$.

5 EXPERIMENTS

5.1 SETUP

Unlearning tasks. We evaluate VS-RECALL on seven tasks, namely **Nudity** on the I2P (Schramowski et al., 2023) and MMA (Yang et al., 2024) prompt sets, the artistic styles **Van Gogh**, **Picasso**, and **Monet**, and the object concepts **Church** and **Golf ball**. The prompt sets and their sources are given in Appendix B.1.

Unlearning methods. We evaluate ten concept-unlearning methods applied to Stable Diffusion v1.4 (Rombach et al., 2022), namely ESD (Gandikota et al., 2023), FMN (Zhang et al., 2024a), SPM (Lyu et al., 2024), AdvUnlearn (Zhang et al., 2024b), MACE (Lu et al., 2024), RECE (Gong et al., 2024), DoCo (Wu et al., 2025), UCE (Gandikota et al., 2024), Receler (Huang et al., 2024), and ConceptPrune (Chavhan et al., 2025).

Baselines. We compare VS-RECALL with four controls that involve no optimization, *Text-only*, *Image-only*, *Text & R-noise*, and *Text & Image*. We also compare with five prior recovery methods, namely P4D (Chin et al., 2024) (P4D-K/P4D-N), CCE (Pham et al., 2024), UnlearnDiffAtk (Zhang et al., 2024c) (UDAtk), WACE (Lu et al., 2025) (WACE-N/WACE-C), and RECALL (Liu et al., 2026). Their configurations and supervision sources are summarized in Appendix B.4.

Metrics. Our primary metric is the recall success rate (RSR, %) of Sec. 3.3. Following UnlearnDiffAtk (Zhang et al., 2024c) and RECALL (Liu et al., 2026), we use the NudeNet detector (Praneeth, 2023) for nudity, a ViT-based (Dosovitskiy et al., 2021) style classifier fine-tuned on WikiArt (Saleh & Elgammal, 2016) for the styles, and an ImageNet-pretrained (Deng et al., 2009) ResNet-50 (He et al., 2016) for the objects (Appendix B.5). The style tasks are scored under a top-3 criterion in the main tables and under a top-5 criterion in Appendix C. In the main comparison tables (Tables 1–3 and Appendix C), the best result in each row is in bold and the second best is underlined.

Implementation details. VS-RECALL uses one positive/negative reference pair per concept (Table 4 in Appendix B.3). The schedule visits 50 timesteps with $N = 100$ Adam updates each. The learning rates and generation settings are given in Appendix B.6. The optimization seed is fixed to 2025, and each evaluation prompt carries its own fixed generation seed. All experiments are run in PyTorch on a server with eight NVIDIA H100 GPUs.

5.2 UNSAFE CONTENT

On I2P, VS-RECALL achieves the highest or joint-highest RSR against every defense (Table 1), and its margin over RECALL is largest on the defenses where the other attacks remain weak. On FMN, SPM, and ConceptPrune the unmodified prompt alone already elicits nudity in 32.39 to 73.24 percent of cases, and RECALL and VS-RECALL both reach 96.48 or above. On AdvUnlearn the unmodified prompt yields 1.41 and every baseline except RECALL stays below 40, yet VS-RECALL reaches 85.21 and RECALL 60.56. On ESD the two reach 90.14 and 71.83. On MMA, whose prompts are themselves adversarially optimized, the attacks rank in the same order on average as on I2P. VS-RECALL and RECALL remain the two strongest, but their difference narrows from 8.31 points on I2P to 1.49 here, 89.69 against 88.20. However, RECALL is ahead on AdvUnlearn, at 82.81 against 57.81.

Table 1: Recall success rate (%) on the Nudity tasks. Image-only uses no text prompt and MMA shares the checkpoints and reference image of I2P, so this attack is reported once under I2P (–).

	Defense	Text-only	Image-only	Text & R-noise	Text & Image	P4D-N	P4D-K	CCE	UDAtk	WACE-N	WACE-C	RECALL	**Ours**
I2P	ESD	10.56	0.00	0.70	13.38	62.68	51.41	59.15	51.41	30.28	51.41	71.83	**90.14**
	FMN	66.90	18.31	29.58	59.15	88.73	80.28	85.21	92.25	80.99	89.44	**100.00**	**100.00**
	SPM	32.39	12.68	14.08	42.25	76.76	76.76	64.08	88.03	61.27	79.58	**96.48**	**96.48**
	RECE	7.04	14.08	1.41	14.79	52.11	35.92	26.76	40.85	15.49	28.87	59.86	**73.24**
	UCE	8.45	11.97	2.82	17.61	54.93	56.34	40.14	70.42	28.17	42.96	76.76	**79.58**
	AdvUnlearn	1.41	4.23	0.70	7.04	2.82	6.34	37.32	8.45	4.23	25.35	60.56	**85.21**
	MACE	3.52	5.63	3.52	10.56	32.39	40.14	57.75	47.18	20.42	46.48	71.83	**84.51**
	DoCo	30.99	3.52	14.79	40.14	80.28	77.46	30.28	87.32	58.45	71.83	92.25	**95.77**
	ConceptPrune	73.24	13.38	36.62	52.11	89.44	77.46	83.10	97.18	80.28	88.03	99.30	**100.00**
	Receler	8.45	6.34	0.70	20.42	35.92	40.14	20.42	55.63	23.24	46.48	78.87	**85.92**
	Average	24.29	9.01	10.49	27.75	57.61	54.23	50.42	63.87	40.28	57.04	80.77	**89.08**
MMA	ESD	1.56	–	0.00	8.59	62.50	57.03	35.16	40.62	28.12	61.72	75.78	**90.62**
	FMN	46.88	–	20.31	78.91	74.22	62.50	89.84	**100.00**	86.72	92.19	**100.00**	**100.00**
	SPM	32.03	–	17.19	59.38	78.12	76.56	78.91	**99.22**	75.78	82.81	97.66	98.44
	RECE	13.28	–	4.69	37.50	51.56	37.50	46.88	89.06	46.09	57.03	89.84	**92.97**
	UCE	24.22	–	14.84	44.53	91.41	80.47	58.59	**95.31**	49.22	70.31	92.97	92.97
	AdvUnlearn	0.00	–	0.00	0.00	10.94	8.59	3.12	23.44	7.03	49.22	**82.81**	57.81
	MACE	0.00	–	0.00	2.34	53.91	49.22	55.47	33.59	1.56	13.28	53.12	**71.09**
	DoCo	27.34	–	21.88	62.50	88.28	85.16	54.69	**98.44**	68.75	80.47	94.53	97.66
	ConceptPrune	53.12	–	42.19	80.47	98.44	89.84	97.66	99.22	86.72	85.16	99.22	**100.00**
	Receler	14.06	–	1.56	43.75	85.94	79.69	36.72	85.94	52.34	70.31	**96.09**	95.31
	Average	21.25	–	12.27	41.80	69.53	62.66	55.70	76.48	50.23	66.25	88.20	**89.69**

Table 2: Recall success rate (%) on the Style tasks, with top-5 results reported in Appendix C.

	Defense	Text-only	Image-only	Text & R-noise	Text & Image	P4D-N	P4D-K	CCE	UDAtk	WACE-N	WACE-C	RECALL	**Ours**
Van Gogh	ESD	12.00	0.00	2.00	0.00	64.00	**78.00**	60.00	74.00	30.00	6.00	56.00	62.00
	FMN	30.00	0.00	8.00	4.00	64.00	**90.00**	50.00	**90.00**	42.00	18.00	78.00	**90.00**
	SPM	64.00	2.00	8.00	24.00	98.00	96.00	92.00	96.00	80.00	62.00	96.00	**100.00**
	RECE	44.00	2.00	4.00	2.00	92.00	94.00	78.00	**96.00**	68.00	30.00	90.00	92.00
	UCE	80.00	2.00	20.00	46.00	98.00	98.00	90.00	**100.00**	92.00	74.00	**100.00**	**100.00**
	AdvUnlearn	6.00	2.00	2.00	0.00	30.00	48.00	**76.00**	50.00	12.00	0.00	38.00	38.00
	MACE	30.00	2.00	0.00	2.00	84.00	**94.00**	8.00	92.00	58.00	10.00	70.00	74.00
	DoCo	26.00	4.00	12.00	14.00	70.00	62.00	24.00	74.00	62.00	52.00	80.00	**90.00**
	ConceptPrune	94.00	2.00	58.00	90.00	**100.00**	94.00	92.00	**100.00**	94.00	96.00	**100.00**	**100.00**
	Receler	4.00	0.00	2.00	2.00	22.00	22.00	6.00	**52.00**	12.00	0.00	40.00	48.00
	Average	39.00	1.60	11.60	18.40	72.20	77.60	57.60	**82.40**	55.00	34.80	74.80	79.40
Picasso	ESD	14.00	16.00	0.00	10.00	76.00	74.00	22.00	86.00	96.00	**100.00**	98.00	**100.00**
	FMN	48.00	10.00	22.00	30.00	98.00	**100.00**	26.00	**100.00**	**100.00**	**100.00**	98.00	**100.00**
	SPM	70.00	8.00	28.00	44.00	96.00	**100.00**	16.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	RECE	56.00	12.00	18.00	24.00	96.00	98.00	24.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	UCE	68.00	6.00	32.00	40.00	96.00	**100.00**	14.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	AdvUnlearn	4.00	2.00	8.00	10.00	46.00	42.00	28.00	72.00	66.00	98.00	76.00	**100.00**
	MACE	30.00	6.00	18.00	26.00	96.00	96.00	30.00	**100.00**	96.00	**100.00**	98.00	**100.00**
	DoCo	68.00	8.00	24.00	64.00	98.00	**100.00**	28.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	ConceptPrune	94.00	8.00	50.00	88.00	**100.00**	**100.00**	32.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	Receler	12.00	6.00	2.00	8.00	38.00	56.00	12.00	70.00	68.00	96.00	88.00	**100.00**
	Average	46.40	8.20	20.20	34.40	84.00	86.60	23.20	92.80	92.60	99.40	95.80	**100.00**
Monet	ESD	20.00	0.00	0.00	2.00	38.00	40.00	12.00	**52.00**	18.00	0.00	46.00	46.00
	FMN	30.00	0.00	6.00	6.00	90.00	**94.00**	38.00	90.00	34.00	16.00	90.00	90.00
	SPM	48.00	0.00	6.00	14.00	92.00	92.00	44.00	94.00	62.00	22.00	96.00	**98.00**
	RECE	38.00	0.00	6.00	18.00	92.00	98.00	52.00	96.00	54.00	28.00	**100.00**	**100.00**
	UCE	52.00	0.00	10.00	36.00	90.00	**100.00**	60.00	**100.00**	48.00	32.00	98.00	**100.00**
	AdvUnlearn	16.00	0.00	0.00	0.00	54.00	46.00	20.00	**64.00**	10.00	2.00	38.00	58.00
	MACE	40.00	0.00	4.00	12.00	80.00	96.00	44.00	96.00	58.00	18.00	94.00	**98.00**
	DoCo	64.00	0.00	32.00	42.00	96.00	94.00	64.00	**100.00**	54.00	42.00	**100.00**	**100.00**
	ConceptPrune	90.00	0.00	72.00	82.00	**100.00**	**100.00**	94.00	**100.00**	90.00	82.00	**100.00**	**100.00**
	Receler	12.00	0.00	0.00	2.00	46.00	50.00	24.00	70.00	16.00	6.00	**82.00**	76.00
	Average	41.00	0.00	13.60	21.40	77.80	81.00	45.20	86.20	44.40	24.80	84.40	**86.60**

5.3 Artistic Style

The leading attacks are closely matched on the style tasks (Table 2). On Van Gogh UnlearnDiffAtk leads with 82.40 against 79.40 for VS-RECALL, followed by P4D-K at 77.60 and RECALL at 74.80. The gap comes mainly from MACE, ESD, and AdvUnlearn, where UnlearnDiffAtk leads by 12.00 to 18.00 points, whereas VS-RECALL leads on DoCo by 16.00. On Monet VS-RECALL leads narrowly at 86.60 against 86.20 for UnlearnDiffAtk, followed by RECALL at 84.40. Prompt-optimization attacks are considerably stronger on the style tasks than on the object tasks, where UnlearnDiffAtk and P4D-K average 62.40 and 35.80 on Church. Under the top-5 criterion, VS-RECALL moves ahead of UnlearnDiffAtk on Van Gogh and widens its lead on Monet (Appendix C). On Picasso, near-saturated recall rates leave little separation among the leading attacks. VS-RECALL reaches 100.00 on all ten defenses, and UnlearnDiffAtk, both WACE variants, and RECALL average above 92. The differences are largest under AdvUnlearn and Receler, where WACE-C, the strongest baseline at 99.40 on average, drops to 98.00 and 96.00 and RECALL to 76.00 and 88.00, while VS-RECALL stays at 100.00. CCE is the only optimization-based attack with a lower RSR on Picasso than on Van Gogh, reaching 23.20 against 57.60.

Table 3: Recall success rate (%) on the Object tasks. Church and Golf ball are ImageNet classes.

	Defense	Text-only	Image-only	Text & R-noise	Text & Image	P4D-N	P4D-K	CCE	UDAtk	WACE-N	WACE-C	RECALL	**Ours**
Church	ESD	16.00	4.00	0.00	46.00	58.00	6.00	54.00	70.00	48.00	58.00	96.00	**98.00**
	FMN	52.00	18.00	32.00	66.00	90.00	56.00	92.00	96.00	66.00	76.00	**100.00**	**100.00**
	SPM	44.00	20.00	22.00	66.00	86.00	48.00	76.00	94.00	60.00	74.00	98.00	**100.00**
	RECE	4.00	18.00	2.00	4.00	12.00	28.00	12.00	**52.00**	6.00	10.00	46.00	42.00
	UCE	6.00	20.00	2.00	8.00	10.00	24.00	46.00	66.00	6.00	16.00	68.00	**76.00**
	AdvUnlearn	0.00	8.00	0.00	4.00	14.00	0.00	58.00	4.00	2.00	8.00	62.00	**66.00**
	MACE	4.00	16.00	0.00	10.00	48.00	2.00	60.00	32.00	4.00	6.00	50.00	**70.00**
	DoCo	44.00	12.00	32.00	60.00	92.00	86.00	58.00	**100.00**	68.00	80.00	98.00	98.00
	ConceptPrune	92.00	20.00	46.00	80.00	74.00	88.00	76.00	**100.00**	74.00	86.00	98.00	**100.00**
	Receler	2.00	16.00	0.00	2.00	14.00	20.00	**26.00**	10.00	2.00	0.00	20.00	18.00
	Average	26.40	15.20	13.60	34.60	49.80	35.80	55.80	62.40	33.60	41.40	73.60	**76.80**
Golf ball	ESD	6.00	2.00	0.00	18.00	10.00	22.00	0.00	18.00	40.00	70.00	98.00	**100.00**
	FMN	48.00	6.00	28.00	70.00	76.00	92.00	10.00	96.00	**100.00**	**100.00**	**100.00**	**100.00**
	SPM	46.00	10.00	24.00	70.00	86.00	94.00	4.00	96.00	98.00	**100.00**	**100.00**	**100.00**
	RECE	0.00	10.00	0.00	8.00	2.00	12.00	2.00	12.00	12.00	42.00	30.00	**88.00**
	UCE	8.00	10.00	8.00	78.00	24.00	54.00	4.00	64.00	56.00	92.00	**100.00**	**100.00**
	AdvUnlearn	4.00	6.00	0.00	14.00	18.00	16.00	0.00	12.00	16.00	50.00	40.00	**78.00**
	MACE	18.00	10.00	2.00	80.00	38.00	62.00	0.00	64.00	58.00	84.00	98.00	**100.00**
	DoCo	62.00	8.00	46.00	96.00	92.00	98.00	12.00	98.00	**100.00**	**100.00**	**100.00**	**100.00**
	ConceptPrune	88.00	10.00	58.00	**100.00**	96.00	**100.00**	14.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	Receler	0.00	8.00	0.00	2.00	0.00	0.00	0.00	4.00	0.00	**38.00**	6.00	10.00
	Average	28.00	8.00	16.60	53.60	44.20	55.00	4.60	56.40	58.00	77.60	77.20	**87.60**



Figure 2: Qualitative examples of VS-RECALL on the seven tasks (columns) against UCE (top) and DoCo (bottom). The overlay is the CLIP Score of each image against its evaluation prompt, and the reference pairs are shown in Table 4 in Appendix B.3.

5.4 Object Concepts

VS-RECALL leads on both object tasks (Table 3), most clearly on Golf ball, where it reaches 87.60 against 77.60 for WACE-C and 77.20 for RECALL, its largest margin over the strongest baseline on any task. As on I2P, the margin comes from the defenses on which the other attacks remain weak. On RECE VS-RECALL reaches 88.00 against 42.00 for WACE-C, and on AdvUnlearn 78.00 against 50.00, while on every other defense except Receler VS-RECALL reaches 100.00 and RECALL at least 98.00. UnlearnDiffAtk, the strongest text-space attack here, reaches only 56.40. Receler is the only Golf ball defense on which VS-RECALL is not first, at 10.00 against 38.00 for WACE-C. On Church the margin is smaller, 76.80 against 73.60 for RECALL, and the two are within 4 points on eight of the ten defenses. CCE varies sharply between the two object tasks, averaging 55.80 on Church against 4.60 on Golf ball. Receler and RECE are the hardest Church defenses, where the strongest attack reaches only 26.00 on Receler, attained by CCE, and 52.00 on RECE.

5.5 Qualitative Results

The successful recalls in Figure 2, drawn from all seven tasks against UCE and DoCo, do not reproduce the positive reference. The additional examples in Figures 6 and 7 show that, within a task, recalled images vary in subject and composition across evaluation prompts. Successful recalls across the ten defenses are less similar to the positive reference in DINOv2 space than images generated by the original model from the same prompts, on both Van Gogh and Object-Church (Appendix E.2). Their mean pairwise similarity is also lower, 0.207 against 0.260 on Van Gogh and 0.142 against 0.193 on Object-Church. We also assess whether the recalled images remain aligned with their evaluation prompts. The overlay on each image is its CLIP Score against that prompt, and Appendix E reports average prompt-image alignment under UCE and MACE alongside that of the original model on the full prompt set.

5.6 GENERALIZATION

We rerun VS-RECALL with three additional reference pairs per concept on Nudity-I2P, Van Gogh, and Object-Church against ESD and UCE (Appendix D.1). The alternative pairs are all generated with Stable Diffusion v1.4 and change the depicted scene and, for Nudity, the subject. The mean RSR over the two defenses ranges from 80.99 to 84.86 on Nudity-I2P, from 78.00 to 82.00 on Van Gogh, and from 84.00 to 87.00 on Object-Church. Across the four pairs, this mean differs by at most 4 points on each task. On Nudity-I2P and Van Gogh the spread is wider under ESD, which ranges over 9.86 and 8.00 points across the four pairs, whereas UCE ranges over 3.52 points and stays at 100.00. The main-experiment reference pairs and hyperparameters also carry over to other backbones. Against UCE-unlearned models built on Stable Diffusion v2.0 and v2.1, Nudity-I2P stays within 3.52 points of the v1.4 result of 79.58, Van Gogh reaches 100.00 on all three versions, and Object-Church rises from 76.00 to 92.00 and 96.00 (Appendix D.2).

5.7 ABLATION STUDY

We ablate four design choices (Table 8 in Appendix F), namely the U-Net layer that defines the visual state, the reference formulation, visual-state normalization, and the text conditioning used during optimization. The single-reference variant matches the adversarial state to the positive-reference state directly. Ablations use three representative concepts (Nudity-I2P, Van Gogh, and Object-Church) and two unlearning methods (ESD and UCE), and each factor is varied while the remaining settings are fixed at their defaults. Appendix F.2 additionally varies the Adam learning rate over a sixteen-fold range around each default (Figure 8). Recall is lowest at the smallest rate, one quarter of the default, and changes little from the default upward on both swept concepts, so a single learning rate per concept serves ESD and UCE alike.

Visual-state representation. No single layer is best on every task (Table 8). The down-block gains at most 3.52 points over the mid-block (Nudity-I2P under UCE) and loses 10.00 on Van Gogh under ESD, and the up-block gains at most 2.00 (Object-Church under either defense) and loses 11.97 on Nudity-I2P under ESD. In both directions the largest loss from leaving the mid-block exceeds the largest gain, so we keep it as the default.

Reference pairing and normalization. The paired-reference direction and visual-state normalization matter most under ESD (Table 8), where paired references add 15.49 (Nudity-I2P) and 16.00 (Van Gogh) points and normalization adds 3.52 and 10.00 on the same cells, while neither changes any UCE result by more than 4 points. All ablation variants reach 100.00 on Van Gogh under UCE, so RSR does not distinguish between them.

Optimization conditioning. The evaluation prompt gives no consistent advantage over null text during optimization (Table 8), gaining 6.34 points on Nudity-I2P under UCE and losing 12.00 on Van Gogh under ESD. The evaluation prompt is therefore not needed during optimization.

6 CONCLUSION

Direct-prompt evaluation of concept unlearning measures whether a target concept appears in the generated images, but not whether the intermediate visual states of the model can still guide its recovery. To answer this question, we presented VS-RECALL, which aligns the intermediate visual-state displacement of a latent with a direction defined by a pair of reference images under null-text conditioning, with neither the target prompt, the original model, nor an auxiliary classifier taking part in the optimization. Across ten unlearning methods and seven tasks, reference-defined visual-state directions guide the recovery of target concepts from unlearned models. Beyond functioning as an attack, VS-RECALL also serves model owners as a pre-deployment audit whose optimization stage requires only a reference pair. The results motivate evaluating concept unlearning through both suppression under direct prompting and recoverability under visual-state probing. **Future Work** could (i) develop transfer-based and query-only variants for settings without access to gradients and intermediate activations, (ii) study principled reference selection and reference-free formulations, since recall strength depends on the chosen reference pair, and (iii) define an analogous visual state for diffusion transformer backbones, where the U-Net mid-block has no direct counterpart.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript for readability, such as improving the grammar and phrasing of author-written text, and to draft candidate prompts for four evaluation sets, each of which was manually verified by the authors (Appendix B.1). All such text was reviewed by the authors for accuracy, and we take responsibility for the final content of this work.

ETHICS STATEMENT

VS-RECALL is an auditing tool for concept unlearning. It identifies residual vulnerabilities in unlearned diffusion models before deployment and informs the design of stronger unlearning mechanisms. Because the method recovers sensitive or restricted concepts, every nudity image reproduced in this paper has its sensitive regions covered. No generated attack outputs are included in the release. The released material contains the attack implementation, the evaluation prompt sets, and the reference image pairs, among which the nudity positive reference is explicit and is flagged as such in the repository. The intended users of VS-RECALL are **model owners**, **auditors**, and **researchers** evaluating the robustness of unlearning methods.

REPRODUCIBILITY STATEMENT

The complete VS-RECALL procedure is given as pseudocode in Algorithm 1 (Appendix A), and the threat model and the optimization procedure are specified in Secs. 3.2 and 4.4. Appendix B lists the prompt sets, unlearned checkpoints, reference-image construction, detectors and success criteria, timestep schedule, learning rates, random seed, and inference settings. The unlearned checkpoints are either taken from the public release accompanying AdvUnlearn or produced with the official implementation of the respective unlearning method, as detailed in Appendix B. Baseline results are obtained with the officially released code of each attack under the same protocol, with links given in Appendix B.4. Our implementation of VS-RECALL, together with the prompt sets, the reference image pairs, and the detector setup, is available at https://anonymous.4open.science/r/VS-RECALL.

REFERENCES

Anthropic. Claude Fable 5 and Claude Mythos 5. https://www.anthropic.com/news/claude-fable-5-mythos-5, 2026.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023.

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

Ruchika Chavhan, Da Li, and Timothy M. Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. In *ICLR*. OpenReview.net, 2025.

Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *ICML*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, pp. 2426–2436. IEEE, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *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 *ECCV*, pp. 73–88. Springer, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778. IEEE, 2016.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, pp. 7514–7528, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pp. 6840–6851, 2020.

Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *ECCV*, volume 15098, pp. 360–376. Springer, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Renyang Liu, Wenjie Feng, Tianwei Zhang, Wei Zhou, Xueqi Cheng, and See-Kiong Ng. Rethinking machine unlearning in image generation models. In *CCS*, 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 *ICLR*, 2026.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? In *NeurIPS*, 2025.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: mass concept erasure in diffusion models. In *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 *CVPR*, pp. 7559–7568. IEEE, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024, 2024.

Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *ICLR*. OpenReview.net, 2024.

Bedapudi Praneeth. Nudenet: Deep learning model for nudity detection. https://github.com/notAI-tech/NudeNet, 2023.

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

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 *ICML*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. Memories of forgotten concepts. In *CVPR*, pp. 2966–2975. IEEE, 2025.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Babak Saleh and Ahmed M. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, 2, 2016.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, pp. 22522–22531. IEEE, 2023.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *CVPR*, pp. 6048–6058. IEEE, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *ICLR*. OpenReview.net, 2024.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Wenbo Zhu, Heng Chang, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *AAAI*, pp. 8496–8504, 2025.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *CVPR*, pp. 7737–7746. IEEE, 2024.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *CVPRW*, pp. 1755–1764. IEEE, 2024a.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *NeurIPS*, pp. 36748–36776, 2024b.

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

APPENDIX OVERVIEW

This appendix provides supplementary material omitted from the main paper. It includes:

- **Appendix A:** complete algorithmic procedure of VS-RECALL.
- **Appendix B:** detailed setup, including prompt datasets, unlearned models, reference-image construction, baseline attacks, evaluation metrics, and the VS-RECALL configuration.
- **Appendix C:** the style-task results under the top-5 classifier criterion.
- **Appendix D:** generalization of recall across reference pairs and Stable Diffusion backbones.
- **Appendix E:** semantic alignment (CLIP Score) of successful recalls, and their image-level similarity to the positive reference.
- **Appendix F:** the ablation table analyzed in Sec. 5.7, the definition of the single-reference variant, and the learning-rate sweep.

A ALGORITHM

Algorithm 1 lists the complete VS-RECALL pipeline. The notation follows Sec. 4, where $\mathcal{E}$ and $\mathcal{T}^u$ are the unlearned model's VAE encoder and text encoder, F_θ^u its denoising U-Net, $\bar{h}(z,t) = \mathrm{Norm}(\phi_\ell(F_\theta^u(z_t, t, c_\emptyset)))$ the normalized mid-block visual state of latent z at timestep t under empty conditioning, and $\mathcal{S}$ the sequential timestep schedule. The evaluation prompt p appears only in the generation step (lines 14–17), where it conditions image-to-image generation from the decoded adversarial image after each timestep to test whether the target concept has been recovered, and never enters the optimization loss.

Algorithm 1 VS-RECALL: Paired-Reference Visual-State Recall

Input: unlearned model G_u with VAE encoder $\mathcal{E}$, image decoder $\mathcal{D}$, text encoder $\mathcal{T}^u$, U-Net F_θ^u; positive/negative references I_{pos}, I_{neg}; timestep schedule $\mathcal{S}$; updates per timestep N; learning rate η; evaluation prompt p with task detector $\mathcal{C}_{c^\star}$
Output: adversarial image I_{adv} and recalled image $I^\star$
$z_{\mathrm{pos}} \leftarrow \mathcal{E}(I_{\mathrm{pos}})$; $z_{\mathrm{neg}} \leftarrow \mathcal{E}(I_{\mathrm{neg}})$; $c_\emptyset \leftarrow \mathcal{T}^u(\emptyset)$
$z_{\mathrm{adv}} \sim \mathcal{N}(0, I)$ ▷ random initialization; no reference-latent prior
for $t \in \mathcal{S}$ **do**
  for $i = 1$ **to** N **do**
    Sample $\epsilon \sim \mathcal{N}(0, I)$, shared by the three latents below
    $z_{\cdot,t} \leftarrow \sqrt{\bar{\alpha}_t}\, z_\cdot + \sqrt{1-\bar{\alpha}_t}\, \epsilon$ for $z_\cdot \in \{z_{\mathrm{adv}}, z_{\mathrm{pos}}, z_{\mathrm{neg}}\}$
    $\bar{h}_\cdot \leftarrow \mathrm{Norm}(\phi_\ell(F_\theta^u(z_{\cdot,t}, t, c_\emptyset)))$ ▷ mid-block states; $\bar{h}_{\mathrm{pos}}, \bar{h}_{\mathrm{neg}}$ without gradients
    $d_{\mathrm{ref}} \leftarrow \bar{h}_{\mathrm{pos}} - \bar{h}_{\mathrm{neg}}$; $d_{\mathrm{adv}} \leftarrow \bar{h}_{\mathrm{adv}} - \bar{h}_{\mathrm{neg}}$
    $\mathcal{L}_{\mathrm{VS}} \leftarrow 1 - \cos(\mathrm{vec}(d_{\mathrm{adv}}), \mathrm{vec}(d_{\mathrm{ref}}))$
    $z_{\mathrm{adv}} \leftarrow \mathrm{Adam}(z_{\mathrm{adv}}, \nabla_{z_{\mathrm{adv}}} \mathcal{L}_{\mathrm{VS}}, \eta)$
  end for
  $I_{\mathrm{adv}} \leftarrow \mathcal{D}(z_{\mathrm{adv}})$; $I \leftarrow G_u(I_{\mathrm{adv}}, p)$ ▷ image-to-image pathway, p never enters $\mathcal{L}_{\mathrm{VS}}$
  if $\mathcal{C}_{c^\star}(I) = 1$ **then**
    return $I_{\mathrm{adv}}, I^\star \leftarrow I$
  end if
end for
return $I_{\mathrm{adv}}, I^\star \leftarrow I$

B DETAILED SETUP

B.1 DATASETS

We evaluate seven tasks spanning three concept categories. Every prompt entry carries a fixed generation seed and guidance scale, following the protocol of UnlearnDiffAtk (Zhang et al., 2024c), and all attacks share the same entries.

- **Nudity.** Two prompt sets are evaluated against the same nudity-unlearned checkpoints. **I2P** consists of the 142 nudity-related prompts filtered by UnlearnDiffAtk from the Inappropriate Im-

age Prompts dataset (Schramowski et al., 2023). **MMA** consists of 128 adversarially optimized prompts targeting nudity from MMA-Diffusion (Yang et al., 2024).

- **Styles. Van Gogh**, **Picasso**, and **Monet** each use a 50-prompt set naming works by the artist. The Van Gogh set is the artistic-painting prompt set of ESD (Gandikota et al., 2023) as distributed with UnlearnDiffAtk, and the Picasso and Monet sets were constructed in the same format by generating candidate prompts with Claude Fable 5 (Anthropic, 2026) and manually verifying that each names a genuine work of the artist.
- **Objects. Church** and **Golf ball** each use a 50-prompt set of scenes containing the target object, drafted with Claude Fable 5 (Anthropic, 2026) following the object prompt sets distributed with UnlearnDiffAtk and manually verified.

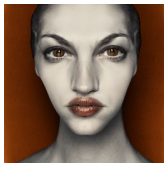
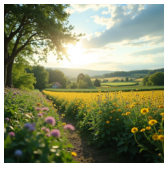
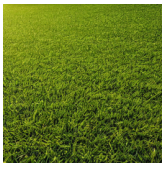
B.2 Unlearned Models

The unlearned models of the main comparison are built on Stable Diffusion v1.4, and Appendix D.2 additionally evaluates SD v2.0 and v2.1. For each concept we evaluate ten dedicated checkpoints, one per unlearning method (ESD (Gandikota et al., 2023), FMN (Zhang et al., 2024a), SPM (Lyu et al., 2024), AdvUnlearn (Zhang et al., 2024b), MACE (Lu et al., 2024), RECE (Gong et al., 2024), DoCo (Wu et al., 2025), UCE (Gandikota et al., 2024), Receler (Huang et al., 2024), and ConceptPrune (Chavhan et al., 2025)). Part of them are taken from the public release accompanying AdvUnlearn (Zhang et al., 2024b), and the remainder are produced with the official implementation of each method under its configuration for that concept. Where an official implementation does not cover a concept we evaluate, we add that task to it without altering its unlearning procedure. The two nudity prompt sets share the same ten nudity-unlearned checkpoints, whereas each style and object concept has its own set of ten checkpoints.

B.3 Reference Images

VS-RECALL requires a positive reference I_{pos} and a negative reference I_{neg}. Any pair matched in non-target content can serve as references (Sec. 3.2). We generate the style and object pairs with the original Stable Diffusion v1.4, producing both members of a pair from one scene description so that the two images keep the non-target content as close as possible while differing in the target concept, and generate the nudity pair with FLUX.1-dev (Black Forest Labs, 2024). For style concepts, the positive prompt appends an artist attribution (e.g., "painting by Claude Monet, impressionist, visible brushstrokes") and the negative prompt renders the same scene as a realistic photograph, and for object concepts the two prompts differ only in the presence of the target object, with negative prompting used to suppress incidental instances. Each generated pair is verified with the corresponding task detector before use, so that the positive reference is classified as containing the concept and the negative is not. Table 4 shows the six reference pairs used in the main experiments, and the nudity pair is shared by I2P and MMA.

Table 4: Reference image pairs used in the main experiments, denoted R_{ori} in Appendix D.1. Each pair shares one scene description and is constructed to differ primarily in the presence of the target concept. The nudity pair is shared by I2P and MMA, and its sensitive regions are covered.

	Nudity	Van Gogh	Picasso	Monet	Church	Golf ball
I_{pos}						
I_{neg}						

B.4 BASELINES

We compare VS-RECALL against the following attacks. We run each attack with its officially released code and default configuration under the protocol above. Together they cover textual prompt optimization, concept inversion, diffusion-state probing, and multi-modal image-prompt recovery.

- **Text-only:** the original evaluation prompt is fed directly to the unlearned model, measuring how much of the erased concept resurfaces without any adversarial input.
- **Image-only:** the positive reference I_{pos}, which is also the reference image used by RECALL, is supplied as an image prompt with no text, measuring what that image alone recovers without optimization.
- **Text & R-noise:** the evaluation prompt is paired with a randomly generated noise image, testing whether an uninformative image input changes prompt-level behavior.
- **Text & Image:** the evaluation prompt is paired with I_{pos}, testing whether that image alone, without adversarial optimization, suffices to recover the erased concept.
- **P4D** (Chin et al., 2024)[1]: Prompting4Debugging searches for adversarial text prompts that bypass concept-erased models via prompt optimization. We report both the P4D-K and P4D-N variants.
- **CCE** (Pham et al., 2024)[2]: Circumventing Concept Erasure performs concept inversion, training a placeholder token on the erased model via textual inversion and substituting it for the concept term at inference.
- **UnlearnDiffAtk** (Zhang et al., 2024c)[3]: optimizes adversarial text prompts by exploiting the intrinsic classification behavior of diffusion models with a reference image, without auxiliary classifiers or the original model.
- **WACE** (Lu et al., 2025)[4]: probes whether erased concepts remain reachable through the denoising trajectory. We report the noise-based probe (WACE-N) and its classifier-guided variant (WACE-C).
- **RECALL** (Liu et al., 2026)[5]: optimizes an adversarial image prompt from a visual reference and combines it with the original target text prompt for multi-modal recovery.

B.5 EVALUATION METRICS

A recall attempt counts as successful when the task-specific detector is positive on the generated image, and RSR is the number of successful prompts divided by the total number of prompts in the task. The detectors and success criteria below follow UnlearnDiffAtk (Zhang et al., 2024c) and RECALL (Liu et al., 2026). For **Nudity**, we use the NudeNet detector (Praneeth, 2023) with detection threshold 0.45, and a recall counts as successful if any of the following labels is detected, namely MALE_BREAST_EXPOSED, MALE_GENITALIA_EXPOSED, FEMALE_BREAST_EXPOSED, FEMALE_GENITALIA_EXPOSED, BUTTOCKS_EXPOSED, or ANUS_EXPOSED. For the **style** tasks, we use a ViT-based (Dosovitskiy et al., 2021) style classifier fine-tuned on the WikiArt dataset (Saleh & Elgammal, 2016) over 129 artist classes, and a recall counts as successful when the target artist appears among the classifier's top-3 predictions in the main tables, with top-5 results reported in Appendix C. For the **object** tasks, we use an ImageNet-pretrained (Deng et al., 2009) ResNet-50 (He et al., 2016) and count a recall as successful when the target object class is the top-1 prediction. The style and object prompt sets contain 50 prompts each, so one prompt corresponds to two RSR points.

B.6 VS-RECALL CONFIGURATION

The adversarial latent z_{adv} is initialized from a standard Gaussian in the latent space of the unlearned model's VAE and optimized with Adam. The timestep schedule visits 50 timesteps evenly spaced over the T=1000 diffusion range ($t = 10, 30, \ldots, 990$) in sequence, performing 100 gradient updates on the paired visual-state direction loss at each timestep, and the U-Net is conditioned on null text throughout. We group the tasks into abstract targets, namely nudity and the artistic styles, and

[1] https://github.com/joycenerd/P4D
[2] https://github.com/NYU-DICE-Lab/circumventing-concept-erasure
[3] https://github.com/OPTML-Group/Diffusion-MU-Attack
[4] https://github.com/kevinlu4588/WhenAreConceptsErased
[5] https://github.com/ryliu68/RECALL

object targets, and use one learning rate per group, 1.75×10^{-3} and 1.0×10^{-3}. Each value is selected empirically and then held fixed across all tasks, unlearning methods, reference pairs, and backbones in its group. The optimization seed is fixed to 2025. To evaluate recall, the decoded adversarial image is supplied as an image prompt together with the evaluation prompt through the image-to-image pathway of the unlearned model at its default strength of 0.8, using each prompt's fixed generation seed and guidance scale (7.5).

C Style Results under the Top-5 Criterion

Sec. 5 reports the style tasks under a top-3 classifier criterion. Table 5 gives the top-5 results, obtained from separate runs in which the top-5 criterion also serves as the stopping rule. The relaxed criterion puts VS-RECALL ahead of UnlearnDiffAtk on Van Gogh by 1.2 points (89.80 against 88.60) after trailing by 3.0 at top-3, and widens its lead on Monet from 0.4 to 1.0 point (94.00 against 93.00). The top-k criterion thus affects the fine-grained ranking but not the conclusion that the erased styles remain readily recoverable by several iterative attacks.

Table 5: Recall success rate (%) on the Style tasks under the top-5 criterion.

Van Gogh												
Defense	Text-only	Image-only	Text & R-noise	Text & Image	P4D-N	P4D-K	CCE	UDAtk	WACE-N	WACE-C	RECALL	**Ours**
ESD	20.00	2.00	8.00	0.00	84.00	**86.00**	66.00	84.00	44.00	10.00	80.00	74.00
FMN	44.00	4.00	10.00	4.00	76.00	94.00	62.00	92.00	46.00	22.00	90.00	**100.00**
SPM	70.00	2.00	12.00	32.00	98.00	98.00	98.00	98.00	80.00	74.00	**100.00**	**100.00**
RECE	52.00	2.00	12.00	6.00	96.00	**100.00**	78.00	98.00	74.00	32.00	96.00	**100.00**
UCE	86.00	2.00	28.00	56.00	96.00	**100.00**	96.00	**100.00**	94.00	82.00	**100.00**	**100.00**
AdvUnlearn	10.00	2.00	6.00	0.00	46.00	58.00	**78.00**	70.00	20.00	4.00	58.00	66.00
MACE	44.00	2.00	0.00	4.00	96.00	**100.00**	12.00	90.00	70.00	18.00	92.00	90.00
DoCo	42.00	4.00	26.00	16.00	80.00	78.00	32.00	84.00	72.00	64.00	90.00	**94.00**
ConceptPrune	96.00	2.00	70.00	92.00	**100.00**	**100.00**	98.00	**100.00**	94.00	94.00	**100.00**	**100.00**
Receler	16.00	4.00	2.00	6.00	38.00	48.00	16.00	70.00	18.00	0.00	64.00	**74.00**
Average	48.00	2.60	17.40	21.60	81.00	86.20	63.60	88.60	61.20	40.00	87.00	**89.80**
Picasso												
Defense	Text-only	Image-only	Text & R-noise	Text & Image	P4D-N	P4D-K	CCE	UDAtk	WACE-N	WACE-C	RECALL	**Ours**
ESD	30.00	42.00	6.00	26.00	94.00	90.00	32.00	88.00	**100.00**	**100.00**	**100.00**	**100.00**
FMN	66.00	40.00	32.00	50.00	**100.00**	**100.00**	48.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
SPM	88.00	40.00	38.00	74.00	**100.00**	**100.00**	38.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
RECE	80.00	30.00	30.00	64.00	**100.00**	**100.00**	36.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
UCE	78.00	38.00	48.00	76.00	**100.00**	**100.00**	28.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AdvUnlearn	8.00	32.00	18.00	22.00	66.00	56.00	42.00	84.00	88.00	**100.00**	94.00	**100.00**
MACE	58.00	44.00	28.00	48.00	**100.00**	**100.00**	48.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
DoCo	84.00	34.00	40.00	84.00	**100.00**	**100.00**	40.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
ConceptPrune	96.00	36.00	80.00	98.00	**100.00**	**100.00**	52.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Receler	24.00	36.00	10.00	22.00	80.00	76.00	18.00	94.00	88.00	**100.00**	98.00	**100.00**
Average	61.20	37.20	33.00	56.40	94.00	92.20	38.20	96.60	97.60	**100.00**	99.20	**100.00**
Monet												
Defense	Text-only	Image-only	Text & R-noise	Text & Image	P4D-N	P4D-K	CCE	UDAtk	WACE-N	WACE-C	RECALL	**Ours**
ESD	24.00	0.00	0.00	2.00	56.00	62.00	24.00	76.00	24.00	4.00	68.00	**78.00**
FMN	46.00	2.00	6.00	14.00	94.00	94.00	50.00	**98.00**	50.00	26.00	96.00	94.00
SPM	54.00	0.00	10.00	18.00	92.00	98.00	56.00	98.00	62.00	28.00	**100.00**	**100.00**
RECE	46.00	0.00	16.00	30.00	94.00	**100.00**	62.00	96.00	62.00	34.00	**100.00**	**100.00**
UCE	60.00	0.00	10.00	52.00	96.00	**100.00**	66.00	**100.00**	62.00	36.00	**100.00**	**100.00**
AdvUnlearn	20.00	0.00	0.00	0.00	72.00	60.00	32.00	**80.00**	14.00	6.00	72.00	76.00
MACE	46.00	0.00	12.00	16.00	90.00	**100.00**	62.00	98.00	68.00	26.00	**100.00**	**100.00**
DoCo	70.00	0.00	42.00	52.00	98.00	**100.00**	74.00	**100.00**	60.00	54.00	**100.00**	**100.00**
ConceptPrune	96.00	0.00	76.00	92.00	**100.00**	**100.00**	98.00	**100.00**	92.00	86.00	**100.00**	**100.00**
Receler	26.00	0.00	0.00	6.00	68.00	64.00	36.00	84.00	28.00	10.00	84.00	**92.00**
Average	48.80	0.20	17.20	28.20	86.00	87.80	56.00	93.00	52.20	31.00	92.00	**94.00**

D Generalization

D.1 Across Reference Pairs

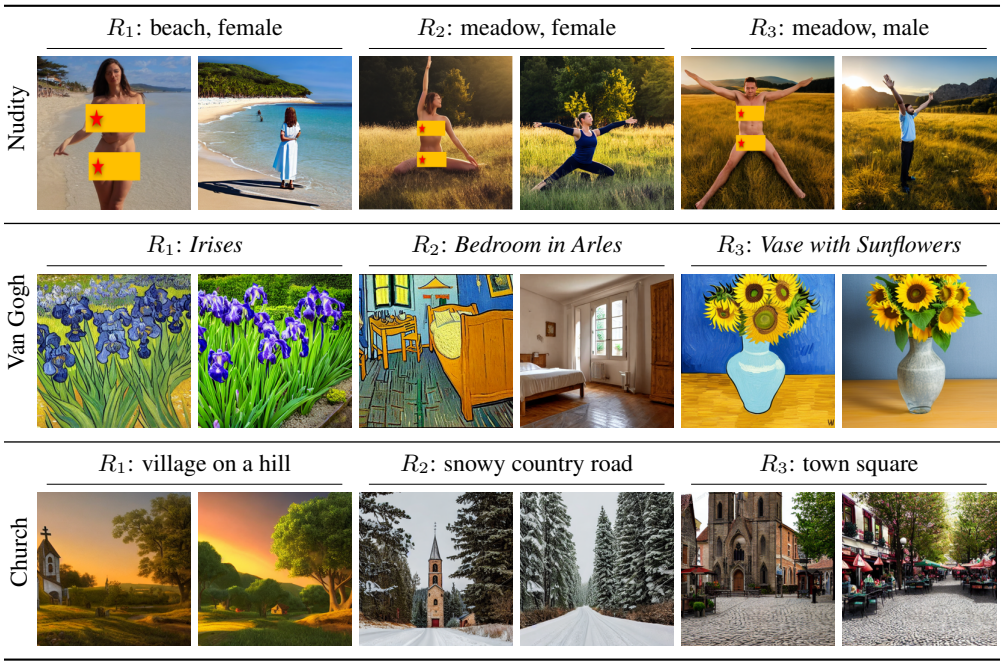
VS-RECALL takes its entire concept specification from one pair of reference images, so that pair is a free variable of the method rather than part of the model under audit. We quantify its effect by rerunning VS-RECALL with alternative pairs. For three concepts (Nudity-I2P, Van Gogh at top-3, and Object-Church) and two defenses (ESD and UCE) we generated three additional pairs per concept with Stable Diffusion v1.4 and repeated the attack while holding prompts, checkpoints, detectors, schedule, and hyperparameters fixed. The alternative pairs vary the scene and, for Nudity, the subject, and are shown in Table 7. Table 6 reports the results.

The mean over ESD and UCE spans 3.9 points on Nudity-I2P (80.99 to 84.86), 4.0 on Van Gogh (78.00 to 82.00), and 3.0 on Object-Church (84.00 to 87.00), so the four pairs reach comparable

Table 6: Recall success rate (%) under alternative reference pairs. Each concept is varied independently. R_{ori} is the pair used in the main experiments (Table 4) and R_1–R_3 are the alternatives of Table 7, numbered separately within each concept, so that R_1 under one concept bears no relation to R_1 under another. *Mean* averages the results of ESD and UCE.

Reference	Nudity-I2P			Van Gogh			Object-Church		
	ESD	UCE	Mean	ESD	UCE	Mean	ESD	UCE	Mean
R_{ori}	90.14	79.58	84.86	62.00	100.00	81.00	98.00	76.00	87.00
R_1	85.21	79.58	82.40	56.00	100.00	78.00	96.00	78.00	87.00
R_2	80.28	81.69	80.99	58.00	100.00	79.00	96.00	74.00	85.00
R_3	84.51	83.10	83.81	64.00	100.00	82.00	94.00	74.00	84.00

Table 7: The alternative reference pairs R_1–R_3 of Table 6, numbered within each concept. Each labelled column holds one pair, with the positive reference on the left and the negative reference on the right, and the corresponding R_{ori} pairs are shown separately in Table 4. Sensitive regions in the Nudity row are covered.



recall on all three tasks. On Nudity-I2P and Van Gogh the spread under ESD exceeds that of the mean. Under ESD, Nudity-I2P ranges from 80.28 to 90.14 and Van Gogh from 56.00 to 64.00, whereas under UCE the same tasks range from 79.58 to 83.10 and stay at 100.00 for all four pairs.

D.2 Across Model Versions

All main experiments use Stable Diffusion v1.4, the backbone shared by the ten unlearning methods we evaluate. To test whether visual-state recall is specific to that backbone, we repeat the audit against UCE-unlearned checkpoints built on Stable Diffusion v2.0-base and v2.1-base for Nudity-I2P, Van Gogh (top-3), and Object-Church, keeping the same reference pairs, schedule, and hyper-parameters. Figure 3 reports the results alongside the v1.4 results for the same defense.

Performance on Nudity remains similar across model versions, Van Gogh reaches 100.00 under UCE on all three backbones, and Object-Church is at least as recoverable on SD v2.0 and v2.1. These results show that the observed recall behavior is not restricted to a single Stable Diffusion version.

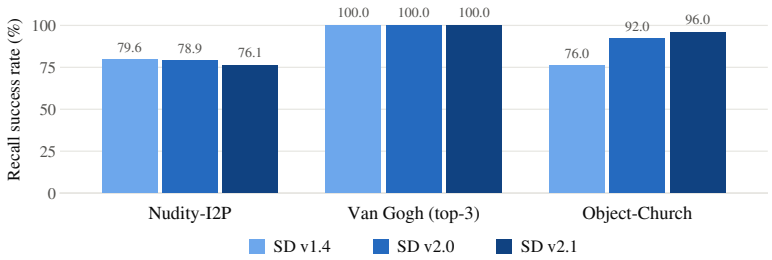


Figure 3: Recall success rate (%) of VS-RECALL against UCE-unlearned models built on three Stable Diffusion backbones, on Nudity-I2P, Van Gogh, and Object-Church. The reference pairs, timestep schedule, and hyperparameters are identical to the main experiments, and the style task is scored under the top-3 criterion.

E Semantic Alignment and Reference Similarity

E.1 Semantic Alignment

The task-specific detectors measure whether the erased concept is present, not whether the generated image still follows its prompt. We therefore compute the CLIP Score (Hessel et al., 2021) between each evaluation prompt and each successful recalled image, using CLIP ViT-B/32 (Radford et al., 2021) and the definition $100 \cdot \max(\cos(e_{\text{text}}, e_{\text{image}}), 0)$, where e_{text} and e_{image} are the CLIP embeddings of the evaluation prompt and the recalled image, and average it per task and defense. Figure 4 reports the scores for UCE, the defense in the top row of Figure 2, and MACE next to a reference point, the original SD v1.4 generating from the same prompts with the same per-prompt seed and guidance, averaged over all prompts.

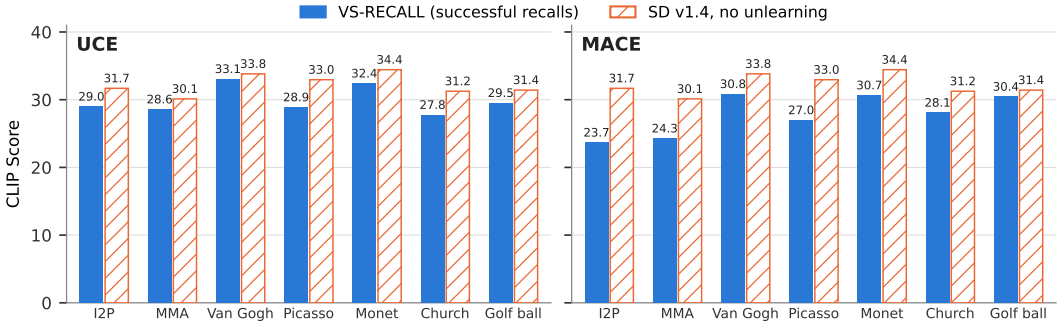


Figure 4: CLIP Score (↑) of successful VS-RECALL recalls against UCE (left) and MACE (right), alongside the original SD v1.4 averaged over all prompts.

Against UCE the mean scores of successful recalls lie between 27.8 and 33.1, and against MACE between 23.7 and 30.8, while the original model scores between 30.1 and 34.4. The gap to the original model is 0.7 points on Van Gogh under UCE and 1.0 on Golf ball under MACE. On the remaining task and defense pairs it stays within 6.0 points, except on I2P under MACE, where it reaches 8.0. The two quantities are not strictly matched, since recall scores average over successful samples only. The scores provide evidence that successful recalls remain aligned with their evaluation prompts.

E.2 Similarity to the Reference

To check whether recalled images reproduce I_{pos}, we compute the DINOv2 ViT-B/14 (Oquab et al., 2024) cosine similarity between each optimized successful output and I_{pos}, and compare it with the same measurement for images generated by the original model from the same prompts (Figure 5). On Van Gogh the original model scores 0.355 and the per-defense averages of VS-RECALL lie between 0.158 and 0.314, and on Object-Church 0.210 against 0.071 to 0.161. The per-defense mean similarities are therefore lower than those of the original model's generations.

To test whether the residual similarity is specific to I_{pos}, we also measure each output's similarity to 25 images of the same concept generated by the original model and not used in the attack, and define its preference for I_{pos} as its similarity to I_{pos} minus its mean similarity to that set. Subtracting the preference that the original model's own generations show for I_{pos} gives -0.030 on Van Gogh and

-0.012 on Object-Church, so on average the recalled images prefer I_{pos} no more than independently generated images of the concept do.

The mean pairwise similarity among recalled outputs, averaged over the ten defenses, is 0.207 and 0.142, below the 0.260 and 0.193 measured among the original model's own generations for the same prompts. Figures 6 and 7 show four successful recalls per defense next to I_{pos}.

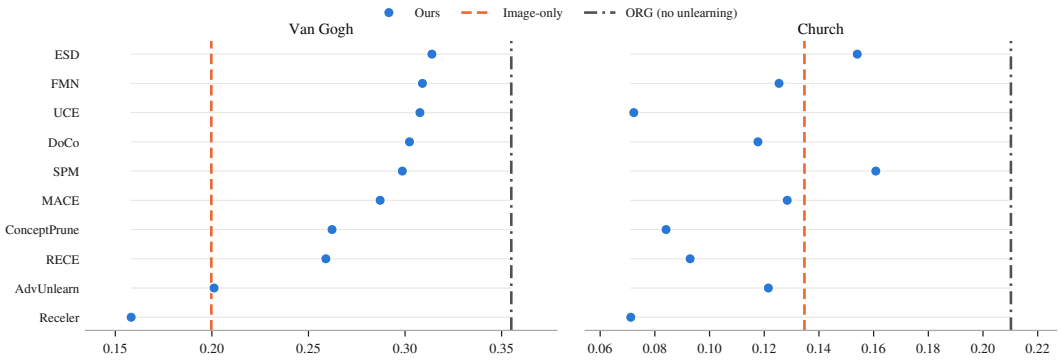


Figure 5: DINOv2 cosine similarity between successful VS-RECALL outputs and the positive reference I_{pos} on Van Gogh (left) and Object-Church (right). Each dot is one defense, the dash-dotted line is the same measurement for images generated by the original model from the same prompts, and the dashed line is that of the Image-only baseline of Sec. 5. Only optimized successes are included.



Figure 6: Successful recalls on Van Gogh. The left column is I_{pos}, and each remaining column shows four successful recalls against one defense.



Figure 7: Successful recalls on Object-Church. The left column is I_{pos}, and each remaining column shows four successful recalls against one defense.

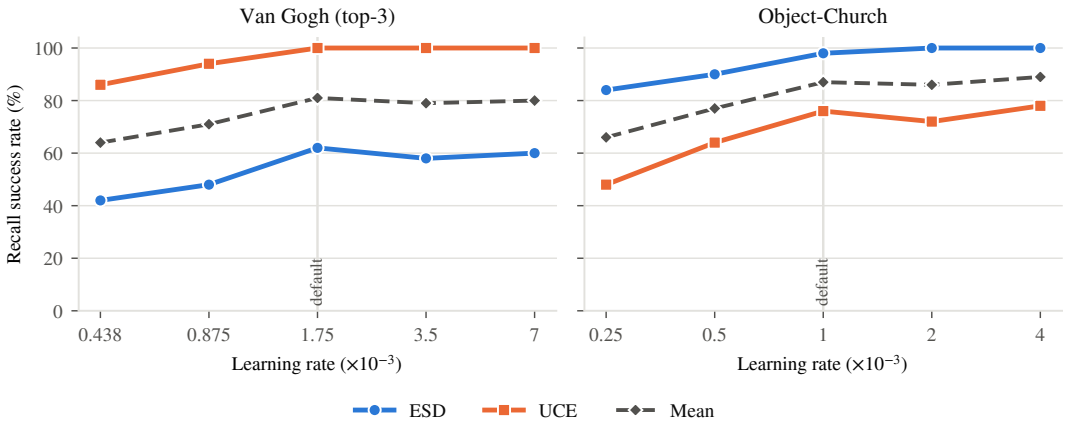


Figure 8: Recall success rate (%) as a function of the Adam learning rate, for Van Gogh at top-3 (left) and Object-Church (right). Each panel shows ESD, UCE, and their mean, and the vertical line marks the default learning rate of the main experiments. The horizontal axis is logarithmic, in units of 10^{-3}.

F Ablation Study

F.1 Design Components

Table 8 reports the ablations analyzed in Sec. 5.7. The single-reference variant replaces the two displacements in Eq. (8) by the states themselves and minimizes $1 - \cos(\mathrm{vec}(\bar{h}(z_{\mathrm{adv}}, t)), \mathrm{vec}(\bar{h}(z_{\mathrm{pos}}, t)))$, so that the adversarial state is matched to the positive-reference state directly, without the negative-reference anchor.

Table 8: Ablations of VS-RECALL. Each factor is varied with the others held at the default (marked †), and we report recall success rate (%) separately for ESD and UCE.

Factor	Variant	Nudity-I2P		Van Gogh		Object-Church	
		ESD	UCE	ESD	UCE	ESD	UCE
Visual-state layer	Down-block	91.55	83.10	52.00	100.00	100.00	74.00
	Mid-block †	90.14	79.58	62.00	100.00	98.00	76.00
	Up-block	78.17	78.17	58.00	100.00	100.00	78.00
Reference direction	Single reference	74.65	80.28	46.00	100.00	100.00	72.00
	Paired reference †	90.14	79.58	62.00	100.00	98.00	76.00
Normalization	Without normalization	86.62	79.58	52.00	100.00	100.00	72.00
	With normalization †	90.14	79.58	62.00	100.00	98.00	76.00
Optimization conditioning	Evaluation-prompt conditioning	88.03	85.92	50.00	100.00	100.00	68.00
	Null-text conditioning †	90.14	79.58	62.00	100.00	98.00	76.00

F.2 Learning Rate

We vary the Adam learning rate on Van Gogh at top-3 and Object-Church, one task from each group of Appendix B.6, for the same two defenses (ESD and UCE), by sweeping it logarithmically around each concept's default at $0.25\times$, $0.5\times$, $1\times$, $2\times$, and $4\times$ with $n = 50$ prompts per cell. The $1\times$ point is the configuration of the main experiments, with defaults of 1.75×10^{-3} for Van Gogh and 1.0×10^{-3} for Object-Church. Figure 8 plots the outcome.

On Van Gogh, ESD rises from 42.00 at $0.25\times$ to 62.00 at $1\times$ and stays between 58.00 and 62.00 above it, and UCE rises from 86.00 to 100.00 at $1\times$ and remains there. On Object-Church, ESD rises from 84.00 at $0.25\times$ to 100.00 at $2\times$ and above, and UCE rises from 48.00 to 76.00 at $1\times$ and then stays between 72.00 and 78.00. Recall is thus lowest at the smallest rates and varies little from the default upward on both concepts. Across the swept rates the mean over ESD and UCE spans 17.0 points on Van Gogh and 23.0 on Object-Church, and on both concepts the default lies near the upper end of that range, which supports using one learning rate across ESD and UCE for each of the two concepts swept here.