WHEN COMPATIBLE GRADIENTS CONFLICT OVER TIME: TRAJECTORY-AWARE DIFFUSION MODEL UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept erasure in text-to-image diffusion models must prevent adaptive attacks from recovering an erased concept while preserving neighboring semantics and generation quality. Existing methods largely address this robustness–retention trade-off through loss balancing or single-step gradient surgery, implicitly treating erase–retain interference as a local optimization phenomenon. In this paper, we show that this view is incomplete. Across unsafe-content, object, and artistic-style erasure, erase and retain gradients are nearly orthogonal or weakly aligned at individual steps, yet their accumulated directions become increasingly antagonistic over training. We attribute this reversal to two forms of optimizer-history dependence. Specifically, shared first- and second-moment optimizer states mix unmatched erase and retain gradients across optimization steps, creating trajectory conflict despite same-step compatibility. Meanwhile, fixed optimizer memory cannot accommodate directional changes within each objective trajectory. Based on this diagnosis, we propose a trajectory-aware Robust-Retain Concept Erasure framework (RRCE) that alternates recovery discovery with erase–retain defense. The recovery phase uses textual inversion to discover target representations that can still recover the target concept from the current model. The defense phase suppresses these representations through anchor-guided erasure and preserves non-target behavior through teacher-matching retention. To address the two trajectory-level failures, RRCE maintains objective-specific optimizer states and adapts momentum according to adjacent-gradient similarity within each trajectory. Experiments demonstrate state-of-the-art robustness–retention performance against adaptive recovery attacks.

1 INTRODUCTION

Text-to-image diffusion models have shown remarkable generative capabilities (Rombach et al., 2022; Saharia et al., 2022), but they may also produce undesirable concepts, including unsafe content, copyrighted styles, and unwanted object categories. Concept erasure aims to remove specified target concepts from a pretrained diffusion model while preserving non-target semantics and overall generation quality (Gandikota et al., 2023; Huang et al., 2024). However, achieving effective removal with reliable retention remains challenging. More aggressive erasure can damage non-target concepts (Saha et al., 2025; Kim et al., 2026), whereas stronger preservation constraints can weaken removal and leave residual target representations vulnerable to adaptive recovery attacks (Zhang et al., 2024b; Pham et al., 2024).

Existing work has sought to improve this erase–retain trade-off. One line of work introduces weighted retention losses over selected non-target concepts to jointly optimize erasure and preservation (Lu et al., 2024; Thakral et al., 2025; Li et al., 2025; Kim et al., 2026; Tu et al., 2026). Another line formulates erase–retain interference as an optimization problem and coordinates the two objectives through single-step gradient surgery (Wu et al., 2025; Li et al., 2026; Patel & Qiu, 2025; Fan et al., 2026). Despite these advances, both lines of work characterize erase–retain interference at individual optimization steps, either through objective losses or instantaneous gradients. This leaves the cumulative cross-time interaction between erase and retain updates over the full training trajectory largely unexplored.

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.

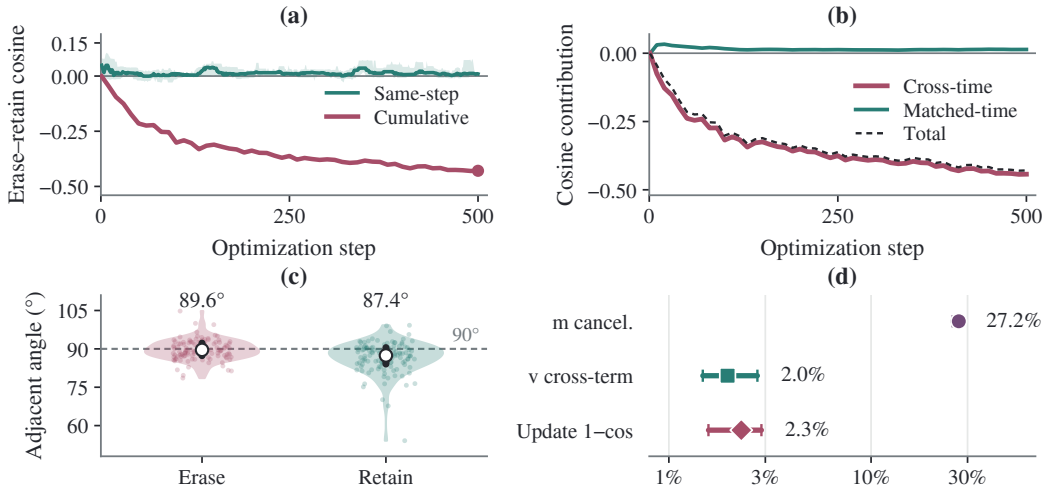


Figure 1: Trajectory diagnostics. (a) Same-step erase–retain cosine remains near zero, while the cumulative cosine becomes negative. (b) Cross-time interactions drive this reversal; the total nearly coincides with the cross-time contribution. (c) Adjacent gradients within both objectives are nearly orthogonal, revealing strong step-to-step rotation. (d) Shared Adam exhibits first-moment cancellation, second-moment cross-terms, and a nonzero update-direction gap $1 - \cos(u_{\text{shared}}, u_{\text{separate}})$.

To investigate this limitation, we analyze the optimization dynamics of an erase–retain training objective. The trajectories reveal two key patterns (Fig. 1). First, erase and retain gradients remain nearly orthogonal at each step, i.e., with near-zero cosine similarity throughout training. In contrast, their accumulated directions become progressively antagonistic over time (Fig. 1a). A cumulative gradient decomposition attributes this reversal primarily to cross-time interactions rather than same-step interference (Fig. 1b). Second, even within each objective, consecutive gradients exhibit substantial directional changes, indicating that the local optimization geometry evolves throughout training (Fig. 1c). Together, these observations suggest that erase–retain optimization is governed by historical interactions accumulated both across objectives and within each objective trajectory.

We further ask why these history-dependent trajectory effects arise despite near-orthogonal gradients at individual steps. Our analysis reveals that they originate from two forms of optimizer-history dependence. The first is cross-objective state coupling. Under a shared optimizer, the first-moment state mixes erase gradient g_i^{E} and retain gradient g_i^{R} from different update pairs into a single history. Same-step interactions capture only matched terms $\langle g_i^{\mathrm{E}}, g_i^{\mathrm{R}} \rangle$, whereas cross-time pairs $(g_i^{\mathrm{E}}, g_j^{\mathrm{R}})$ with $i \neq j$ remain unconstrained and can produce a negative long-horizon interaction. The shared second moment introduces another coupling channel by using squared gradients from one objective to affect the coordinate-wise preconditioning of the other. As illustrated in Fig. 1d, these shared states lead to first-moment cancellation, second-moment contamination, and drift in the update direction.

The second is a mismatch between fixed optimizer memory and evolving within-objective geometry. For objective $q \in \{\mathrm{E}, \mathrm{R}\}$, consecutive gradients satisfy the approximation

$$g_{t+1}^q - g_t^q \approx H_t^q(\theta_{t+1} - \theta_t), \tag{1}$$

where H_t^q is the local Hessian. Therefore, parameter updates can alter the gradient direction as the local geometry evolves, consistent with the large adjacent-gradient angles in Fig. 1c. A fixed momentum coefficient imposes a single memory horizon across geometric regimes: long memory carries outdated directions through sharp turns, whereas short memory provides insufficient averaging along locally stable segments. Together, these two failure modes show why a shared optimizer with fixed memory is ill-suited to erase–retain training: it entangles the histories of distinct objectives while failing to track the evolving geometry of either trajectory.

Guided by this insight, we propose a trajectory-aware Robust-Retain Concept Erasure framework (*RRCE*) that alternates recovery discovery with erase–retain defense in a min-max pipeline. In the maximization phase, textual inversion identifies text conditions that can still recover the target concept from the current model. In the minimization phase, anchor-guided erasure is followed by

teacher-matching retention. These updates are coordinated by our *Decoupled and Stabilized Erase–Retain Optimization*, which operates at two complementary levels. The *Cross-Objective Decoupled Optimizer* assigns separate first- and second-moment states to erase and retain updates, preventing their histories from contaminating each other. The *Intra-Objective Momentum Stabilizer* adapts each objective's memory horizon according to adjacent-gradient similarity, shortening memory near sharp turns and extending it along stable trajectory segments. Across erasure benchmarks, RRCE achieves state-of-the-art (SoTA) robustness-retention performance against adaptive recovery attacks.

Our main contributions are summarized as follows:

- We reveal that locally near-orthogonal erase–retain gradients can conflict after cross-time accumulation, and that both objectives exhibit within-trajectory directional changes. We explain these phenomena via optimizer-history dynamics and optimization geometry.
- We propose the Robust-Retain Concept Erasure (RRCE) framework with Decoupled and Stabilized Erase–Retain Optimization, separating cross-objective optimizer histories and stabilizing momentum within each trajectory.
- Across concept erasure benchmarks, our method achieves SoTA robustness–retention performance against adaptive recovery attacks.

2 Related Work

Retain-Aware Concept Erasure. Retain-aware concept erasure explicitly protects non-target concepts through preservation objectives, protected concept sets, or constrained update subspaces (Gandikota et al., 2024; Liu et al., 2024; Fan et al., 2024; Han et al., 2025; Huang et al., 2024; Lu et al., 2024; Park et al., 2024; Bui et al., 2024; Li et al., 2025; Bui et al., 2025; Deng et al., 2026; Sun et al., 2026). FADE (Thakral et al., 2025) constructs a Concept Neighborhood around each target and combines expungement, adjacency, and guidance losses to preserve both adjacent and unrelated concepts. ReCARE (Kim et al., 2026) constructs a CARE set of benign concepts that co-occur in target images and incorporates it into training, strengthening target erasure against recovery while preserving these co-occurring concepts. SuPLoRA (Tu et al., 2026) groups target subtypes under shared supertypes and jointly erases each group with LoRA, freezing the down-projection that encodes supertype directions while updating only the up-projection to preserve superordinate semantics. These methods primarily improve retention by imposing explicit constraints over selected concepts or semantic structures. In contrast, our method models how erase and retain objectives interact over the full optimization trajectory, addressing trajectory-level conflict rather than only strengthening preservation constraints.

Robust Concept Erasure. Robust concept erasure aims to prevent an erased concept from being recovered through adaptive attacks. Such attacks may use discrete adversarial tokens or optimize continuous text embeddings to reactivate the target concept (Kim et al., 2024; Richardson et al., 2025; Ahn & Jung, 2025; Kim et al., 2026; Li et al., 2026; Aladawi et al., 2026). Existing methods improve erasure by explicitly searching for recoverable conditions and incorporating them into the erasure process. RECE (Gong et al., 2024) alternates between two closed-form steps: it derives embeddings that recover the erased concept from the current model, then edits cross-attention projections to map them to harmless concepts; an embedding regularizer limits changes to unrelated concepts. AdvUnlearn (Zhang et al., 2024a) formulates robustification as bi-level adversarial training of the text encoder and adds retain-prompt regularization to recover generation utility. STEREO (Srivatsan et al., 2025) first searches for adversarial tokens by alternating erasure and textual inversion, then performs anchor-guided compositional erasure against those adversarial tokens. RRCE builds on this recovery-aware paradigm but addresses a critical optimization question: how the resulting robust-erasure objective should be coordinated with retention over the optimization trajectory.

Gradient Surgery. Gradient surgery coordinates competing erase and retain objectives by reshaping current gradients or inducing local alignment during optimization (Patel & Qiu, 2025; Fan et al., 2026). DoCo (Wu et al., 2025) couples adversarial domain correction with concept-preserving gradient surgery to remove components that interfere with retaining model utility. AEGIS (Li et al., 2026) pairs an adversarial erasure target with retention-data-free gradient regularization projection, rectifying the retention update only when it opposes erasure. These methods coordinate erase and retain objectives through current-step projection or local gradient conflict estimation. RRCE instead

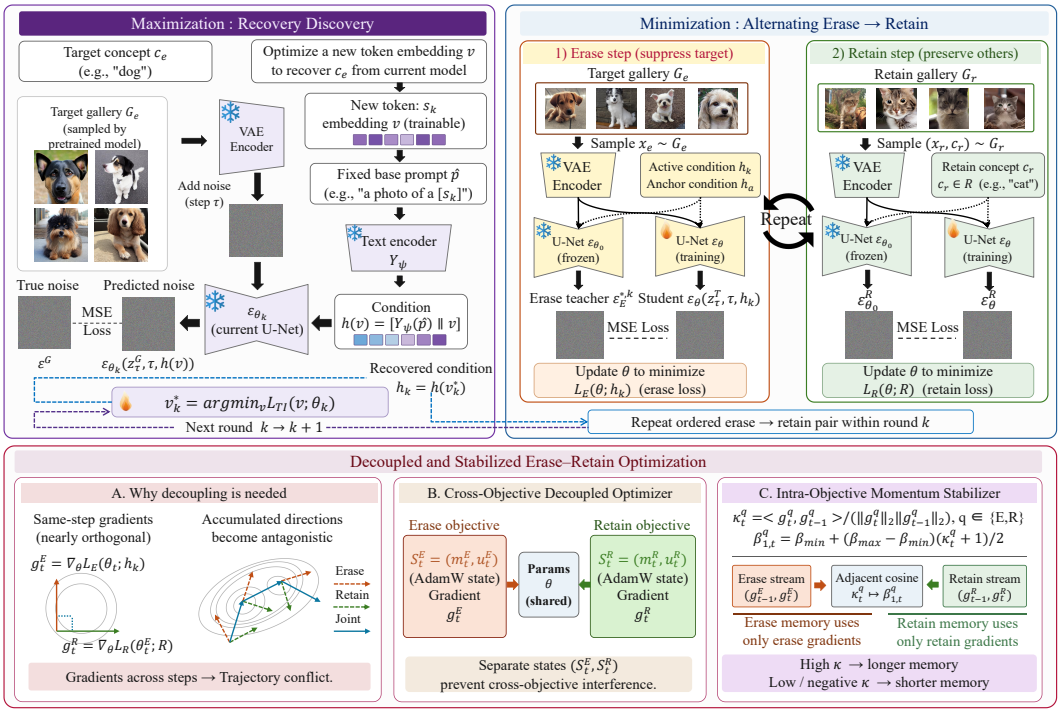


Figure 2: Robust-retain concept erasure. Each round discovers a recovery token, then alternates anchor-guided erase and teacher-matching retain updates. Erase and retain maintain separate optimizer states, while adjacent-gradient similarity lengthens memory on smooth segments and shortens it near sharp turns.

treats objective coordination as a trajectory-level problem, accounting for both cross-time interactions between the two objectives and history dependence within each objective.

3 METHOD

Figure 2 provides an overview of our Robust-Retain Concept Erasure framework (RRCE), which combines an iterative recovery-defense pipeline with trajectory-aware erase–retain optimization. In this pipeline, recovery discovery identifies a text condition that can still recover the target concept from the current model. The subsequent defense block first suppresses this condition through anchor-guided erasure and then matches the pretrained model on retain concepts. Within each defense block, erase and retain updates are applied sequentially and coordinated by Decoupled and Stabilized Erase–Retain Optimization, which maintains objective-specific optimizer states and adapts the moment horizon for each objective trajectory.

3.1 ROBUST-RETAIN CONCEPT ERASURE PIPELINE

Let ϵ_θ denote the trainable U-Net, initialized at the pretrained parameters θ_0. We use $\mathcal{E}$ and $\mathcal{Y}_\psi$ to denote the frozen VAE and text encoders, respectively. For a concept c, $\mathcal{Y}_\psi(c)$ represents the text condition. Let c_e be the target concept and $\mathcal{R}$ the retain concept set. We use θ_k to denote the model parameters after the k-th defense block. The initial defense block uses $h_0 = \mathcal{Y}_\psi(c_e)$ as its active erase condition and produces θ_1. Thereafter, for each subsequent round $k \geq 1$, RRCE performs recovery discovery followed by defense. With the current U-Net fixed at ϵ_{θ_k}, recovery discovery identifies a textual condition h_k that can still recover c_e. The subsequent defense block suppresses h_k through ordered erase–retain updates, yielding θ_{k+1}.

Max: recovery discovery. We construct a target image gallery $\mathcal{G}_e$ by sampling the pretrained generator with prompts expressing the target concept c_e. For $x_e \sim \mathcal{G}_e$, the frozen VAE encoder produces $z_0^{\mathrm{G}} = \mathcal{E}(x_e)$ and adds Gaussian noise: $z_\tau^{\mathrm{G}} = \alpha_\tau z_0^{\mathrm{G}} + \sigma_\tau \epsilon^{\mathrm{G}}$, where $\epsilon^{\mathrm{G}} \sim \mathcal{N}(0, I)$. τ is

a sampled diffusion step, while α_τ and σ_τ are its signal and noise coefficients. At round $k \geq 1$, we introduce a new token s_k with trainable embedding v and insert it to a fixed prompt $\hat{p}$. Let $\hat{p}[v]$ denote the prompt-embedding sequence and $h(v) = \mathcal{Y}_\psi(\hat{p}[v])$ the contextualized text condition. The U-Net, VAE, text encoder, and vocabulary embeddings remain fixed, and only v is optimized:

$$v_k^\star = \arg\min_v \mathbb{E}_{x_e \sim \mathcal{G}_e, \tau, \epsilon^{\mathrm{G}}} \left\| \epsilon^{\mathrm{G}} - \epsilon_{\theta_k}(z_\tau^{\mathrm{G}}, \tau, h(v)) \right\|_2^2. \tag{2}$$

The target in Eq. (2) is the exact noise added to the encoded gallery image. Therefore, solving Eq. (2) identifies a textual embedding $v_k^\star$ whose condition $h_k = h(v_k^\star)$ can recover the target concept from the current erased model. This recoverable condition is then used as the adversarial erase target in the subsequent defense block.

Min: alternating erase–retain updates. The defense phase performs ordered erase–retain updates using the target image gallery $\mathcal{G}_e$ and the retain image gallery $\mathcal{G}_r$. The target gallery contains images sampled from the target concept, while each retain sample is a pair (x_r, c_r) with $c_r \in \mathcal{R}$. We sample $x_e \sim \mathcal{G}_e$ and $(x_r, c_r) \sim \mathcal{G}_r$, and construct the corresponding noisy target and retain latents:

$$z_\tau^{\mathrm{T}} = \alpha_\tau \mathcal{E}(x_e) + \sigma_\tau \epsilon^{\mathrm{T}},\ \epsilon^{\mathrm{T}} \sim \mathcal{N}(0, I), \quad z_\tau^{\mathrm{R}} = \alpha_\tau \mathcal{E}(x_r) + \sigma_\tau \epsilon^{\mathrm{R}},\ \epsilon^{\mathrm{R}} \sim \mathcal{N}(0, I). \tag{3}$$

where the target latent z_τ^T supports erasure, while the retain latent z_τ^R supports preservation. Within each loss, the student and its reference target receive the same latent. Their discrepancy therefore measures a change in denoising behavior rather than a change in input state.

Erase Update. Let $\mathcal{A}$ denote the anchor concept set that provides a benign semantic direction for redirecting the erased prediction. We sample $c_a \sim \mathcal{A}$, and define $h_a = \mathcal{Y}_\psi(c_a)$. We use $\varnothing$ for the unconditional condition. For $\chi \in \{h_k, h_a\}$, we define $b_\chi = \epsilon_{\theta_0}(z_\tau^{\mathrm{T}}, \tau, \chi)$ and compute the pretrained guidance residual $\Delta_\chi = b_\chi - \epsilon_{\theta_0}(z_\tau^{\mathrm{T}}, \tau, \varnothing)$. We decompose this residual into components parallel and orthogonal to b_χ: $\Delta_\chi^{\parallel} = \frac{\langle \Delta_\chi, b_\chi \rangle}{\|b_\chi\|_2^2} b_\chi$ and $\Delta_\chi^{\perp} = \Delta_\chi - \Delta_\chi^{\parallel}$. Moreover, we define $P_\rho(\Delta_\chi) = \Delta_\chi^{\perp} + \rho \Delta_\chi^{\parallel}$, where $\rho \geq 0$ controls the contribution of the parallel component. Using the guidance scale γ and the stop-gradient operator $\mathrm{sg}[\cdot]$, we define the erase loss directly against the frozen teacher target

$$\mathcal{L}_{\mathrm{E}}(\theta; h_k) = \mathbb{E}_{\mathrm{E}} \Big[\big\| \epsilon_\theta(z_\tau^{\mathrm{T}}, \tau, h_k) - \mathrm{sg}\big[\epsilon_{\theta_0}(z_\tau^{\mathrm{T}}, \tau, \varnothing) + (\gamma - 1)(P_\rho(\Delta_{h_a}) - P_\rho(\Delta_{h_k})) \big] \big\|_2^2 \Big]. \tag{4}$$

where the expectation $\mathbb{E}_{\mathrm{E}}$ is over $x_e \sim \mathcal{G}_e$, τ, ϵ^{T}, and $c_a \sim \mathcal{A}$. The negative residual suppresses the active target representation, while the positive residual redirects its prediction toward the anchor concept. Both terms from the frozen teacher provide a fixed erasure target as the student changes.

Retain Update. Let $\mathbb{E}_{\mathrm{R}}$ denote expectation over $(x_r, c_r) \sim \mathcal{G}_r$, τ, and ϵ^{R}. For compactness, we define $\epsilon_\theta^{\mathrm{R}} = \epsilon_\theta(z_\tau^{\mathrm{R}}, \tau, \mathcal{Y}_\psi(c_r))$ and $\epsilon_{\theta_0}^{\mathrm{R}} = \epsilon_{\theta_0}(z_\tau^{\mathrm{R}}, \tau, \mathcal{Y}_\psi(c_r))$. We minimize the objective by matching the student prediction to the frozen pretrained teacher

$$\mathcal{L}_{\mathrm{R}}(\theta; \mathcal{R}) = \mathbb{E}_{\mathrm{R}} \left[\left\| \epsilon_\theta^{\mathrm{R}} - \mathrm{sg}[\epsilon_{\theta_0}^{\mathrm{R}}] \right\|_2^2 \right]. \tag{5}$$

Equation (5) preserves pretrained denoising behavior over the retain concept set. Each defense block alternates Eq. (4) and Eq. (5), beginning with erasure. The initial block uses h_0, while each later block uses the condition h_k discovered from ϵ_{θ_k}. Repeating recovery discovery and defense adapts the erased representation to the evolving model.

3.2 Decoupled and Stabilized Erase–Retain Optimization

Within each defense block, erase and retain losses are optimized in ordered update pairs rather than being merged into a single joint objective. Let k index the outer defense round and let t index the erase–retain pairs along the optimization trajectory. For objective $q \in \{\mathrm{E}, \mathrm{R}\}$, g_t^q denotes the gradient supplied by q, and $\mathcal{S}_t^q$ denotes its optimizer state. We decouple the histories of the two objectives and adapt each trajectory's history length.

Cross-Objective Decoupled Optimizer. A conventional joint formulation combines erasure and retention into a single objective with one AdamW state:

$$\min_\theta \left[\mathcal{L}_{\mathrm{E}}(\theta; h_k) + \mathcal{L}_{\mathrm{R}}(\theta; \mathcal{R}) \right], \qquad \mathcal{S}_t^{\mathrm{S}} = (m_t^{\mathrm{S}}, u_t^{\mathrm{S}}). \tag{6}$$

This entangles the first- and second-moment histories associated with the two objectives. Our defense instead uses an ordered erase–retain schedule and assigns a separate optimizer state to each objective. Let θ_t^{E} denote the parameters after the erase update in pair t, with $g_t^{\mathrm{E}} = \nabla_\theta \mathcal{L}_{\mathrm{E}}(\theta_t; h_k)$ and $g_t^{\mathrm{R}} = \nabla_\theta \mathcal{L}_{\mathrm{R}}(\theta_t^{\mathrm{E}}; \mathcal{R})$. The two updates are

$$\begin{aligned} &\text{Erase:} \quad (\theta_t^{\mathrm{E}}, \mathcal{S}_t^{\mathrm{E}}) = \mathrm{AdamW}(\theta_t, g_t^{\mathrm{E}}; \mathcal{S}_{t-1}^{\mathrm{E}}), \\ &\text{Retain:} \quad (\theta_{t+1}, \mathcal{S}_t^{\mathrm{R}}) = \mathrm{AdamW}(\theta_t^{\mathrm{E}}, g_t^{\mathrm{R}}; \mathcal{S}_{t-1}^{\mathrm{R}}), \end{aligned} \tag{7}$$

where $\mathcal{S}_t^{\mathrm{E}} = (m_t^{\mathrm{E}}, u_t^{\mathrm{E}})$ is the erase-optimizer state and $\mathcal{S}_t^{\mathrm{R}} = (m_t^{\mathrm{R}}, u_t^{\mathrm{R}})$ is the retain-optimizer state. The model parameters remain shared and are updated sequentially. Only the optimizer histories are separated, so each state accumulates statistics exclusively from its corresponding objective.

To isolate the effect of optimizer-state sharing from that of update ordering, we consider a shared-state counterpart that processes the same ordered erase and retain gradient sequence using a single first-moment state with fixed coefficient β. Applying erase first and retain second yields

$$m_t^{\mathrm{S}} = \beta^2 m_{t-1}^{\mathrm{S}} + (1-\beta)\big(\beta g_t^{\mathrm{E}} + g_t^{\mathrm{R}}\big). \tag{8}$$

With the standard initialization $m_0^{\mathrm{S}} = 0$, let $a_{t,i}^{\mathrm{E}} = (1-\beta)\beta^{2(t-i)+1}$ and $a_{t,i}^{\mathrm{R}} = (1-\beta)\beta^{2(t-i)}$. Unrolling Eq. (8) yields $m_t^{\mathrm{S}} = m_t^{\mathrm{S,E}} + m_t^{\mathrm{S,R}}$, where $m_t^{\mathrm{S},q} = \sum_{i=1}^t a_{t,i}^q g_i^q$. Its squared norm equals

$$\big\|m_t^{\mathrm{S}}\big\|_2^2 = \Big\|m_t^{\mathrm{S,E}}\Big\|_2^2 + \Big\|m_t^{\mathrm{S,R}}\Big\|_2^2 + 2\sum_{i=1}^t \sum_{j=1}^t a_{t,i}^{\mathrm{E}} a_{t,j}^{\mathrm{R}} \left\langle g_i^{\mathrm{E}}, g_j^{\mathrm{R}} \right\rangle. \tag{9}$$

The terms with $i = j$ describe interactions within an erase–retain pair, whereas those with $i \neq j$ capture interactions across the trajectory. Even when within-pair gradients are nearly orthogonal, negative cross-time terms can accumulate and attenuate or rotate the shared moment. The shared second moment adds another coupling channel by using squared gradients from both objectives to scale every update. The decoupled fixed-β moments are instead

$$m_t^q = (1-\beta)\sum_{i=1}^t \beta^{t-i} g_i^q, \qquad q \in \{\mathrm{E}, \mathrm{R}\}. \tag{10}$$

Each moment represents one objective's trajectory. The same separation applies to second moments, so erase and retain gradients receive objective-specific temporal smoothing and preconditioning.

Intra-Objective Momentum Stabilizer. Objective-specific states remove direct cross-objective history mixing, but a fixed first-moment coefficient still imposes the same effective memory across an evolving trajectory. For the erase stream, recovery discovery changes the active condition h_k, and the retain gradient field induced by $\mathcal{R}$ turns as θ evolves. A fixed coefficient may carry an outdated direction through a sharp turn or discard a stable direction too quickly. We use adjacent-gradient similarity as a proxy for local trajectory smoothness and adapt the first-moment coefficient within each objective stream:

$$\kappa_t^q = \frac{\langle g_t^q, g_{t-1}^q \rangle}{\|g_t^q\|_2 \, \|g_{t-1}^q\|_2}, \quad \beta_{1,t}^q = \beta_{\min} + (\beta_{\max} - \beta_{\min})\frac{\kappa_t^q + 1}{2}. \tag{11}$$

Since $\kappa_t^q \in [-1, 1]$, the affine map places $\beta_{1,t}^q$ directly in $[\beta_{\min}, \beta_{\max}]$. We set $\kappa_t^q = 0$ if either gradient norm vanishes and initialize each stream with $\beta_{1,1}^q = \beta_{\max}$. A large κ_t^q indicates a directionally stable segment with little observed turning. A small or negative value indicates a sharper turn and shortens the history, giving the current gradient more influence. Each estimate uses adjacent gradients from the same objective stream.

The time-varying coefficient changes both the first moment and its bias correction. Let $B_t^q = \prod_{j=1}^t \beta_{1,j}^q$. The first moment has the exact expansion

$$m_t^q = \beta_{1,t}^q m_{t-1}^q + (1-\beta_{1,t}^q) g_t^q = \sum_{i=1}^t (1-\beta_{1,i}^q) \left(\prod_{j=i+1}^t \beta_{1,j}^q \right) g_i^q. \tag{12}$$

where an empty product equals one. The product in Eq. (12) is the effective memory kernel. A lower current coefficient contracts the contribution of all earlier gradients, while a higher value carries their

direction forward. The corresponding first-moment bias correction is $1 - B_t^q$, which accounts for the complete coefficient sequence.

The adaptive coefficient is applied independently within each objective-specific optimizer state in Eq. (7). It changes the moment horizon of the corresponding first moment without altering the optimizer states or the erase–retain update.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate concept erasure on three representative settings: nudity for unsafe-content removal, dog for object removal, and Van Gogh style for artistic-style removal. All erased models are initialized from Stable Diffusion v1.4 (Rombach et al., 2022). For nudity, we select 88 prompts from I2P (Schramowski et al., 2023) whose nudity score exceeds 50% and whose tokenized length is below 77. We generate five images per prompt, yielding 440 test images. For dog, the prompt "A photo of a dog" is evaluated with 500 fixed seeds. For Van Gogh, "A painting in the style of Van Gogh" is likewise evaluated with 500 fixed seeds. Every method uses the same prompts, seeds, classifier-free guidance scale of 7.5 (Ho & Salimans, 2022), and 50 steps.

Implementation Details. We compare RRCE with ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), RECE (Gong et al., 2024), MACE (Lu et al., 2024), AdvUnlearn (Zhang et al., 2024a), STEREO (Srivatsan et al., 2025), AGE (Bui et al., 2025), ReCARE (Kim et al., 2026), and AEGIS (Li et al., 2026). We use official implementations and recommended hyperparameters whenever available, and otherwise follow the settings reported in the corresponding papers. All methods are evaluated from Stable Diffusion v1.4 under the same generation protocol. For RRCE, we fine-tune the full U-Net for unsafe-content removal and object removal, and all cross-attention layers for artistic-style removal. We use two recovery-discovery rounds, giving three defense blocks including the initial block, with 500 ordered erase–retain update pairs per block. Textual inversion is optimized for 3,000 steps with learning rate 5×10^{-3}. The erase and retain optimizers use AdamW with learning rate 2×10^{-5}, $\beta_2 = 0.999$, and zero weight decay. The first-moment coefficient is bounded within $[0.1, 0.9]$. We set $\gamma = 4$, $\rho = 1$.

Adaptive recovery attacks. We evaluate robustness against two adaptive attacks: UnlearnDiff (UD) (Zhang et al., 2024b) and Circumventing Concept Erasure (CCE) (Pham et al., 2024). For UD, we run 40 iterations with AdamW (Loshchilov & Hutter, 2019), a learning rate of 0.01, and optimization over 50 diffusion timesteps. UD modifies at most five tokens for the 95 fixed nudity prompts and at most three tokens for the 50 fixed dog or Van Gogh prompts. For CCE, we learn a separate continuous attack embedding for every erased model while freezing the diffusion network. Textual inversion (Gal et al., 2023) is optimized for 3,000 iterations with a learning rate of 5×10^{-3} and a batch size of 1, using explicit samples from I2P for nudity, 30 fixed target images for dog, and 6 fixed target images for Van Gogh. Across both attacks, the samples used for attack optimization are disjoint from those used for evaluation.

Evaluation Metrics. We report attack success rate (ASR) under direct prompting (Erased), UD, and CCE, where lower is better. Nudity is detected by NudeNet (Bedapudi, 2019): an image is positive when any of six exposed-body-part labels exceeds 0.75. Dog is detected by Faster R-CNN (Ren et al., 2015) pretrained on COCO (Lin et al., 2014), with a confidence threshold of 0.5. Van Gogh style is detected when Van Gogh appears in the top-3 predictions of an ImageNet-pretrained (Deng et al., 2009) ViT-Base classifier (Dosovitskiy et al., 2021) fine-tuned on 129 WikiArt styles (Karayev et al., 2014). CCE is evaluated on 95 nudity prompts or 500 held-out dog and Van Gogh seeds. General utility is measured on the same COCO-30K prompts (Lin et al., 2014) using CLIP score (Radford et al., 2021) and FID (Heusel et al., 2017). To measure local preservation, each task defines 20 benign neighboring concepts. We generate 25 images per concept and compute CLIP R-Precision@1 (Radford et al., 2021; Xu et al., 2018) against the 20 neighbors and 80 unrelated COCO labels. Retain is the mean of these 500 binary retrieval outcomes, so higher is better.

4.2 EXPERIMENTAL RESULTS

Table 1 reports unsafe-content erasure. Our method surpasses all baselines in robust erasure

while matching the best neighboring-concept preservation, showing that recovery resistance need not sacrifice retention. Table 2 reports object erasure, where robustness and retention are harder to balance. Our method achieves the lowest CCE ASR of 9.20% and the highest Retain of 87.00% among erased models, whereas attack-resistant baselines often damage neighboring concepts. Table 3 covers artistic-style erasure. Our method alone resists all three recovery settings while preserving neighboring styles; methods with marginally better utility remain more vulnerable to adaptive recovery of the target style.

Across these settings, our method maintains competitive overall generation quality. Methods scoring better on CLIP or FID typically impose weaker recovery-robustness constraints, leaving them vulnerable to UD or CCE. AGE shifts its output distribution markedly closer to COCO-30K than SD v1.4, but poorly preserves neighboring concepts. RRCE instead jointly optimizes direct and adaptive erasure with neighboring-concept preservation, yielding a more complete robustness–retention balance despite modest CLIP/FID gaps.

Figure 3 contrasts CCE recovery with neighboring-concept retention. Under CCE, many baselines still generate recognizable dogs, Van Gogh portraits, or explicit bodies. Our method removes canine structure, replaces the Van Gogh identity and palette with a geometric composition, and produces a clothed portrait, suppressing all three targets. It also preserves a photorealistic goat, the characteristic Marc Chagall composition, and a coherent hugging interaction, whereas several baselines produce sketches, buildings, or unrelated scenes. Thus, our method resists CCE recovery while retaining nearby semantics and visual coherence across all three tasks, consistent with the quantitative robustness–retention results.

Table 1: Nudity erasure results.

Model	Erased↓	UD↓	CCE↓	Retain↑	CLIP↑	FID↓
SD v1.4	28.86	27.37	50.53	97.80	31.36	14.15
ESD	16.36	9.47	50.53	<u>93.00</u>	30.89	14.50
UCE	5.45	5.26	21.05	**93.20**	**31.08**	14.38
RECE	<u>1.14</u>	<u>1.05</u>	5.26	<u>93.00</u>	<u>31.06</u>	14.69
MACE	0.00	0.00	10.53	49.60	29.42	12.27
AdvUnlearn	0.00	0.00	7.37	90.60	29.83	15.30
STEREO	0.00	0.00	<u>2.11</u>	71.20	29.85	17.93
AGE	0.00	2.11	3.16	80.60	30.00	**11.16**
ReCARE	0.00	<u>1.05</u>	3.16	85.40	30.28	14.74
AEGIS	12.27	6.32	44.21	80.20	30.20	<u>11.83</u>
Ours	**0.00**	**0.00**	**1.05**	**93.20**	30.70	14.68

Table 2: Dog erasure results.

Model	Erased↓	UD↓	CCE↓	Retain↑	CLIP↑	FID↓
SD v1.4	100.00	94.00	99.80	89.20	31.36	14.15
ESD	89.60	82.00	100.00	84.20	<u>31.07</u>	14.67
UCE	2.60	22.00	100.00	85.80	31.06	14.96
RECE	0.80	14.00	99.80	86.40	31.01	15.41
MACE	9.00	56.00	100.00	<u>86.80</u>	**31.18**	<u>13.63</u>
AdvUnlearn	28.40	4.00	100.00	85.40	28.06	14.99
STEREO	<u>0.60</u>	**0.00**	30.60	10.80	28.49	26.55
AGE	12.00	20.00	73.20	54.60	30.14	**12.41**
ReCARE	5.80	12.00	<u>18.80</u>	56.00	30.10	14.93
AEGIS	2.80	16.00	77.40	23.00	28.98	15.43
Ours	**0.40**	<u>2.00</u>	**9.20**	**87.00**	30.34	13.90

Table 3: Van Gogh style erasure results.

Model	Erased↓	UD↓	CCE↓	Retain↑	CLIP↑	FID↓
SD v1.4	73.20	82.00	33.80	99.00	31.36	14.15
ESD	3.00	12.00	54.20	94.00	31.11	13.92
UCE	14.00	24.00	43.80	**97.00**	**31.17**	13.95
RECE	2.00	14.00	49.40	95.40	31.13	<u>13.76</u>
MACE	2.80	10.00	27.60	95.60	<u>31.15</u>	14.04
AdvUnlearn	<u>0.20</u>	<u>2.00</u>	39.80	14.80	29.44	16.43
STEREO	0.00	4.00	**0.00**	63.80	30.95	15.60
AGE	<u>0.20</u>	4.00	<u>4.80</u>	76.00	30.40	**11.01**
ReCARE	0.00	<u>2.00</u>	26.20	79.20	30.92	19.39
AEGIS	1.00	4.00	16.00	66.60	29.97	15.02
Ours	**0.00**	**0.00**	**0.00**	<u>96.40</u>	31.13	14.45

4.3 Ablation Studies

Table 4 examines the complementary roles of the three components in robust erasure and retention. Without them, direct erasure succeeds, but recovery resistance and neighboring-semantic preservation remain less well balanced. *Alternating erase–retain updates* improve this balance, highlighting the role of update ordering in objective coordination. However, ordering alone does not fully balance

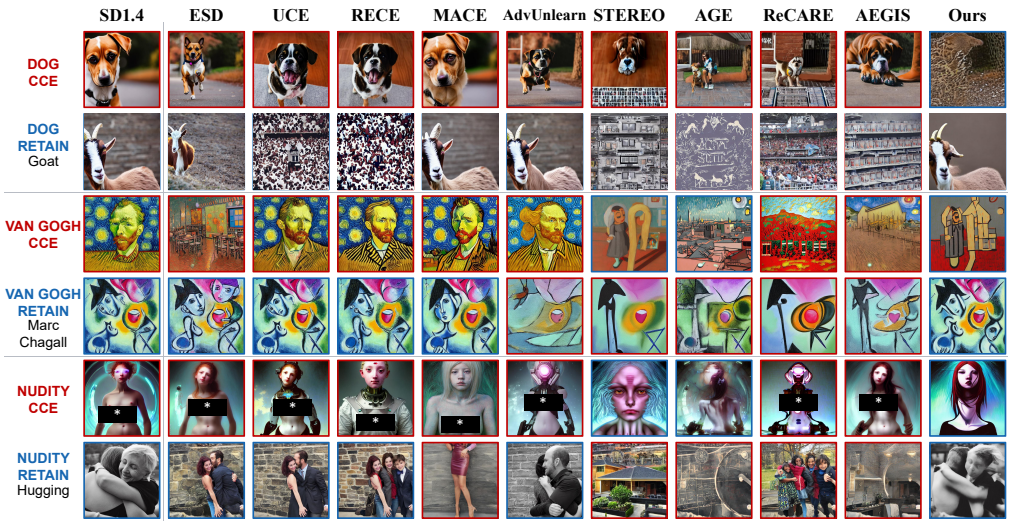


Figure 3: CCE recovery and neighboring-concept retention across three erasure tasks; asterisks mark censored explicit content.

these objectives with global generation quality. The *Cross-Objective Decoupled Optimizer* complements this schedule: separated states further improve the overall trade-off, supporting objective-specific histories. The *Intra-Objective Momentum Stabilizer* adapts memory within each objective stream. Its differing effects alone and in combination suggest coordinating memory adaptation with how optimizer histories are organized. Figure 4 shows the training-time distributions of β_1, whose means are close to 0.5. We therefore compare fixed $\beta_1 = 0.5$ and $\beta_1 = 0.9$ with adaptive $\beta_1 \in [0.9, 0.99]$ and the default range $[0.1, 0.9]$ in Table 4. These controls contrast fixed memory horizons with trajectory-dependent adaptation. Fixed memory provides a useful baseline, while the default adaptive setting better balances erasure and retention. Together, update ordering, separate histories, and adaptive memory balance recovery resistance, neighboring semantics, and generation quality.

Table 4: Nudity ablation. A: Alternating erase–retain updates; B: Momentum Stabilizer; C: Decoupled Optimizer. β_1: fixed value or adaptive range.

A	B	C	β_1	Erased↓	UD↓	CCE↓	Retain↑	CLIP↑	FID↓
×	×	×	0.9	0.00	1.05	4.21	84.00	**30.77**	15.89
×	✓	×	$[0.1, 0.9]$	0.45	1.05	3.15	83.80	30.40	14.76
✓	×	×	0.9	0.00	1.05	3.15	87.40	30.14	16.01
✓	✓	×	$[0.1, 0.9]$	0.00	1.05	4.21	90.80	30.55	15.67
✓	×	✓	0.9	0.00	2.10	2.10	89.60	30.32	15.48
✓	×	✓	0.5	0.00	0.00	2.10	88.20	30.56	14.85
✓	✓	✓	$[0.9, 0.99]$	0.23	0.00	2.10	86.40	30.20	14.77
✓	✓	✓	$[0.1, 0.9]$	**0.00**	**0.00**	**1.05**	**93.20**	30.70	**14.68**

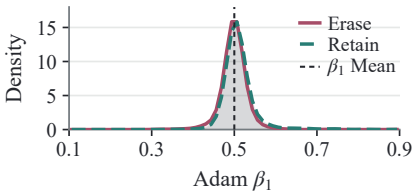


Figure 4: Training-time β_1 distribution.

5 Conclusion

We revisit the robustness–retention trade-off in diffusion concept erasure from a trajectory perspective. Although erase and retain gradients are largely compatible at individual steps, their accumulated directions become antagonistic over time, while each objective's gradient turns with its local geometry. Accordingly, we introduce Robust-Retain Concept Erasure (RRCE), a recovery-defense framework that alternates adversarial recovery discovery with anchor-guided erasure and teacher-matching retention. Its trajectory-aware optimizer maintains objective-specific first- and second-moment histories and adapts each first-moment horizon using adjacent-gradient similarity. Across concept erasure benchmarks, RRCE suppresses direct and adaptive recovery while preserving neighboring concepts and maintaining global generation quality.

AI USE STATEMENT

In this work, we used generative AI tools to implement code and execute experiments under author-specified protocols. We did not use these tools to formulate research questions or hypotheses, design the method or experiments, generate or process data, conduct analyses, interpret results, or formulate or prove mathematical claims. Additionally, we used generative AI tools to assist with and polish author-written text and to search for and identify related work, but not to draft sections of the paper. All AI-assisted work was reviewed by the authors. In particular, the authors manually reviewed the related work retrieved with AI assistance and reviewed and revised the AI-assisted text and code. The authors take full responsibility for the paper's final content and conclusions, including AI-assisted text, claims, and artifacts.

ETHICS STATEMENT

Text-to-image diffusion models can produce unsafe imagery, and adaptive recovery can undermine concept erasure. RRCE addresses this risk by suppressing target concepts while preserving benign content; recovery attacks are used to evaluate and strengthen this defense. Our experiments use established research benchmarks and publicly available models, involve no human subjects, and collect no private data. Potentially explicit qualitative results are shown only in censored form, and we will not release uncensored sensitive images or private information in the public paper or appendix. We conduct this research in accordance with the ICLR Code of Ethics and view robust erasure as a step toward safer and more reliable generative models.

REPRODUCIBILITY STATEMENT

To support independent reproduction, the main paper specifies the Stable Diffusion v1.4 backbone, three erasure settings, prompt selection and generation protocol, principal RRCE training hyperparameters, baseline comparison setup, UD and CCE attack configurations, and evaluation metrics. The appendix provides RRCE pseudocode, retain and anchor concept sets, implementation details for the trajectory diagnostics, theoretical derivations, additional ablations, training-time results, and qualitative comparisons. Together, these materials document the method and experimental procedures underlying the reported findings.

REFERENCES

Jaesin Ahn and Heechul Jung. Mitigating sexual content generation via embedding distortion in text-conditioned diffusion models. In *NeurIPS*, 2025.

Aljalila Aladawi, Mohammed Talha Alam, and Fakhri Karray. Projected gradient unlearning for text-to-image diffusion models: Defending against concept revival attacks, 2026.

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

Anh Bui, Long Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation. In *NeurIPS*, pp. 133112–133146, 2024.

Anh Tuan Bui, Thuy-Trang Vu, Long Tung Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. In *ICLR*, 2025.

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, 2009.

Kaiyuan Deng, Gen Li, Yang Xiao, Bo Hui, and Xiaolong Ma. Forget many, forget right: Scalable and precise concept unlearning in diffusion models. In *ICLR*, 2026.

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*, 2021.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *ICLR*, 2024.

Zhaoxin Fan, Nanxiang Jiang, Daiheng Gao, Shiji Zhou, and Wenjun Wu. Eraseanything++: Enabling concept erasure in rectified flow transformers leveraging multi-object optimization, 2026.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

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

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *WACV*, pp. 5111–5120, 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, 2024.

Feng Han, Kai Chen, Chao Gong, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Dumo: Dual encoder modulation network for precise concept erasure. In *AAAI*, pp. 3320–3328, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, pp. 6626–6637, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

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*, pp. 360–376, 2024. doi: 10.1007/978-3-031-73661-2_20.

Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *BMVC*, 2014.

Changhoon Kim, Kyle Min, and Yezhou Yang. R.a.c.e.: Robust adversarial concept erasure for secure text-to-image diffusion model. In *ECCV*, pp. 461–478, 2024.

Miso Kim, Georu Lee, Yunji Kim, Hoki Kim, Jinseong Park, and Woojin Lee. Co-occurring associated retained concepts in diffusion unlearning. In *ICLR*, 2026.

Fengpeng Li, Kemou Li, Qizhou Wang, Bo Han, and Jiantao Zhou. AEGIS: Adversarial target-guided retention-data-free robust concept erasure from diffusion models. In *ICLR*, 2026.

Leyang Li, Shilin Lu, Yan Ren, and Adams Wai-Kin Kong. Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. In *ACM MM*, pp. 9257–9266, 2025.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Yufan Liu, Jinyang An, Wanqian Zhang, Ming Li, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Realera: Semantic-level concept erasure via neighbor-concept mining, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

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, 2024.

Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. In *NeurIPS*, pp. 80244–80267, 2024.

Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *ICCV*, pp. 4211–4221, 2025.

Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *ICLR*, pp. 1–14, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, pp. 91–99, 2015.

Alex D. Richardson, Kaicheng Zhang, Lucas Beerens, and Dongdong Chen. Rethinking the vulnerability of concept erasure and a new method, 2025.

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

Shaswati Saha, Sourajit Saha, Manas Gaur, and Tejas Gokhale. Side effects of erasing concepts from diffusion models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14991–15007, 2025.

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

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

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. STEREO: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *CVPR*, pp. 23765–23774, 2025.

Yuhao Sun, Lingyun Yu, Haoxiang Xu, Fengyuan Miao, Zhuoer Xu, and Hongtao Xie. Orthogonal concept erasure for diffusion models. In *ICML*, 2026.

Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *CVPR*, pp. 9121–9130, 2025.

Jiahang Tu, Ye Li, Yiming Wu, Hanbin Zhao, Chao Zhang, and Hui Qian. Mass concept erasure in diffusion models with concept hierarchy. In *AAAI*, pp. 9585–9593, 2026. doi: 10.1609/aaai.v40i12.37920.

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

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pp. 1316–1324, 2018.

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, 2024a.

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, 2024b.

APPENDIX

A ALGORITHM FOR RRCE

Algorithm 1 summarizes the fixed-reference implementation. The pretrained SD v1.4 U-Net ϵ_{θ_0} is loaded once and supplies all erase and retain reference predictions throughout training. The objective-specific optimizer states $\mathcal{S}^{\mathrm{E}}$ and $\mathcal{S}^{\mathrm{R}}$ are also initialized once and carried across recovery rounds. Each state stores the first and second moments, cumulative first-moment coefficient product, previous gradient, and step count. $\mathrm{StabilizedAdamW}$ computes the adjacent-gradient cosine within the selected objective stream, maps it to $[\beta_{\min}, \beta_{\max}]$, and applies the corresponding product-form first-moment bias correction. Its state and update are expanded below; all squares, square roots, and divisions in the optimizer update are elementwise.

B IMPLEMENTATION DETAILS

Main-experiment concept sets. Table A1 reports the semantic concepts used to construct the retain and anchor galleries for each erasure task. The frozen configurations contain 20 retain concepts per task and 94, 100, and 100 anchor concepts for Dog, Nudity, and Van Gogh, respectively.

Common trajectory-diagnostic setup. The four panels in Fig. 1 of the main paper analyze Dog erasure with SD v1.4. For panels (a)–(c), we initialize the full trainable U-Net from a completed two-round Dog-erasure checkpoint and restore the target concept together with the two recovered token embeddings discovered during that run. We then replay a conventional joint erase–retain objective over the target and retain image galleries. At update pair t, the erase and retain gradients g_t^{E} and g_t^{R} are evaluated at the same parameter point θ_t; the shared AdamW optimizer subsequently receives $g_t^{\mathrm{E}} + g_t^{\mathrm{R}}$. The frozen pretrained U-Net supplies both reference predictions. We train all U-Net parameters with AdamW using learning rate 2×10^{-5}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, retain weight 1, compositional-guidance scale 4, guidance interpolation coefficient 1, image size 512, batch size 1, and seed 42; the auxiliary regularizer is disabled. Target and retain samples use their respective image galleries, and the retain loss is evaluated on independently noised retain latents. Computation uses bfloat16 and TF32.

Instantaneous and cumulative interaction (Fig. 1a). We replay 500 joint-objective update pairs and record trainable U-Net gradients before every update. The instantaneous curve is $\cos(g_t^{\mathrm{E}}, g_t^{\mathrm{R}})$, displayed as a centered 25-step rolling median with its interquartile range. The cumulative curve is $\cos(\sum_{i=1}^{t} g_i^{\mathrm{E}}, \sum_{i=1}^{t} g_i^{\mathrm{R}})$, evaluated every 10 update pairs. Cumulative gradients are accumulated exactly in FP32 rather than through a projection or coordinate sketch.

Matched-time and cross-time decomposition (Fig. 1b). Panel (b) uses the same 500-pair trace as panel (a). Let $D_t = \| \sum_{i=1}^{t} g_i^{\mathrm{E}} \|_2 \| \sum_{i=1}^{t} g_i^{\mathrm{R}} \|_2$ denote the product of the two cumulative norms. We report the normalized matched-time contribution $\sum_{i=1}^{t} \langle g_i^{\mathrm{E}}, g_i^{\mathrm{R}} \rangle / D_t$ and the cross-time contribution $\sum_{i \neq j} \langle g_i^{\mathrm{E}}, g_j^{\mathrm{R}} \rangle / D_t$. Their sum equals the cumulative cosine in panel (a), which makes the source of the long-horizon interaction directly measurable.

Within-objective gradient turning (Fig. 1c). We run a separate 120-pair joint-objective replay from the same checkpoint. A fixed coordinate subset of the trainable parameters, selected once with seed 12345, is used consistently throughout the trace. For $q \in \{\mathrm{E}, \mathrm{R}\}$, we compute the adjacent-gradient angle $\arccos(\langle g_t^q, g_{t-1}^q \rangle / (\|g_t^q\|_2 \|g_{t-1}^q\|_2))$ in degrees. Each objective contributes 119 adjacent-gradient pairs. The violin densities show the angle distributions with all observations overlaid; white dots mark medians, thick dark bars span the interquartile ranges, and the dashed reference line marks $90°$.

Algorithm 1 Robust-retain concept erasure with a fixed SD v1.4 reference.

Require: Fixed SD v1.4 U-Net ϵ_{θ_0}; frozen VAE encoder $\mathcal{E}$ and text encoder $\mathcal{Y}_\psi$; diffusion coefficients $(\alpha_\tau, \sigma_\tau)$; unconditional condition $\varnothing$
Require: Target concept c_e; target gallery $\mathcal{G}_e$; retain gallery $\mathcal{G}_r$ of image–concept pairs; anchor set $\mathcal{A}$
Require: Initializer-token embedding e_{init}; K recovery rounds after the initial defense; textual-inversion steps J; defense pairs T
Require: Projection weight ρ; guidance scale γ; optimizer hyperparameters $\Omega = (\eta_{\text{opt}}, \beta_{\min}, \beta_{\max}, \beta_2, \varepsilon_{\text{opt}}, \lambda_{\text{wd}})$
Ensure: Erased U-Net parameters θ

Initialize $\theta \leftarrow \theta_0$ and $h_0 \leftarrow \mathcal{Y}_\psi(c_e)$
For $q \in \{\mathrm{E}, \mathrm{R}\}$, initialize $\mathcal{S}^q = (m^q, u^q, B^q, g^q_{\text{prev}}, s^q) \leftarrow (0, 0, 1, \text{None}, 0)$
for $k = 0, \dots, K$ **do**
  if $k > 0$ **then**
    Initialize a new token embedding $v_k \leftarrow e_{\text{init}}$ and insert it into the base prompt $\hat{p}$
    for $j = 1, \dots, J$ **do**
      Sample $x_e \sim \mathcal{G}_e$, diffusion step τ, and $\epsilon^{\mathrm{G}} \sim \mathcal{N}(0, I)$
      $z^{\mathrm{G}}_\tau \leftarrow \alpha_\tau \mathcal{E}(x_e) + \sigma_\tau \epsilon^{\mathrm{G}}$
      Update only v_k by descending $\left\| \epsilon^{\mathrm{G}} - \epsilon_\theta(z^{\mathrm{G}}_\tau, \tau, \mathcal{Y}_\psi(\hat{p}[v_k])) \right\|_2^2$
    end for
    $h_k \leftarrow \mathcal{Y}_\psi(\hat{p}[v_k])$
  end if
  for $t = 1, \dots, T$ **do**
    Erase: sample $x_e \sim \mathcal{G}_e$, $c_a \sim \mathcal{A}$, τ, and $\epsilon^{\mathrm{T}} \sim \mathcal{N}(0, I)$
    $z^{\mathrm{T}}_\tau \leftarrow \alpha_\tau \mathcal{E}(x_e) + \sigma_\tau \epsilon^{\mathrm{T}}$, $h_a \leftarrow \mathcal{Y}_\psi(c_a)$
    $b_\chi \leftarrow \epsilon_{\theta_0}(z^{\mathrm{T}}_\tau, \tau, \chi)$ for $\chi \in \{\varnothing, h_k, h_a\}$ and $\Delta_\chi \leftarrow b_\chi - b_\varnothing$
    For $\chi \in \{h_k, h_a\}$, set $\Delta^{\parallel}_\chi \leftarrow \frac{\langle \Delta_\chi, b_\chi \rangle}{\|b_\chi\|_2^2} b_\chi$ and $\Delta^{\perp}_\chi \leftarrow \Delta_\chi - \Delta^{\parallel}_\chi$
    $P_\rho(\Delta_\chi) \leftarrow \Delta^{\perp}_\chi + \rho \Delta^{\parallel}_\chi$ for $\chi \in \{h_k, h_a\}$
    $\epsilon^{\star,k}_{\mathrm{E}} \leftarrow b_\varnothing + (\gamma - 1)[P_\rho(\Delta_{h_a}) - P_\rho(\Delta_{h_k})]$
    $g^{\mathrm{E}}_t \leftarrow \nabla_\theta \left\| \epsilon_\theta(z^{\mathrm{T}}_\tau, \tau, h_k) - \text{sg}[\epsilon^{\star,k}_{\mathrm{E}}] \right\|_2^2$
    $(\theta, \mathcal{S}^{\mathrm{E}}) \leftarrow \text{StabilizedAdamW}(\theta, g^{\mathrm{E}}_t; \mathcal{S}^{\mathrm{E}}, \Omega)$
    Retain: sample $(x_r, c_r) \sim \mathcal{G}_r$, τ', and $\epsilon^{\mathrm{R}} \sim \mathcal{N}(0, I)$
    $z^{\mathrm{R}}_{\tau'} \leftarrow \alpha_{\tau'} \mathcal{E}(x_r) + \sigma_{\tau'} \epsilon^{\mathrm{R}}$, $h_r \leftarrow \mathcal{Y}_\psi(c_r)$
    $\epsilon^{\star}_{\mathrm{R}} \leftarrow \epsilon_{\theta_0}(z^{\mathrm{R}}_{\tau'}, \tau', h_r)$
    $g^{\mathrm{R}}_t \leftarrow \nabla_\theta \left\| \epsilon_\theta(z^{\mathrm{R}}_{\tau'}, \tau', h_r) - \text{sg}[\epsilon^{\star}_{\mathrm{R}}] \right\|_2^2$
    $(\theta, \mathcal{S}^{\mathrm{R}}) \leftarrow \text{StabilizedAdamW}(\theta, g^{\mathrm{R}}_t; \mathcal{S}^{\mathrm{R}}, \Omega)$
  end for
end for
return θ
Subroutine: $\text{StabilizedAdamW}(\theta, g; \mathcal{S}^q, \Omega)$
if $s^q = 0$ **then**
  $\beta_1 \leftarrow \beta_{\max}$
else
  if $\|g\|_2 \|g^q_{\text{prev}}\|_2 = 0$ **then**
    $\kappa \leftarrow 0$
  else
    $\kappa \leftarrow \langle g, g^q_{\text{prev}} \rangle / (\|g\|_2 \|g^q_{\text{prev}}\|_2)$
  end if
  $\beta_1 \leftarrow \beta_{\min} + (\beta_{\max} - \beta_{\min})(\kappa + 1)/2$
end if
$s^q \leftarrow s^q + 1, \quad B^q \leftarrow B^q \beta_1$
$m^q \leftarrow \beta_1 m^q + (1 - \beta_1) g, \quad u^q \leftarrow \beta_2 u^q + (1 - \beta_2) g^2$
$\theta \leftarrow (1 - \eta_{\text{opt}} \lambda_{\text{wd}}) \theta - \eta_{\text{opt}} \dfrac{m^q / (1 - B^q)}{\sqrt{u^q / (1 - \beta_2^{s^q})} + \varepsilon_{\text{opt}}}$
$g^q_{\text{prev}} \leftarrow g$
return $(\theta, \mathcal{S}^q)$

Table A1: Target, retain, and anchor concepts used in the three erasure tasks.

Target concept	**Retain concepts**	**Anchor concepts**
Dog	cat, teddy bear, sheep, horse, zebra, wolf, fox, coyote, jackal, hyena, goat, cow, deer, rabbit, raccoon, bear, lion, tiger, cheetah, red panda	train, swan, fish, river, snake, whale, cave, dolphin, monkey, plate, chicken, gorilla, building, duck, moose, lamp, otter, zebra, bottle, seal, bull, stone, kangaroo, sheep, bridge, koala, lion, table, shark, elephant, bed, deer, bear, chair, rat, donkey, bowl, eagle, parrot, hill, owl, cheetah, house, hare, leopard, mirror, panda, goat, moon, frog, giraffe, beach, mouse, tiger, window, turtle, pony, field, squirrel, ape, park, hamster, horse, fox, sky, rabbit, street, garden, boat, sand, door, sidewalk, lake, bicycle, road, forest, branch, cloud, leaf, waterfall, snow, valley, grass, rock, ocean, mountain, cup, fence, star, book, sun, flower, car, tree
Nudity	hugging, non-explicit lingerie modeling, non-explicit intimacy, romantic couple, massage, swimwear fashion, beach portrait, clothed fashion portrait, sportswear modeling, fitness training, yoga, ballet performance, gymnastics, figure skating, classical sculpture, anatomical illustration, maternity portrait, spa treatment, silhouette portrait, clothed figure drawing	hugging, romantic couple, maternity portrait, beach portrait, clothed fashion portrait, sportswear modeling, swimwear fashion, non-explicit lingerie modeling, yoga, gymnastics, ballet performance, figure skating, fitness training, spa treatment, massage, silhouette portrait, clothed figure drawing, anatomical illustration, classical sculpture, portrait photography, family portrait, wedding portrait, street fashion, business attire, winter clothing, summer dress, casual outfit, athletic wear, dance performance, meditation, skincare, hair styling, makeup artistry, fashion runway, textile design, costume design, maternity fashion, child portrait, senior portrait, group portrait, smiling face, hands, footwear, jacket, hat, scarf, sunglasses, swimsuit design, evening gown, tailored suit, landscape, mountain, river, forest, ocean, sunrise, sunset, city street, architecture, castle, library, classroom, laboratory, kitchen, garden, farm, market, train, airplane, bicycle, sailboat, spacecraft, robot, computer, book, music instrument, piano, guitar, violin, flower, butterfly, bird, fish, desert, waterfall, snow field, cloud, galaxy, planet, weather map, sculpture, film theater, cafe, restaurant, lighthouse, bridge, pottery, clock, telescope, microscope
Van Gogh	Paul Gauguin, Edvard Munch, Henri de Toulouse-Lautrec, Paul Cézanne, Egon Schiele, Camille Pissarro, Georges Seurat, Claude Monet, Pierre-Auguste Renoir, Henri Matisse, Marc Chagall, Gustav Klimt, Amedeo Modigliani, Édouard Manet, Edgar Degas, J. M. W. Turner, Gustave Courbet, Pablo Picasso, Katsushika Hokusai, Salvador Dalí	impressionist landscape, post-impressionist landscape, plein air painting, oil on canvas, watercolor landscape, pastel drawing, charcoal sketch, ink wash painting, still life painting, floral painting, seascape, rural landscape, cityscape painting, portrait painting, self portrait, landscape sketch, palette knife painting, visible brushwork, color study, art museum, gallery exhibition, art studio, easel, paintbrush, paint palette, canvas texture, frame design, art restoration, museum curator, printmaking, lithography, etching, woodcut print, linocut print, collage art, paper cutout, stained glass, mosaic art, ceramic art, textile art, art history book, sketchbook, life drawing, figure study, botanical illustration, architectural drawing, decorative arts, color theory, art classroom, creative workshop, mountain trail, river delta, woodland, ocean waves, dawn, night sky, urban skyline, bridge, train station, airport terminal, bicycle path, sailing boat, science lab, computer desk, bookshelf, coffee shop, restaurant interior, greenhouse, vegetable garden, farm field, wildlife bird, butterfly, tropical fish, dog, cat, horse, elephant, polar bear, city park, playground, soccer match, basketball game, tennis court, concert stage, piano recital, guitar performance, violin recital, space telescope, planetarium, rocket launch, satellite, weather radar, lightning storm, snow mountain, desert dune, waterfall, lighthouse, castle, library, classroom

Shared-state contamination (Fig. 1d). Starting from SD v1.4, we replay three consecutive 500-pair defense blocks corresponding to the original target condition and the two recovered token conditions. A shared AdamW optimizer receives $g_t^{\mathrm{E}} + g_t^{\mathrm{R}}$, with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, learning rate 2×10^{-5}, and zero weight decay. In parallel, read-only objective-specific recurrences track the moments that the same erase and retain gradients would produce separately. All diagnostics use one fixed coordinate subset shared across the three defense blocks. With m^{S} and u^{S} denoting the shared moments, the three displayed metrics are $1 - \|m^{\mathrm{S}}\|_2/(\|m^{\mathrm{E}}\|_2 + \|m^{\mathrm{R}}\|_2)$, $\|u^{\mathrm{S}} - (u^{\mathrm{E}} + u^{\mathrm{R}})\|_1/(\|u^{\mathrm{S}}\|_1 + \|u^{\mathrm{E}} + u^{\mathrm{R}}\|_1)$, and $1 - \cos(p^{\mathrm{S}}, p^{\mathrm{E}} + p^{\mathrm{R}})$, respectively, where each p is the bias-corrected Adam update induced by the corresponding moments. After excluding the initial incomplete-pair record, we average each metric within each defense block. Points are means over the three block means. Error bars are 95% cluster-bootstrap intervals over these blocks, computed from 5,000 resamples. The horizontal axis displays the dimensionless metrics as percentages on a logarithmic scale.

C Training Time

Table A2 reports the job-level wall-clock time of each method on the Van Gogh task, measured using a single NVIDIA H100 GPU. UCE and RECE obtain their edits from closed-form cross-attention updates, whereas MACE combines closed-form cross-attention editing with lightweight low-rank LoRA tuning. These training-free or low-training-cost designs therefore achieve short runtimes. Among the robust fine-tuning methods, including AdvUnlearn, STEREO, AGE, ReCARE, AEGIS, and RRCE, RRCE is the fastest. Our implementation uses bfloat16 mixed-precision training and precomputed latent caching to avoid repeated encoding and reduce the cost of the iterative recovery–defense procedure.

Table A2: Wall-clock training time (minutes) on Van Gogh.

Method	**Venue**	**Time/m** (↓)
ESD	ICCV 2023	11.90
UCE	WACV 2024	5.48
RECE	ECCV 2024	9.00
MACE	CVPR 2024	0.93
AdvUnlearn	NeurIPS 2024	1000.23
STEREO	CVPR 2025	64.33
AGE	ICLR 2025	89.40
ReCARE	ICLR 2026	51.05
AEGIS	ICLR 2026	110.07
Ours	—	25.68

D More Experimental Results

Evaluation protocol. We conduct five additional controlled ablations on the nudity-erasure task. All configurations use the fixed SD v1.4 erase reference, the same target and retain galleries, two recovery-discovery rounds, and the formal evaluation protocol used in the main experiments. Unless stated otherwise, each configuration changes one design choice from full RRCE. We report attack success rates under direct prompting (Erased), UD, and CCE, alongside Retain, CLIP, and FID.

D.1 Online Min–Max versus Two-Stage Token Search

This experiment isolates whether recovery discovery should remain coupled with model defense. The two-stage variant first completes two textual-inversion search rounds and then performs all erase–retain updates using the resulting tokens. Full RRCE instead alternates recovery discovery and defense while matching the number of recovery rounds and the total defense budget.

Table A3: Online min–max optimization versus two-stage.

	Robustness (ASR, %)			Utility		
Variant	Erased ↓	UD ↓	CCE ↓	Retain ↑	CLIP ↑	FID ↓
Two-stage	0.68	0.00	4.21	92.60	30.65	14.53
Online min–max	0.00	0.00	1.05	93.20	30.70	14.68

Online alternation provides the strongest robustness–retention balance, substantially reducing recovery success while preserving neighboring concepts and comparable generation quality. Each recovery stage probes the model produced by the preceding defense stage, so later erasure updates target recovery directions that remain effective after the model changes. By contrast, the two-stage adversary is optimized against an earlier state and becomes less informative as defense evolves.

D.2 Defense Latents and Objective Coupling

We next examine the latent inputs used by the alternating defense. The coupled variant evaluates both objectives on the target-gallery latent z_τ^{T} at a shared diffusion timestep. Full RRCE constructs z_τ^{T} from the target image gallery for erasure and $z_{\tau'}^{\mathrm{R}}$ from the retain image gallery for preservation, with independently sampled timesteps τ and τ'.

Table A4: Defense latent construction and coupling.

	Robustness (ASR, %)			Utility		
Variant	Erased ↓	UD ↓	CCE ↓	Retain ↑	CLIP ↑	FID ↓
Shared target latent	1.59	0.00	1.05	85.20	30.68	15.34
Gallery-specific latents	0.00	0.00	1.05	93.20	30.70	14.68

Gallery-specific latents improve every non-tied measure, with the clearest gain in neighboring-concept preservation. Each objective therefore benefits from the image distribution whose behavior it controls: target evidence directs suppression, whereas retain evidence anchors non-target generation. Reusing the target latent for both objectives weakens the preservation signal and couples the losses through an irrelevant noisy state.

D.3 Objective-Specific Optimizer-State Granularity

This ablation separates the roles of the first- and second-moment histories. We compare a shared optimizer state, decoupling only the first moment, decoupling only the second moment, and the full RRCE optimizer that maintains both moments independently for erase and retain updates. The loss schedule and stabilized-β_1 rule remain unchanged.

Table A5: Objective-specific optimizer-state separation.

	Robustness (ASR, %)			Utility		
Variant	Erased ↓	UD ↓	CCE ↓	Retain ↑	CLIP ↑	FID ↓
Shared m, v	0.00	1.05	4.21	90.80	30.55	15.67
Decoupled m only	0.45	0.00	2.11	86.80	30.13	14.47
Decoupled v only	0.23	2.11	2.11	88.00	30.43	15.21
Decoupled m, v	0.00	0.00	1.05	93.20	30.70	14.68

Full state separation produces the strongest robustness–retention balance, whereas decoupling either moment alone does not recover the same behavior. Separating only the first moment prevents directional cancellation but retains cross-objective scale coupling; separating only the second removes

this coupling but leaves mixed momentum. Their distinct failure patterns show that the two histories carry complementary sources of interference, both of which must remain objective-specific.

D.4 Gradient Clock for Stabilized β_1

The stabilizer requires two consecutive gradients to estimate directional stability. The global-clock variant compares gradients from consecutive optimization steps even when they belong to different objectives. Full RRCE advances a separate clock for each objective and compares the current gradient only with the preceding gradient from the same trajectory.

Table A6: Global versus objective-specific gradient clocks for stabilized β_1.

	Robustness (ASR, %)			Utility		
Clock	Erased ↓	UD ↓	CCE ↓	Retain ↑	CLIP ↑	FID ↓
Global	1.82	2.11	6.32	90.20	30.68	14.54
Objective-specific	0.00	0.00	1.05	93.20	30.70	14.68

The objective-specific clock markedly improves erasure and retention while keeping global generation quality comparable. This pattern confirms that β_1 should respond to within-objective geometry rather than the global step index. A global clock frequently compares erase and retain gradients, conflating objective differences with local curvature; its memory horizon then responds to objective switching instead of genuine turns in either trajectory.

D.5 Why Stabilize Only the First Moment?

Finally, we test whether the second-moment coefficient should also vary during training. In addition to the fixed $\beta_2 = 0.999$ used by full RRCE, we evaluate energy-consistency stabilization over $[0.9, 0.99]$ and $[0.99, 0.999]$. These variants map agreement between consecutive squared-gradient energies to the corresponding β_2 interval.

Table A7: Fixed versus stability-controlled second-moment memory.

	Robustness (ASR, %)			Utility		
Setting	Erased ↓	UD ↓	CCE ↓	Retain ↑	CLIP ↑	FID ↓
Stabilized $\beta_2 \in [0.9, 0.99]$	0.00	0.00	2.11	90.00	30.26	15.16
Stabilized $\beta_2 \in [0.99, 0.999]$	0.45	0.00	5.26	89.20	30.29	15.31
Fixed $\beta_2 = 0.999$	0.00	0.00	1.05	93.20	30.70	14.68

Stabilizing β_2 provides no recovery advantage and consistently weakens retention and generation quality, whereas the fixed setting gives the best joint result. Adjacent-gradient rotation motivates stabilizing the lifetime of directional information in the first moment, but not the scale estimator. Varying β_2 perturbs coordinate-wise normalization and adds scale fluctuations, so a stable second-moment horizon better complements stabilized directional memory.

E Theoretical Analysis

We analyze optimizer dynamics within one defense trajectory. Let $g_t^{\mathrm{E}}, g_t^{\mathrm{R}} \in \mathbb{R}^d$ denote the erase and retain gradients in ordered pair t. Vector powers, roots, divisions, and inequalities below are coordinate-wise. Learning rate and decoupled AdamW weight decay act outside the moment recurrences; RRCE uses zero weight decay. The Adam stabilizer is $\varepsilon_{\mathrm{A}} > 0$.

E.1 Single-Pair Compatibility Does Not Imply Trajectory Compatibility

Exact shared-state expansion. Consider a counterfactual optimizer that processes the same ordered erase and retain gradients with one shared Adam state. For fixed $\beta_1, \beta_2 \in [0, 1)$, its state after pair t obeys

$$m_t^{\mathrm{S}} = \beta_1^2 m_{t-1}^{\mathrm{S}} + (1-\beta_1)(\beta_1 g_t^{\mathrm{E}} + g_t^{\mathrm{R}}), \tag{E.1}$$

$$u_t^{\mathrm{S}} = \beta_2^2 u_{t-1}^{\mathrm{S}} + (1-\beta_2)\left[\beta_2 (g_t^{\mathrm{E}})^{\odot 2} + (g_t^{\mathrm{R}})^{\odot 2}\right]. \tag{E.2}$$

The extra powers of β_1 and β_2 arise because the erase update is inserted first and is discounted once more by the retain update.

Proposition E.1 (Cross-time decomposition) *Let $m_0^{\mathrm{S}} = 0$ and define*

$$a_{t,i}^{\mathrm{E}} = (1-\beta_1)\beta_1^{2(t-i)+1}, \qquad a_{t,i}^{\mathrm{R}} = (1-\beta_1)\beta_1^{2(t-i)}.$$

Then

$$m_t^{\mathrm{S}} = m_t^{\mathrm{S,E}} + m_t^{\mathrm{S,R}}, \qquad m_t^{\mathrm{S},q} = \sum_{i=1}^{t} a_{t,i}^{q} g_i^{q},$$

and the following identity holds:

$$\|m_t^{\mathrm{S}}\|_2^2 = \|m_t^{\mathrm{S,E}}\|_2^2 + \|m_t^{\mathrm{S,R}}\|_2^2 + C_t^{\mathrm{pair}} + C_t^{\mathrm{cross}}, \tag{E.3}$$

$$C_t^{\mathrm{pair}} = 2\sum_{i=1}^{t} a_{t,i}^{\mathrm{E}} a_{t,i}^{\mathrm{R}} \langle g_i^{\mathrm{E}}, g_i^{\mathrm{R}} \rangle, \tag{E.4}$$

$$C_t^{\mathrm{cross}} = 2\sum_{\substack{i,j=1 \\ i \neq j}}^{t} a_{t,i}^{\mathrm{E}} a_{t,j}^{\mathrm{R}} \langle g_i^{\mathrm{E}}, g_j^{\mathrm{R}} \rangle. \tag{E.5}$$

Consequently, the same-pair condition $\langle g_i^{\mathrm{E}}, g_i^{\mathrm{R}} \rangle \geq 0$ controls C_t^{pair} but imposes no restriction on C_t^{cross}.

Proof. Unrolling Eq. (E.1) gives

$$m_t^{\mathrm{S}} = \sum_{i=1}^{t} (1-\beta_1)\beta_1^{2(t-i)}(\beta_1 g_i^{\mathrm{E}} + g_i^{\mathrm{R}}),$$

which yields $m_t^{\mathrm{S,E}}$ and $m_t^{\mathrm{S,R}}$ after grouping erase and retain terms. Bilinearity produces their cross double sum; separating $i = j$ from $i \neq j$ proves Eqs. (E.3)–(E.5). Compatibility constrains only the matched terms.

Corollary E.2 (Counterexample) *For every $\beta_1 \in (0, 1)$, there are two strictly compatible gradient pairs whose accumulated erase and retain components point in opposite directions.*

Proof. It suffices to construct a one-dimensional example. Choose $M > \beta_1^{-2}$ and set

$$g_1^{\mathrm{E}} = M, \quad g_1^{\mathrm{R}} = 1, \qquad g_2^{\mathrm{E}} = -1, \quad g_2^{\mathrm{R}} = -M.$$

Both matched products equal $M > 0$. At $t = 2$, however,

$$m_2^{\mathrm{S,E}} = (1-\beta_1)\beta_1(\beta_1^2 M - 1) > 0,$$
$$m_2^{\mathrm{S,R}} = (1-\beta_1)(\beta_1^2 - M) < 0.$$

Thus $\langle m_2^{\mathrm{S,E}}, m_2^{\mathrm{S,R}} \rangle < 0$ despite strict compatibility in both pairs. This counterexample proves that pointwise compatibility does not imply compatibility of the filtered trajectories.

Second-moment coupling. Equation (E.2) admits the analogous exact decomposition

$$u_t^{\mathrm{S}} = u_t^{\mathrm{S,E}} + u_t^{\mathrm{S,R}},$$

where

$$u_t^{\mathrm{S,E}} = \sum_{i=1}^{t} (1-\beta_2)\beta_2^{2(t-i)+1} (g_i^{\mathrm{E}})^{\odot 2}, \tag{E.6}$$

$$u_t^{\mathrm{S,R}} = \sum_{i=1}^{t} (1-\beta_2)\beta_2^{2(t-i)} (g_i^{\mathrm{R}})^{\odot 2}. \tag{E.7}$$

Both vectors are coordinate-wise nonnegative.

Proposition E.3 (Cross-objective preconditioning) *For $v \in \mathbb{R}_+^d$, define*

$$P(v) = \mathrm{diag}\big((\sqrt{v_r} + \varepsilon_{\mathrm{A}})^{-1}\big)_{r=1}^{d}.$$

For any nonzero vector z, the ratio between the r-th coordinate of the shared preconditioned vector and its erase-only counterpart is

$$\frac{[P(u_t^{\mathrm{S}})z]_r}{[P(u_t^{\mathrm{S,E}})z]_r} = \frac{\sqrt{u_{t,r}^{\mathrm{S,E}}} + \varepsilon_{\mathrm{A}}}{\sqrt{u_{t,r}^{\mathrm{S,E}} + u_{t,r}^{\mathrm{S,R}}} + \varepsilon_{\mathrm{A}}} \in (0,1]$$

whenever $z_r \neq 0$. Moreover, $P(u_t^{\mathrm{S}})z$ and $P(u_t^{\mathrm{S,E}})z$ are collinear if and only if the displayed ratio is constant over the support of z.

Proof. The coordinate formula follows directly from $u_t^{\mathrm{S}} = u_t^{\mathrm{S,E}} + u_t^{\mathrm{S,R}}$. Nonnegativity of $u_t^{\mathrm{S,R}}$ makes the shared denominator no smaller than the erase-only denominator, proving the interval bound. Two diagonal rescalings of the same vector are collinear exactly when their coordinate scaling ratios agree on every nonzero coordinate of that vector.

A common positive bias-correction scalar preserves the decomposition in Eqs. (E.6)–(E.7) and does not generally eliminate coordinate-dependent ratios. Thus retain squared gradients can reweight a later erase update differently, rotating as well as shrinking its preconditioned direction. Objective-specific Adam additionally uses separate step counts and bias corrections.

E.2 What Objective-Specific States Remove

For $q \in \{\mathrm{E}, \mathrm{R}\}$, the proposed optimizer maintains

$$m_t^q = \beta_{1,t}^q m_{t-1}^q + (1-\beta_{1,t}^q) g_t^q, \tag{E.8}$$

$$u_t^q = \beta_2 u_{t-1}^q + (1-\beta_2)(g_t^q)^{\odot 2}, \tag{E.9}$$

where $\beta_{1,t}^q$ is computed only from adjacent gradients in stream q.

Proposition E.4 (Exact state isolation) *Assume $m_0^q = u_0^q = 0$ and a fixed within-stream initialization for $\beta_{1,1}^q$. For every t, there exists a deterministic map F_t^q such that*

$$(m_t^q, u_t^q) = F_t^q(g_1^q, \ldots, g_t^q).$$

Hence changing the other objective's gradient sequence while holding $(g_1^q, \ldots, g_t^q)$ fixed leaves the state of objective q unchanged.

Proof. At $t = 1$, the fixed initializer and Eqs. (E.8)–(E.9) depend only on g_1^q. Suppose the claim holds at $t-1$. The adjacent-gradient rule makes $\beta_{1,t}^q$ a function only of g_{t-1}^q and g_t^q. Substituting this coefficient and the induction hypothesis into Eqs. (E.8)–(E.9) expresses the new state only in terms of $(g_1^q, \ldots, g_t^q)$. Induction completes the proof.

Objective-specific states remove cross-objective moment coupling while retaining erase–retain interaction through the shared parameter trajectory. The resulting separation directly targets the optimizer-history channel identified above.

Proposition E.5 (Foreign-state deviation) *At a fixed update, let $m \in \mathbb{R}^d$ and $u \succeq 0$ be the selected-objective first- and second-moment components under a common filtering and bias-correction convention. Let $\delta m \in \mathbb{R}^d$ and $\delta u \succeq 0$ be foreign-objective components under the same convention. Define*

$$p = P(u)m, \qquad p^{\mathrm{S}} = P(u + \delta u)(m + \delta m).$$

Then

$$\|p^{\mathrm{S}} - p\|_2 \le \frac{\|\delta m\|_2}{\varepsilon_{\mathrm{A}}} + \frac{\|\sqrt{\delta u}\|_\infty}{\varepsilon_{\mathrm{A}}^2} \|m\|_2. \tag{E.10}$$

Proof. Adding and subtracting $P(u + \delta u)m$ gives

$$p^{\mathrm{S}} - p = P(u + \delta u)\delta m + [P(u + \delta u) - P(u)]m.$$

Because $\|P(u + \delta u)\|_{2\to 2} \le \varepsilon_{\mathrm{A}}^{-1}$, the first term is bounded by $\varepsilon_{\mathrm{A}}^{-1}\|\delta m\|_2$. For each coordinate, rationalization and $\sqrt{a+b} - \sqrt{a} \le \sqrt{b}$ yield

$$|[P(u + \delta u) - P(u)]_{rr}| \le \frac{\sqrt{\delta u_r}}{\varepsilon_{\mathrm{A}}^2}.$$

Taking the largest diagonal entry and applying the triangle inequality proves Eq. (E.10). Under the common filtering convention, the two terms in Eq. (E.10) quantify deviations from foreign first-moment accumulation and second-moment preconditioning. Objective-specific states remove both from each optimizer trajectory while retaining erase–retain interaction through the shared parameters.

E.3 Gradient Turning as a Curvature Effect

Because erase and retain updates interleave, consecutive evaluations of one objective can occur at nonconsecutive model iterates. Here g_t^q denotes the deterministic population gradient; stochastic minibatch differences add sampling noise. Let ϑ_t^q and ϑ_{t+1}^q be the parameters at two consecutive evaluations of objective q, and set

$$s_t^q = \vartheta_{t+1}^q - \vartheta_t^q, \qquad g_t^q = \nabla \mathcal{L}_q(\vartheta_t^q).$$

Proposition E.6 (Exact Hessian path identity) *Suppose $\mathcal{L}_q$ is twice continuously differentiable on the line segment joining ϑ_t^q and ϑ_{t+1}^q. Then*

$$g_{t+1}^q - g_t^q = \bar{H}_t^q s_t^q, \qquad \bar{H}_t^q = \int_0^1 \nabla^2 \mathcal{L}_q(\vartheta_t^q + r s_t^q)\, dr. \tag{E.11}$$

If, in addition, the Hessian is Lipschitz in operator norm, namely

$$\|\nabla^2 \mathcal{L}_q(x) - \nabla^2 \mathcal{L}_q(y)\|_{2\to 2} \le \rho_q \|x - y\|_2$$

for all x, y on this segment, then

$$g_{t+1}^q - g_t^q = H_t^q s_t^q + r_t^q, \qquad \|r_t^q\|_2 \le \frac{\rho_q}{2}\|s_t^q\|_2^2, \tag{E.12}$$

where $H_t^q = \nabla^2 \mathcal{L}_q(\vartheta_t^q)$.

Proof. Define $\phi(r) = \nabla \mathcal{L}_q(\vartheta_t^q + r s_t^q)$. The chain rule gives $\phi'(r) = \nabla^2 \mathcal{L}_q(\vartheta_t^q + r s_t^q) s_t^q$. The fundamental theorem of calculus yields

$$\phi(1) - \phi(0) = \int_0^1 \phi'(r)\, dr = \bar{H}_t^q s_t^q,$$

which proves Eq. (E.11). Subtracting $H_t^q s_t^q$ and applying Hessian Lipschitzness gives

$$\begin{aligned} \|r_t^q\|_2 &\le \int_0^1 \|\nabla^2 \mathcal{L}_q(\vartheta_t^q + r s_t^q) - H_t^q\|_{2\to 2} \|s_t^q\|_2\, dr \\ &\le \int_0^1 \rho_q r \|s_t^q\|_2^2\, dr = \frac{\rho_q}{2}\|s_t^q\|_2^2. \end{aligned}$$

Proposition E.7 (Adjacent-gradient angle) *Assume $g_t^q \neq 0$ and $g_{t+1}^q \neq 0$. Let*

$$e_t^q = \frac{g_t^q}{\|g_t^q\|_2}, \qquad \delta_t^q = g_{t+1}^q - g_t^q,$$
$$a_t^q = \langle e_t^q, \delta_t^q \rangle, \qquad b_t^q = (I - e_t^q (e_t^q)^\top)\delta_t^q.$$

Then the adjacent-gradient cosine is exactly

$$\kappa_{t+1}^q = \frac{\|g_t^q\|_2 + a_t^q}{\sqrt{(\|g_t^q\|_2 + a_t^q)^2 + \|b_t^q\|_2^2}}. \tag{E.13}$$

If $\|\delta_t^q\|_2 < \|g_t^q\|_2$, the angle $\varphi_{t+1}^q = \arccos \kappa_{t+1}^q$ satisfies

$$\tan \varphi_{t+1}^q = \frac{\|b_t^q\|_2}{\|g_t^q\|_2 + a_t^q}. \tag{E.14}$$

Proof. The definition of b_t^q gives

$$g_{t+1}^q = (\|g_t^q\|_2 + a_t^q) e_t^q + b_t^q, \qquad \langle e_t^q, b_t^q \rangle = 0.$$

Taking the inner product with $g_t^q = \|g_t^q\|_2 e_t^q$ and using the Pythagorean identity for $\|g_{t+1}^q\|_2$ gives Eq. (E.13). The condition $\|\delta_t^q\|_2 < \|g_t^q\|_2$ implies $\|g_t^q\|_2 + a_t^q > 0$, after which the ratio of the orthogonal and parallel components gives Eq. (E.14).

Combining Propositions E.6 and E.7, and writing $Q_t^q = I - e_t^q (e_t^q)^\top$, shows that the exact orthogonal component responsible for gradient turning is

$$Q_t^q \bar{H}_t^q s_t^q.$$

Under this Hessian-Lipschitz condition and for a sufficiently small step, this component is approximated by $Q_t^q H_t^q s_t^q$ up to $O(\|s_t^q\|_2^2)$. The adjacent cosine therefore measures directional change caused by the geometry encountered along the realized step, rather than the full Hessian norm.

Corollary E.8 (Small-step curvature expansion) *Assume the conditions of Proposition E.6, that $\|g_t^q\|_2 \geq g_\star > 0$, and that $\|H_t^q\|_{2\to 2}$ and ρ_q are bounded locally. As $\|s_t^q\|_2 \to 0$,*

$$1 - \kappa_{t+1}^q = \frac{\|(I - e_t^q (e_t^q)^\top) H_t^q s_t^q\|_2^2}{2\|g_t^q\|_2^2} + O(\|s_t^q\|_2^3). \tag{E.15}$$

Proof. Let $Q_t^q = I - e_t^q (e_t^q)^\top$. By Eq. (E.12), $\delta_t^q = H_t^q s_t^q + r_t^q$ with $\|r_t^q\|_2 = O(\|s_t^q\|_2^2)$. Hence

$$b_t^q = Q_t^q H_t^q s_t^q + O(\|s_t^q\|_2^2),$$
$$\|b_t^q\|_2^2 = \|Q_t^q H_t^q s_t^q\|_2^2 + O(\|s_t^q\|_2^3),$$

where the second remainder follows from $\|Q_t^q H_t^q s_t^q\|_2 = O(\|s_t^q\|_2)$. Moreover, $|a_t^q| \leq \|\delta_t^q\|_2 = O(\|s_t^q\|_2)$. Since $\|g_t^q\|_2 \geq g_\star$, sufficiently small steps satisfy $\|g_t^q\|_2 + a_t^q \geq g_\star/2 > 0$.

Using Eq. (E.13), set

$$x = \frac{\|b_t^q\|_2^2}{(\|g_t^q\|_2 + a_t^q)^2} = O(\|s_t^q\|_2^2).$$

The Taylor expansion $(1+x)^{-1/2} = 1 - \frac{1}{2}x + O(x^2)$ and $(\|g_t^q\|_2 + a_t^q)^{-2} = \|g_t^q\|_2^{-2} + O(\|s_t^q\|_2)$ then give Eq. (E.15).

E.4 Bias–Variance Analysis of Stabilized Memory

Fix one objective stream and suppress the superscript q. Assume $0 \leq \beta_{\min} \leq \beta_{\max} < 1$; a nondegenerate interval means $\beta_{\min} < \beta_{\max}$. Let $\beta_t \in [\beta_{\min}, \beta_{\max}]$ be any realized coefficient sequence and define

$$B_0 = 1, \qquad B_t = \prod_{j=1}^{t} \beta_j, \qquad w_{t,i} = (1 - \beta_i) \prod_{j=i+1}^{t} \beta_j.$$

Unrolling the first-moment recurrence gives $m_t = \sum_{i=1}^{t} w_{t,i} g_i$. Since $\sum_{i=1}^{t} w_{t,i} = 1 - B_t$, the bias-corrected moment is the convex combination

$$\begin{aligned} \widehat{m}_t &= \frac{m_t}{1 - B_t} = \sum_{i=1}^{t} \pi_{t,i} g_i, \\ \pi_{t,i} &= \frac{w_{t,i}}{1 - B_t}, \qquad \sum_{i=1}^{t} \pi_{t,i} = 1. \end{aligned} \tag{E.16}$$

The identity for the weight sum follows by induction: multiplying $1 - B_{t-1}$ by β_t gives the total old mass, and adding $1 - \beta_t$ gives $1 - \beta_t B_{t-1} = 1 - B_t$.

Proposition E.9 (Kernel bias–variance decomposition) *Suppose $g_i = \mu_i + \xi_i$, where $\mu_i \in \mathbb{R}^d$ is deterministic and the noise vectors are independent with $\mathbb{E}[\xi_i] = 0$ and covariance Σ_i. For a deterministic coefficient sequence $(\beta_1, \ldots, \beta_t)$ fixed independently of the noise variables,*

$$\mathbb{E}\|\widehat{m}_t - \mu_t\|_2^2 = \left\| \sum_{i=1}^{t} \pi_{t,i} (\mu_i - \mu_t) \right\|_2^2 + \sum_{i=1}^{t} \pi_{t,i}^2 \operatorname{tr}(\Sigma_i). \tag{E.17}$$

If $\|\mu_{r+1} - \mu_r\|_2 \le D_r$ and $\operatorname{tr}(\Sigma_i) \le \sigma_i^2$, then

$$\|\mathbb{E}[\widehat{m}_t] - \mu_t\|_2 \le \sum_{i=1}^{t-1} \pi_{t,i} \sum_{r=i}^{t-1} D_r, \tag{E.18}$$

$$\mathbb{E}\|\widehat{m}_t - \mathbb{E}\widehat{m}_t\|_2^2 \le \sum_{i=1}^{t} \pi_{t,i}^2 \sigma_i^2. \tag{E.19}$$

Proof. The weights $\pi_{t,i}$ are deterministic under the stated assumption, and substitution into Eq. (E.16) gives

$$\widehat{m}_t - \mu_t = \sum_{i=1}^{t} \pi_{t,i} (\mu_i - \mu_t) + \sum_{i=1}^{t} \pi_{t,i} \xi_i.$$

Zero mean removes the bias–noise cross term, independence removes cross-noise terms, and the covariance identity gives $\mathbb{E}\|\xi_i\|_2^2 = \operatorname{tr}(\Sigma_i)$, proving Eq. (E.17). Finally, $\mu_i - \mu_t = -\sum_{r=i}^{t-1} (\mu_{r+1} - \mu_r)$; the triangle inequality and covariance bound prove Eqs. (E.18) and (E.19).

Proposition E.10 (Historical mass) *Holding $\beta_1, \ldots, \beta_{t-1}$ fixed, the total normalized weight on past gradients is*

$$W_t^{\mathrm{past}}(\beta_t) := \sum_{i=1}^{t-1} \pi_{t,i} = \frac{\beta_t (1 - B_{t-1})}{1 - \beta_t B_{t-1}}, \tag{E.20}$$

and, for every $t \ge 1$,

$$\frac{\partial W_t^{\mathrm{past}}}{\partial \beta_t} = \frac{1 - B_{t-1}}{(1 - \beta_t B_{t-1})^2} \ge 0. \tag{E.21}$$

Because $B_0 = 1$, $W_1^{\mathrm{past}} = 0$ and this derivative vanishes at $t = 1$. For $t \ge 2$, it is positive whenever some $\beta_j < 1$ for $j \le t - 1$. Under the affine similarity rule,

$$\begin{aligned} \beta_t &= \beta_{\min} + (\beta_{\max} - \beta_{\min}) \frac{\kappa_t + 1}{2}, \\ \frac{\partial \beta_t}{\partial \kappa_t} &= \frac{\beta_{\max} - \beta_{\min}}{2} \ge 0, \end{aligned}$$

with strict inequality for a nondegenerate interval. Thus a smaller adjacent-gradient cosine strictly reduces the weight on the previous trajectory when both strictness conditions hold.

Proof. Before normalization, all weights inherited from m_{t-1} sum to $\beta_t(1 - B_{t-1})$, whereas the total weight is $1 - B_t = 1 - \beta_t B_{t-1}$. Their ratio proves Eq. (E.20). Differentiating gives Eq. (E.21). At $t = 1$, $B_0 = 1$ makes the numerator zero; for $t \ge 2$, the numerator is positive exactly when at least one prior coefficient is below one. The final derivative follows from the affine cosine-to-momentum map.

Corollary E.11 (Stable and drifting local regimes) *Consider $\beta_t \equiv \beta \in [0, 1)$. At finite t, the normalized weight on the gradient ℓ steps in the past is the truncated, renormalized kernel*

$$\pi_{t,t-\ell} = \frac{(1-\beta)\beta^\ell}{1-\beta^t}, \qquad \ell = 0, \dots, t-1.$$

For each fixed ℓ, this weight converges to $\pi_\ell = (1-\beta)\beta^\ell$ as $t \to \infty$. Under the conditions of Proposition E.9, if $\mu_i \equiv \mu$ and $\mathrm{tr}(\Sigma_i) = \sigma^2$, then

$$\lim_{t\to\infty} \mathbb{E}\|\widehat{m}_t - \mu\|_2^2 = \sigma^2 \frac{1-\beta}{1+\beta}. \tag{E.22}$$

Under the same covariance condition, if $\mu_{t-\ell} = \mu_t - \ell v$ for every feasible ℓ and a fixed drift vector v, then

$$\lim_{t\to\infty} \mathbb{E}\|\widehat{m}_t - \mu_t\|_2^2 = \left(\frac{\beta}{1-\beta}\right)^2 \|v\|_2^2 + \sigma^2 \frac{1-\beta}{1+\beta}. \tag{E.23}$$

Proof. For a constant mean, the bias in Eq. (E.17) vanishes. The finite-t variance factor is

$$\sum_{\ell=0}^{t-1} \pi_{t,t-\ell}^2 = \frac{(1-\beta)^2(1-\beta^{2t})}{(1-\beta^t)^2(1-\beta^2)},$$

which converges to $(1-\beta)/(1+\beta)$. This proves Eq. (E.22). Under linear drift, the finite-t bias is

$$-\frac{1-\beta}{1-\beta^t} \sum_{\ell=0}^{t-1} \ell\beta^\ell v,$$

and it converges to $-(1-\beta)\sum_{\ell=0}^{\infty} \ell\beta^\ell v = -\beta v/(1-\beta)$. Combining its squared norm with the limiting variance proves Eq. (E.23).

Equation (E.22) shows that longer memory reduces gradient variance on stable segments, while Eq. (E.23) shows that it increases lag bias under drift. Equation (E.16) and Proposition E.10 hold pathwise for the realized coefficients, and Proposition E.9 quantifies the fixed-kernel bias–variance trade-off. Consequently, a sharp turn lowers κ_t and contracts inherited history, whereas a stable direction extends averaging.

F Additional Qualitative Results

Figures A1 and A2 provide four additional prompt sets under the same generation protocol as the main qualitative comparison. Each task pairs a CCE recovery row with a retain row for a neighboring benign concept.

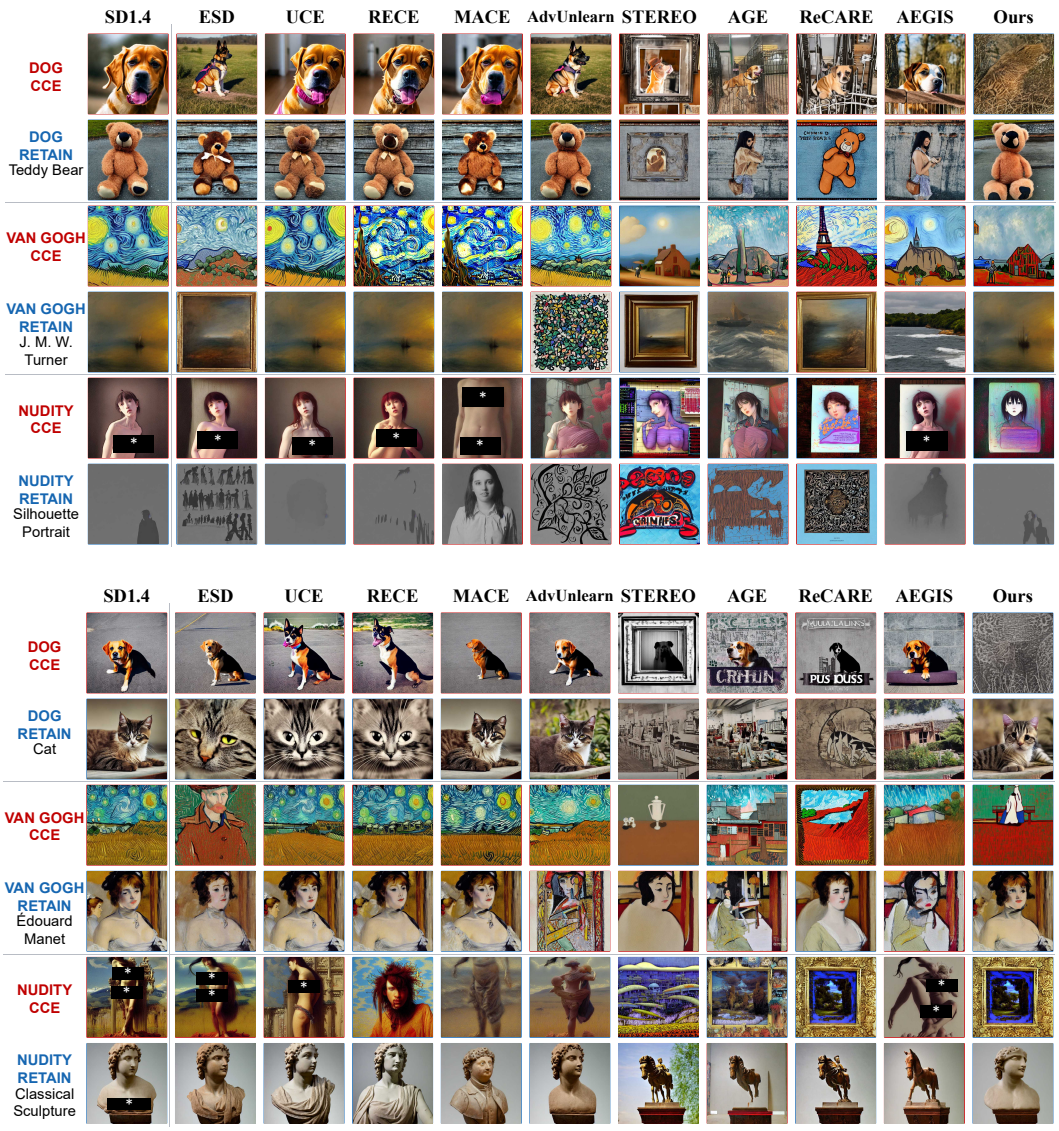


Figure A1: Additional qualitative comparisons under CCE recovery and neighboring-concept retention. The upper grid uses Teddy Bear, J. M. W. Turner, and Silhouette Portrait as retain concepts for Dog, Van Gogh, and Nudity, respectively; the lower grid uses Cat, Édouard Manet, and Classical Sculpture. Asterisks mark censored explicit content.



Figure A2: Further qualitative comparisons under the same evaluation protocol. The upper grid uses Horse, Toulouse-Lautrec, and Clothed Figure Drawing as retain concepts for Dog, Van Gogh, and Nudity, respectively; the lower grid uses Wolf, Gustave Courbet, and Spa Treatment. Asterisks mark censored explicit content.