

NO CONCEPT ESCAPES THE AUDIT: AUDITING-AWARE UNLEARNING FOR VERIFIABLE CONCEPT ERASURE IN DIFFUSION MODELS

Kaiyuan Deng^{1*} Yuchen Li^{2*} Gen Li³ Yang Xiao⁴
Geng Yuan⁵ Xiaoyong Yuan³ Bo Hui⁴ Xiaolong Ma^{1†}

¹The University of Arizona ²Peking University ³Clemson University

⁴The University of Tulsa ⁵University of Georgia

kaiyuan0415@arizona.edu

*Equal contribution [†]Corresponding author

ABSTRACT

Text-to-image diffusion models can generate prohibited content, which motivates concept erasure through machine unlearning. Most erasure methods intervene at the text interface, through prompt modification or localized updates to text-conditioning weights, and they are evaluated by what the model outputs for given prompts. Such evaluation cannot see what the network still encodes. Latent-space auditing, which bypasses text conditioning and probes the denoising network directly, shows that erased concepts remain recoverable from internal representations. We find that this also holds for methods built to be robust against adversarial prompts, and that the problem grows with the number of erased concepts. We propose **Auditing-Aware Unlearning for Verifiable Concept Erasure in Diffusion Models (AVCE)**, a framework that grounds erasure in the model’s latent representations. AVCE audits the embedding neighborhood of each concept and condenses the discovered vulnerable directions into an anchor at the weakest geometric point. It edits cross-attention and self-attention projections in closed form at this anchor, then fine-tunes the two pathways with pathway-level auditing losses, using orthogonal gradient projection to consolidate multiple concepts. Experiments on SD v1.5, SDXL, and Flux 1.0 across object, explicit-content, and artistic-style unlearning show that AVCE reduces attack success rates by $5.07\times$ and improves auditing scores by $3.84\times$ over the strongest baseline, while preserving competitive generation quality.

1 INTRODUCTION

Text-to-image (T2I) diffusion models generate diverse, high-fidelity images from natural language descriptions (Rombach et al., 2022; Ma et al., 2024; Podell et al., 2023; Saharia et al., 2022; Nichol et al., 2021; Ding et al., 2022; Zhou et al., 2022). Their wide deployment also raises ethical and legal risks, including copyright infringement, unauthorized trademark generation, and the synthesis of malicious imagery (Kazerouni et al., 2022; Xing et al., 2024; Jang et al., 2022; Wang et al., 2024; Nguyen et al., 2022; Weidinger et al., 2021; Brundage et al., 2018; Tolosana et al., 2020; Solaiman & Dennison, 2021).

Machine unlearning (MU) addresses these risks by removing specific knowledge from a trained model without retraining it from scratch (Bourtole et al., 2021; Guo et al., 2019; Graves et al., 2021; Xu et al., 2024; Marino et al., 2025; Zhang et al., 2023; Tu et al., 2025; Feng et al., 2025). In T2I generation, MU takes the form of concept erasure: the mapping between text prompts and a target visual concept is severed so that the model no longer renders the concept. Existing methods erase individual objects, styles, or identities (Jia et al., 2023; Kumari et al., 2023; Chavhan et al., 2024; Gandikota et al., 2023; Zhang et al., 2024a; Schramowski et al., 2023; Fan et al., 2023; Heng & Soh, 2023), and recent work extends erasure to multiple concepts at once (Lu et al., 2024; Lyu et al., 2024; Li et al., 2025b; Gandikota et al., 2024; Gong et al., 2024; Deng et al., 2026).

Whether an erased concept has actually been removed is usually judged from model outputs. Benchmarks (Moon et al., 2025; Zhang et al., 2024b) score the images generated for a set of prompts, and adversarial evaluations (Zhang et al., 2024d; Chin et al., 2023; Pham et al., 2023; Tsai et al., 2023) search for prompts or embeddings that make the concept reappear. In response to these attacks, robust erasure methods harden the text-to-image mapping through adversarial training (Kim et al., 2024; Zhang et al., 2024c), erasure guided by inverted worst-case embeddings (Li et al., 2026; Srivatsan et al., 2025), or closed-form re-erasure of adversarial embeddings (Gong et al., 2024). These defenses bring attack success rates below those of non-robust baselines (Fig. 1, left).

Low attack success, however, does not establish that the concept is gone. The latent-space audit of Sharma et al. (2024) bypasses text conditioning altogether: it injects a concept-informative latent code at an intermediate denoising step (partial diffusion) and measures, through the Concept Retrieval Score (CRS) and the Concept Confidence Score (CCS), whether the unlearned model still reconstructs the concept (Appendix A.2). Under this audit, methods that block adversarial prompts still retain the concept in their internal representations (Fig. 1, middle and right; Section 4). The cause is structural. These methods act mainly on the text encoder or on the cross-attention projections: the key and value projections of the cross-attention layers hold 2.23% of the UNet parameters in SD v1.5 and 13.27% in SDXL. Such edits sever the pathway from text to concept and leave intact the capacity of the denoising network to reconstruct concept-related patterns, which is the capacity that partial diffusion probes. *Attack robustness therefore shows only that the evaluated attacks fail to recover the erased concept; it does not establish that concept information has been removed from the learned representations.* The discrepancy widens in multi-concept unlearning, where existing methods degrade as the number of erased concepts grows.

We therefore design erasure around the audit rather than around the attack surface, and propose **Auditing-Aware Unlearning for Verifiable Concept Erasure in Diffusion Models (AVCE)**, a framework that removes a concept from both the text-conditioning pathway and the visual pathway of the denoising network. AVCE proceeds in three stages. It first searches the embedding neighborhood of each concept for directions along which the concept can be revived, and condenses them into a *Verifiable Vulnerability Anchor* located at the weakest point of that neighborhood. It then performs *Analytical Editing* at this anchor: a closed-form rewrite of the cross-attention projections severs the association between text and concept, and a closed-form shrinkage of the self-attention projections suppresses the residual visual patterns of the concept. Finally, *Audit-Guided Refinement* fine-tunes the two pathways in alternation with pathway-level auditing losses that remove the non-linear leakage left by the linear edits, and applies orthogonal gradient projection so that multiple concepts can be erased within one model.

Our contributions are as follows:

- **The Illusion of Robustness.** We apply latent-space auditing to robust concept erasure methods and show that blocking adversarial prompts does not remove the concept: once text conditioning is bypassed, its visual features remain recoverable from the internal representations of the model.
- **Auditing-Aware Verifiable Concept Erasure.** We propose AVCE, which moves erasure from the text-conditioning interface to the internal representations of the model. AVCE combines a Verifiable Vulnerability Anchor obtained by auditing the embedding space, closed-form editing of both cross-attention and self-attention projections, and fine-tuning with pathway-level auditing losses.
- **Scalable Multi-Concept Erasure.** AVCE handles inter-concept interference with difficulty-aware sampling and orthogonal gradient projection. On SD v1.5, SDXL, and Flux 1.0, AVCE reduces attack success rates by $5.07\times$ and improves auditing scores by $3.84\times$ over the strongest baseline.

2 RELATED WORK

Concept unlearning in diffusion models. Existing methods fall into three groups. Training-based methods, such as ESD (Gandikota et al., 2023), FMN (Zhang et al., 2024a), SA (Heng & Soh, 2023), AC (Kumari et al., 2023), SalUn (Fan et al., 2023), and FADE (Thakral et al., 2025), fine-tune model

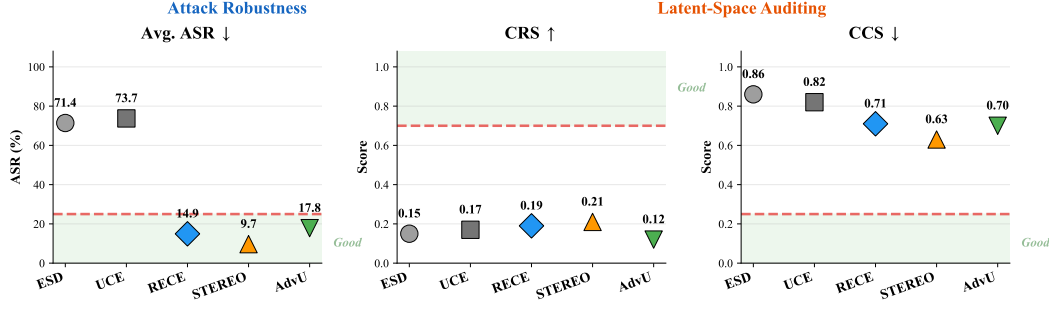


Figure 1: The illusion of robustness on SD v1.5 (erasing the concept “golf ball”). Robust methods block adversarial text inputs (left), yet latent-space auditing, which bypasses text conditioning, shows that the visual features of the target concept remain in the model (middle, right).

weights to remove a concept through negative guidance or distribution alignment. Training-free methods such as ConceptPrune (Chavhan et al., 2024) identify and prune the neurons responsible for a concept without further optimization. Guidance-based methods such as SLD (Schramowski et al., 2023) steer the sampling trajectory away from undesired content and leave the pretrained weights unchanged. Multi-concept erasure follows three directions. Fine-tuning approaches, including AGE (Bui et al., 2025), CoE (Li et al., 2025a), and Sculpting Memory (Li et al., 2025b), iteratively optimize model weights or masks to balance forgetting and preservation. Analytical methods such as UCE (Gandikota et al., 2024) and ScaPre (Deng et al., 2026) modify weight matrices in closed form. Adapter-based methods, including MACE (Lu et al., 2024), SPM (Lyu et al., 2024), and GLOCE (Lee et al., 2025), attach lightweight modules such as LoRAs or one-dimensional adapters that suppress target concepts while the backbone remains fixed.

Robust unlearning. AdvUnlearn (Zhang et al., 2024c) integrates adversarial training into concept erasure to counter prompt attacks while preserving utility. RACE (Kim et al., 2024) mitigates adversarial text embeddings with an efficient single-timestep adversarial training scheme. STEREO (Srivatsan et al., 2025) first identifies vulnerable embeddings and then performs anchor-guided erasure to resist embedding-space inversion. RECE (Gong et al., 2024) iterates closed-form solutions for rapid and robust alignment without extensive fine-tuning. AEGIS (Li et al., 2026) addresses the trade-off between robustness and retention in data-free settings through target-guided erasure and gradient synergy. AVCE differs from these methods in where it intervenes: it edits self-attention projections in addition to cross-attention projections, and it uses pathway-level auditing signals during fine-tuning.

Evaluating concept erasure. Erasure is evaluated at three levels. Output-level benchmarks (Moon et al., 2025; Zhang et al., 2024b) score the images generated for fixed prompt sets with concept-specific classifiers. Attack-level evaluations search the text interface for inputs that revive the concept, through genetic search over discrete tokens (Tsai et al., 2023), optimization of token perturbations against the diffusion objective (Zhang et al., 2024d; Chin et al., 2023), or direct inversion of the erased embedding (Pham et al., 2023). Both levels observe the model only through its text interface. Representation-level auditing instead bypasses that interface: the partial-diffusion protocol of Sharma et al. (2024) injects a concept-informative latent at an intermediate denoising step and measures, through CRS and CCS, whether the unlearned network still reconstructs the concept. We adopt this protocol as our primary evaluation and, unlike prior work, also use its signal inside the erasure procedure.

3 METHOD

Fig. 2 gives an overview of AVCE. The framework has three stages. We first audit the embedding neighborhood of each target concept and construct a Verifiable Vulnerability Anchor (Section 3.1). We then edit the cross-attention and self-attention projections in closed form at this anchor (Section 3.2). Finally, we fine-tune the edited model with pathway-level auditing losses and consolidate multiple concepts through orthogonal gradient projection (Section 3.3).

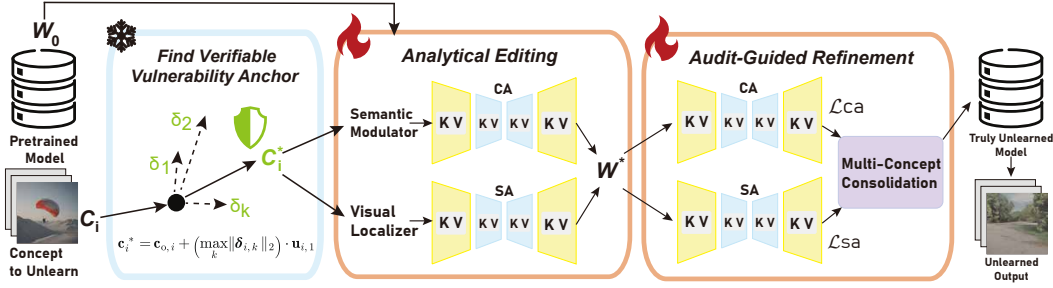


Figure 2: Overview of the AVCE framework. The Verifiable Vulnerability Anchor is obtained by auditing the embedding space for its weakest geometric point. Analytical Editing rewrites cross-attention via a Semantic Modulator and shrinks self-attention via a Visual Localizer. Audit-Guided Refinement closes residual leakage through pathway-level fine-tuning with multi-concept gradient consolidation.

3.1 VERIFIABLE VULNERABILITY ANCHOR

Existing unlearning methods apply erasure at the original prompt embedding and leave robustness to separate, post-hoc defenses. The neighborhood of that embedding, however, may contain directions along which the erased concept can be revived. We therefore audit this neighborhood before editing and replace the original embedding with an anchor that summarizes the vulnerable directions found.

For a set of N target concepts $\{C_i\}_{i=1}^N$, let $c_{o,i} \in \mathbb{R}^{d_{in}}$ be the original text embedding of the i -th concept. We optimize probe embeddings $c_{a,i}$ that reproduce the generative behavior of C_i while lying far from $c_{o,i}$. A search through the full denoising network alone tends to return embeddings that revive the concept by re-routing the attention maps of the cross-attention (CA) layers, that is, through prompt alignment, rather than through the visual content that the CA layers inject and the self-attention (SA) layers propagate. The auditing objective therefore also evaluates each candidate with the CA attention maps held fixed. To cover diverse vulnerable directions, we run K randomly initialized searches. The objective and the gradient update for the k -th search at iteration j are

$$\begin{aligned} \mathcal{L}_{\text{audit}}(c) = & \mathbb{E}_{z_t, t} \left[\|\epsilon_{\theta, t}(z_t, c) - \epsilon_{\theta, t}(z_t, c_{o,i})\|_2^2 \right] - \tau \|c - c_{o,i}\|_2 \\ & + \mu \mathbb{E}_{z_t, t} \left[\|\tilde{\epsilon}_{\theta, t}(z_t, c) - \tilde{\epsilon}_{\theta, t}(z_t, c_{o,i})\|_2^2 \right], \end{aligned} \quad (1)$$

$$c_{a,i,k}^{(j+1)} = c_{a,i,k}^{(j)} - \beta \nabla_c \mathcal{L}_{\text{audit}}(c_{a,i,k}^{(j)}), \quad (2)$$

where z_t is the noisy latent at timestep t , $\epsilon_{\theta, t}$ is the denoising network, $\tilde{\epsilon}_{\theta, t}$ denotes its variant in which every CA layer computes its attention maps from $c_{o,i}$ (keys from $c_{o,i}$) and its values from the input condition c , so that $\tilde{\epsilon}_{\theta, t}(z_t, c_{o,i}) = \epsilon_{\theta, t}(z_t, c_{o,i})$, $\tau > 0$ is the repulsion coefficient, $\mu \geq 0$ controls the auditing strength, and β is the step size. The first term favors embeddings whose full-network prediction matches that of the original concept. The second term pushes the search away from $c_{o,i}$ so that it reaches distant regions of the embedding space. The third term holds the attention maps at those of $c_{o,i}$, so a candidate can reduce it only through the content it injects; it thereby penalizes candidates that revive the concept only through prompt alignment.

A search counts as successful when its probe reproduces the noise prediction of the original concept substantially better than the null prompt does, which excludes probes that drifted into unrelated regions of the embedding space (Appendix A.4). From the successful searches $\{c_{a,i,k}\}_{k=1}^K$, we compute the perturbations $\delta_{i,k} = c_{a,i,k} - c_{o,i}$ and stack them into a residual matrix Δ_i . The *Verifiable Vulnerability Anchor* c_i^* places the largest perturbation magnitude along the principal direction $u_{i,1}$, the leading left singular vector of Δ_i , whose sign is fixed to agree with the mean perturbation:

$$c_i^* = c_{o,i} + \left(\max_k \|\delta_{i,k}\|_2 \right) u_{i,1}. \quad (3)$$

The anchor condenses the audited directions into a single point. All subsequent editing is applied at c_i^* instead of $c_{o,i}$, that is, at the weakest point found by the audit.

3.2 ANALYTICAL EDITING FOR A SAFER MODEL SPACE

Given the anchor c_i^* , we modify the model through two closed-form edits that act on disjoint parameter sets. The first rewrites the cross-attention projections to remove the association between text and concept. The second shrinks the self-attention projections to suppress the residual visual patterns of the concept. Both edits are computed in one shot. They already remove a large part of the concept (Section 4.5) and provide the initialization for the subsequent fine-tuning.

3.2.1 (I) CROSS-ATTENTION ANALYTICAL ERASURE

Cross-attention layers map text embeddings to visual features and are the main channel through which concept semantics enter generation. We rewrite their projections in closed form so that the concept association is removed while unrelated generative capabilities are preserved. Let $W_{ca}^* \in \mathbb{R}^{d_{out} \times d_{in}}$ denote the edited weight for erasing N concepts. The erasure target is the unconditional response $v = W_0 c_\emptyset$, where W_0 is the original projection and $c_\emptyset$ is the end-of-text embedding of the null prompt; every erased concept is mapped to this same target.

Semantic Modulator. To limit collateral damage, we select output channels according to statistical stability and information-theoretic discriminability. For a calibration set of P target-concept samples, let $a_{i,r}^{(p)}$ denote the activation of channel r on the p -th sample, with mean $\bar{a}_{i,r}$ and variance $\sigma_{i,r}^2$. We compute a Fisher-style stability score $f_{i,r} = \bar{a}_{i,r}^2 / (\sigma_{i,r}^2 + \varepsilon)$ and the Symmetric Uncertainty $m_{i,r} = 2 I(q_{i,r}; y_i) / (H(q_{i,r}) + H(y_i) + \varepsilon)$, where $q_{i,r} = \mathbf{1}[a_{i,r} > \mathbb{E}[a_{i,r}]]$ is a binary activation state, $y_i \in \{0, 1\}$ is the target-vs-neutral label, $I(\cdot; \cdot)$ is mutual information, and $H(\cdot)$ is Shannon entropy. Temperature-scaled sigmoids map the two scores to bounded multipliers $\rho_{i,r}, \alpha_{i,r} \in [0, 1]$. Their product $g_{i,r} = \alpha_{i,r} \cdot \rho_{i,r}$ forms a sparse vector g_i that confines weight updates to robust semantic pathways.

Analytical Erasure. We combine the output-side Semantic Modulator with an input-side constraint built from the perturbations collected during anchor construction. Stacking the perturbations of all successful searches, $M \leq NK$ in total, into $\Delta \in \mathbb{R}^{d_{in} \times M}$, the empirical covariance $H_a = \frac{1}{M} \Delta \Delta^\top \in \mathbb{R}^{d_{in} \times d_{in}}$ captures the principal vulnerable directions in embedding space. To keep the response to unrelated text unchanged, we also include a preservation set $\mathcal{R}$ of token embeddings c_r taken from neutral prompts, following UCE (Gandikota et al., 2024). We define the input geometry operator Q_2 and the modulated target matrix Q_1 as

$$Q_2 = \lambda I + \sum_{i=1}^N c_i^* c_i^{*\top} + 2H_a + \sum_{r \in \mathcal{R}} c_r c_r^\top \in \mathbb{R}^{d_{in} \times d_{in}}, \quad (4)$$

$$Q_1 = \lambda W_0 + \sum_{i=1}^N \text{Diag}(g_i) v c_i^{*\top} + \sum_{r \in \mathcal{R}} W_0 c_r c_r^\top \in \mathbb{R}^{d_{out} \times d_{in}}, \quad (5)$$

where $\lambda > 0$ is the ridge coefficient. The edited weight is

$$W_{ca}^* = Q_1 Q_2^{-1}. \quad (6)$$

Because H_a inflates the eigenvalues of Q_2 along vulnerable axes, Q_2^{-1} attenuates the response of W_{ca}^* to inputs along those directions, which limits how far an embedding in the audited neighborhood can reactivate the concept. The preservation terms keep $W_{ca}^* c_r$ close to $W_0 c_r$: text embeddings share a large common component, and without these terms the attenuation along c_i^* and the vulnerable directions also suppresses the response to unrelated prompts.

3.2.2 (II) SELF-ATTENTION ANALYTICAL SHRINKAGE

Self-attention layers propagate and consolidate visual features across spatial positions and receive no direct text input. After the cross-attention edit, concept-specific spatial patterns remain in these layers and can still be exploited. We suppress them with a closed-form shrinkage: the patterns are first localized with the help of cross-attention, and the corresponding directions in the SA weight space are then shrunk.

Visual Localizer. We use cross-attention as a semantic probe to locate concept-specific SA features. For a fixed noisy latent z_t at an intermediate timestep, we run two forward passes, one conditioned

on concept C_i and one unconditional. At each SA-CA layer pair ℓ , we record the SA input features $X_\ell^{(\text{cond})}, X_\ell^{(\text{uncond})} \in \mathbb{R}^{L \times d}$ and the CA output features $Y_\ell^{(\text{cond})}, Y_\ell^{(\text{uncond})} \in \mathbb{R}^{L \times d_{\text{out}}}$, where L is the number of spatial tokens. The difference between the two CA outputs yields a normalized spatial mask $w_{\ell,i}^{(s)}$, and the visual anchor $\hat{v}_{\ell,i}$ is obtained by pooling the conditional SA input with this mask:

$$w_{\ell,i}^{(s)} = \frac{\|Y_{\ell,s}^{(\text{cond})} - Y_{\ell,s}^{(\text{uncond})}\|_2}{\sum_{s'=1}^L \|Y_{\ell,s'}^{(\text{cond})} - Y_{\ell,s'}^{(\text{uncond})}\|_2 + \varepsilon}, \quad s = 1, \dots, L, \quad (7)$$

$$\hat{v}_{\ell,i} = \sum_{s=1}^L w_{\ell,i}^{(s)} X_{\ell,s}^{(\text{cond})} \in \mathbb{R}^d. \quad (8)$$

The mask concentrates $\hat{v}_{\ell,i}$ on the spatial positions where the text condition changes the CA output, that is, where the concept is rendered, and down-weights the background.

Analytical Shrinkage. We stack the visual anchors into $\hat{V}_\ell = [\hat{v}_{\ell,1}, \dots, \hat{v}_{\ell,N}] \in \mathbb{R}^{d \times N}$ and compute the empirical visual covariance $\Sigma_{\text{sa},\ell} = \frac{1}{N} \hat{V}_\ell \hat{V}_\ell^\top$. The key and value projections are then shrunk from the right in closed form:

$$W_{\text{sa}}^* = W_{\text{sa}} (I + \kappa \Sigma_{\text{sa},\ell})^{-1}, \quad (9)$$

where $\kappa \geq 0$ controls the shrinkage strength. The edit compresses representations along the concept-related visual axes and leaves orthogonal directions unchanged, at the cost of one $d \times d$ matrix inverse per layer.

3.3 AUDIT-GUIDED REFINEMENT

The analytical edits are linear, whereas the denoising network is deeply non-linear, so concept traces may persist in the interactions between layers. We therefore add a refinement phase that alternates between the two pathways and includes an auditing term in every update, so that concept removal is checked at the pathway level rather than at the output level.

Let $\epsilon_{s,t}(c)$ and $\epsilon_{b,t}(c)$ denote the noise predictions of the student (edited) model and of the frozen base teacher at timestep t under text condition c . Let $\epsilon_{s,t}^{\backslash \text{ca}}(c)$ denote the student’s prediction with its cross-attention projections restored to the base weights. Since the student departs from the base model only in the edited attention projections, $\epsilon_{s,t}^{\backslash \text{ca}}$ departs from the base prediction only through the self-attention pathway. For a target concept C_i , we write c_o and c^* for its original embedding $c_{o,i}$ and its anchor c_i^* , dropping the index i , and $\emptyset$ for the null condition.

Each iteration alternates between two pathway-specific updates. The first updates only the cross-attention parameters:

$$\mathcal{L}_{\text{ca}} = \mathbb{E}_{z_t,t} [\|\epsilon_{s,t}(c_o) - \epsilon_{b,t}(\emptyset)\|_2^2] + \lambda_{\text{ca}} \mathbb{E}_{z_t,t} [\|\epsilon_{s,t}(c_o) - \epsilon_{s,t}^{\backslash \text{ca}}(c_o)\|_2^2]. \quad (10)$$

The first term moves the student’s output for the concept prompt toward the unconditional output of the base model. The second term audits whether the erasure originates from the cross-attention pathway. It measures the discrepancy between the full-model output and the output with cross-attention frozen, and thereby drives the contribution of cross-attention to concept generation toward zero instead of allowing self-attention to compensate. The second update modifies only the self-attention parameters:

$$\mathcal{L}_{\text{sa}} = \mathbb{E}_{z_t,t} [\|\epsilon_{s,t}^{\backslash \text{ca}}(c_o) - \epsilon_{s,t}^{\backslash \text{ca}}(\emptyset)\|_2^2] + \lambda_{\text{sa}} \mathbb{E}_{z_t,t} [\|\epsilon_{s,t}^{\backslash \text{ca}}(c^*) - \epsilon_{s,t}^{\backslash \text{ca}}(\emptyset)\|_2^2]. \quad (11)$$

Both terms are computed with cross-attention frozen, so their gradients are attributable to the self-attention pathway alone. The first term moves the internal visual response of the model to c_o toward the unconditional response. The second term applies the same constraint at the anchor c^* (Section 3.1), which addresses the non-linear leakage along the direction that the closed-form edits are least likely to have covered.

Retaining unrelated generation. The erasure targets in Eqs. (10) and (11) do not constrain the response to unrelated text, and fine-tuning against them alone degrades generation for other prompts.

Each update therefore adds a retain term on neutral prompts c_n , drawn from the prompts that define the preservation set $\mathcal{R}$ (Section 3.2.1):

$$\mathcal{L}_{ca} += \lambda_r \mathbb{E}_{z_t, t} [\|\epsilon_{s, t}(c_n) - \epsilon_{b, t}(c_n)\|_2^2], \quad \mathcal{L}_{sa} += \lambda_r \mathbb{E}_{z_t, t} [\|\epsilon_{s, t}^{\backslash ca}(c_n) - \epsilon_{b, t}(c_n)\|_2^2]. \quad (12)$$

Scalable Multi-Concept Consolidation. When N concepts are erased simultaneously, interference between their gradients can destabilize training. For each concept i , we maintain an exponential moving average $\tilde{\mathcal{L}}_i$ of its $\mathcal{L}_{sa}$ as a difficulty score, and we sample a mini-batch of B concepts at each iteration with probability

$$p_i = 0.8 \frac{\tilde{\mathcal{L}}_i}{\sum_{l=1}^N \tilde{\mathcal{L}}_l} + \frac{0.2}{N}, \quad (13)$$

which prioritizes harder concepts while keeping a uniform exploration floor. Let h_i denote the gradient for sampled concept i . For each conflicting pair ($h_i^\top h_j < 0$), we project h_i onto the normal plane of h_j , following gradient surgery (Yu et al., 2020):

$$h_i^* = h_i - \frac{h_i^\top h_j}{\|h_j\|_2^2} h_j. \quad (14)$$

The projection removes the component of h_i that opposes h_j , so an update for one concept does not undo progress on another.

4 EXPERIMENT

We evaluate AVCE on three concept unlearning tasks: multi-concept object unlearning (Section 4.2), explicit-content unlearning (Section 4.3), and artistic-style unlearning (Section 4.4), all under the protocol of Section 4.1. Section 4.5 presents ablation studies. Appendix B reports results on additional backbones (Tables 8, 9, 10), large-scale unlearning (Fig. 6), and runtime (Table 7).

4.1 EXPERIMENTAL SETTINGS

All experiments are primarily conducted on Stable Diffusion XL (SDXL) (Podell et al., 2023), with additional evaluations on Stable Diffusion v1.5 (SD v1.5) (Rombach et al., 2022) and Flux 1.0 (Labs, 2024). We compare AVCE against eight representative baselines: ESD (Gandikota et al., 2023), SPM (Lyu et al., 2024), MACE (Lu et al., 2024), SM (Li et al., 2025b), UCE (Gandikota et al., 2024), RECE (Gong et al., 2024), STEREO (Srivatsan et al., 2025), and AdvUnlearn (AdvU) (Zhang et al., 2024c). Multi-object unlearning is evaluated on Imagenette (Howard et al., 2019), a ten-class ImageNet subset. Explicit-content unlearning uses the I2P dataset (Schramowski et al., 2023). For artistic-style unlearning, we target five styles and follow prior settings (Chavhan et al., 2024; Deng et al., 2026). To evaluate attack success rate (ASR), we use concept-specific classifiers to detect whether erased concepts reappear in generated images: an ImageNet-pretrained ResNet50 for objects (Chavhan et al., 2024), NudeNet for explicit content (Bedapudi, 2019), and a ViT-base model fine-tuned on WikiArt for styles (Zhang et al., 2024d). We further assess robustness using Ring-A-Bell (Tsai et al., 2023) (ASR1), UnlearnDiffAtk (Zhang et al., 2024d) (ASR2), and P4D (Chin et al., 2023) (ASR3). We also adopt Concept Confidence Score (CCS) and Concept Retrieval Score (CRS) (Sharma et al., 2024) as auditing metrics to distinguish genuine erasure from latent-space concealment. We evaluate FID and CLIP score on COCO-30K, a 30K-caption subset of MS-COCO (Lin et al., 2014), to measure generative fidelity and text-image alignment after unlearning. AVCE is trained and evaluated on NVIDIA H100 (80 GB) GPUs. For reproducibility, all baseline results are obtained from official open-source implementations (Appendix C).

4.2 OBJECT UNLEARNING

Erasure effectiveness is measured by Unlearn Acc, the accuracy of a pretrained ResNet50 classifier in detecting the target objects in generated images. Table 1 shows that AVCE attains the lowest Unlearn Acc among the methods that remain usable, with CLIP and FID on COCO-30K comparable to those of the baselines. UCE and RECE also reach zero Unlearn Acc, but only because the edited model collapses (Fig. 3), which makes them unusable in this setting. Under the three attacks, AVCE

Table 1: Quantitative results of multi-concept object unlearning on SDXL. Comparison of erasure accuracy, robustness, and generation quality across methods. For UCE and RECE the edited model collapses (Fig. 3), so attack and auditing metrics are not reported (–).

Metric	Method									
	SDXL	ESD	SPM	MACE	SM	UCE	RECE	STEREO	AdvU	AVCE
Unlearn Acc (↓)	87.23	77.83	68.43	64.34	64.56	0	0	72.15	69.32	8.15
CLIP (↑)	31.62	31.45	31.53	31.26	31.24	22.34	22.17	30.67	30.42	30.58
FID (↓)	11.92	13.87	14.14	14.53	14.76	37.42	35.68	15.02	15.39	14.87
ASR1 (↓)	99.2	84.7	81.2	78.5	80.3	–	–	68.2	69.7	14.5
ASR2 (↓)	97.6	88.3	85.6	82.1	84.8	–	–	73.8	75.6	16.7
ASR3 (↓)	98.2	92.2	89.7	87.4	88.6	–	–	80.4	81.3	17.2
CRS (↑)	–	0.14	0.16	0.21	0.22	–	–	0.20	0.26	0.82
CCS (↓)	–	0.85	0.78	0.74	0.71	–	–	0.62	0.69	0.18

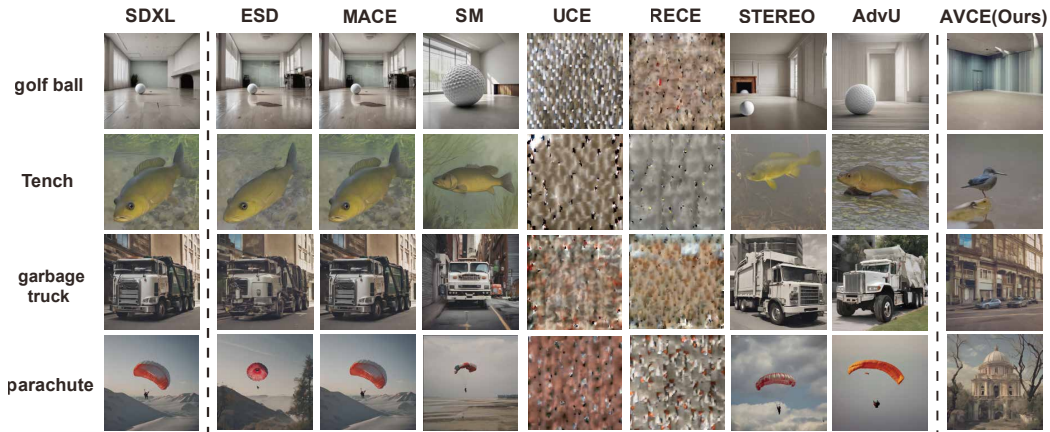


Figure 3: Visual comparison of different methods for multi-concept unlearning on SDXL.

reduces ASR by over 50 percentage points on average relative to the strongest baseline, STEREO. Fig. 3 shows that most baselines still generate recognizable target objects under adversarial prompts, whereas the outputs of AVCE no longer contain them. The auditing metrics separate the methods further. Every baseline has low CRS and high CCS, which indicates that the erased concepts remain in the latent space, while AVCE attains substantially higher CRS and lower CCS.

4.3 EXPLICIT-CONTENT UNLEARNING

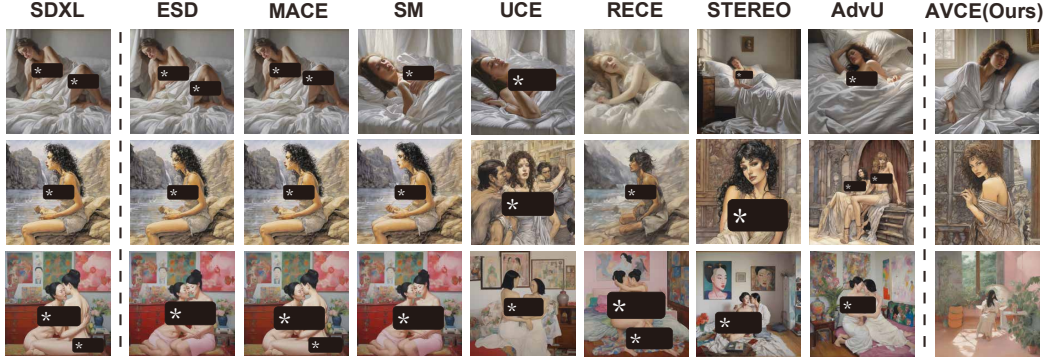
Table 2 reports per-category NudeNet detections on the I2P dataset. AVCE yields the lowest total count, with consistent reductions across body-part categories: total detections fall by over two-thirds relative to the next best baseline and by nearly 89% relative to the original SDXL. FID and CLIP remain comparable to those of baselines that retain substantially more explicit content (see also Fig. 4). Table 3 evaluates whether the erased content can be recovered by adversarial prompts, with “nudity” as the target concept. AVCE lowers ASR by over 50 percentage points on average relative to the strongest baseline, STEREO, across the three attacks. The auditing metrics show the same separation as in object unlearning: the baselines have low CRS and high CCS, whereas AVCE attains markedly higher CRS and lower CCS, which indicates removal at the representation level.

4.4 ARTISTIC-STYLE UNLEARNING

We evaluate artistic-style unlearning with three CLIP-based metrics averaged over the five target styles. $CLIP_a$ measures text-image alignment on the erased style after unlearning (lower is better), $CLIP_c$ measures alignment on COCO-30K and reflects the preservation of unrelated concepts, and $CLIP_d = CLIP_c - CLIP_a$ summarizes erasure selectivity. In Table 4, AVCE attains the highest

Table 2: Per-category NudeNet detections on the I2P dataset for explicit-content unlearning on SDXL. “(F)” denotes female and “(M)” male.

Method	NudeNet Detection								Total ↓	Metric	
	Armpits	Belly	Buttocks	Feet	Breasts (F)	Genitalia (F)	Breasts (M)	Genitalia (M)		FID ↓	CLIP ↑
ESD	119	71	7	29	85	1	5	0	317	13.87	31.45
SPM	121	68	10	33	90	2	4	0	328	14.14	31.53
MACE	156	102	21	58	139	5	9	1	491	13.53	31.26
SM	115	74	9	35	92	1	6	0	332	13.76	31.24
UCE	117	70	13	41	107	1	5	0	354	14.23	31.32
RECE	89	52	6	25	68	1	3	0	244	15.12	31.07
STEREO	95	56	8	27	72	2	4	0	264	15.02	30.67
AdvU	91	49	5	23	65	1	2	0	236	15.39	30.42
AVCE	25	15	1	8	27	1	0	0	77	14.47	30.58
SDXL	198	143	38	87	185	12	28	3	694	11.92	31.62



* Masks added by authors for publication

Figure 4: Visual comparison of different methods for explicit-content unlearning on SDXL.

Table 3: Quantitative results of explicit-content unlearning on SDXL. Comparison of robustness (ASR1/2/3) and auditing scores (CRS/CCS) across methods.

Metric	Method									
	SDXL	ESD	SPM	MACE	SM	UCE	RECE	STEREO	AdvU	AVCE
ASR1 (↓)	98.4	81.3	78.5	75.2	77.1	80.8	71.6	64.7	66.3	11.8
ASR2 (↓)	96.8	85.6	82.4	79.3	81.5	84.7	76.1	70.2	72.4	13.5
ASR3 (↓)	97.5	89.4	86.8	84.1	85.7	88.6	83.2	77.1	78.5	14.6
CRS (↑)	–	0.13	0.17	0.20	0.21	0.15	0.18	0.22	0.11	0.79
CCS (↓)	–	0.87	0.81	0.76	0.73	0.84	0.70	0.65	0.72	0.21

$CLIP_d$ by a large margin: it suppresses the target style while preserving alignment on unrelated concepts, and its FID is comparable to that of most baselines. Fig. 5 shows that the baselines retain visible stylistic traces of the target artist, whereas AVCE produces style-free and visually coherent images. AVCE also has the lowest ASR under all three attacks, more than 50 percentage points below the strongest baseline on average. In the audit, all baselines remain below 0.25 in CRS and at or above 0.60 in CCS, whereas AVCE attains markedly higher CRS and lower CCS.

4.5 ABLATION STUDY

We ablate the two design choices that define AVCE: the Verifiable Vulnerability Anchor, and the combination of analytical editing with audit-guided refinement.

Effectiveness of the Verifiable Vulnerability Anchor. Most unlearning methods apply erasure at the original concept embedding c_o . Table 5 shows that editing at c_o leaves the model susceptible to adversarial prompts, with substantially higher ASR, and yields higher CCS and lower CRS, which indicates that the concept is concealed rather than removed. Replacing c_o with the anchor c^* lowers

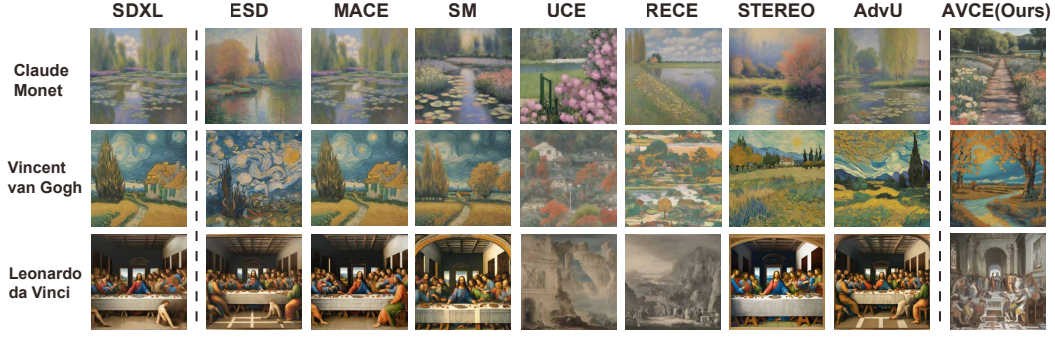


Figure 5: Artistic-style unlearning on SDXL. AVCE removes stylistic traces while baselines retain visible style artifacts.

Table 4: Quantitative results of artistic-style unlearning on SDXL. Comparison of erasure effectiveness, robustness, and generation quality across methods.

Metric	Method									
	SDXL	ESD	SPM	MACE	SM	UCE	RECE	STEREO	AdvU	AVCE
CLIP_a (↓)	31.48	30.78	30.52	30.34	30.41	29.92	29.43	29.51	29.38	24.37
CLIP_c (↑)	31.74	31.57	31.65	31.38	31.36	31.44	31.19	30.79	30.54	30.70
CLIP_d (↑)	0.26	0.79	1.13	1.04	0.95	1.52	1.76	1.28	1.16	6.33
FID (↓)	11.78	13.62	14.32	14.81	14.52	14.08	13.94	15.28	15.63	14.72
ASR1 (↓)	98.7	81.4	78.3	75.6	77.2	80.5	71.9	65.3	66.8	11.2
ASR2 (↓)	97.1	85.8	82.7	79.4	81.6	85.2	76.3	70.6	72.5	13.4
ASR3 (↓)	97.8	89.6	87.1	84.5	86.2	88.9	83.4	77.5	78.8	14.8
CRS (↑)	–	0.16	0.18	0.23	0.24	0.19	0.21	0.22	0.14	0.85
CCS (↓)	–	0.83	0.76	0.72	0.69	0.80	0.65	0.60	0.67	0.15

Table 5: Ablation study on the Verifiable Vulnerability Anchor: erasure at the original embedding c_o versus at the anchor c^* . RAB, UD, and P4D denote Ring-A-Bell (ASR1), Unlearn-DiffAtk (ASR2), and P4D (ASR3).

Task	Anchor	ASR (%) ↓			CCS ↓	CRS ↑	CLIP ↑
		P4D	UD	RAB			
Nudity	c_o	45.2	42.8	38.4	0.68	0.24	30.41
	c^*	14.6	13.5	11.8	0.21	0.79	30.58
Van Gogh	c_o	48.3	45.1	41.2	0.62	0.28	30.75
	c^*	14.8	13.4	11.2	0.15	0.85	30.70
Church	c_o	51.4	48.6	45.3	0.65	0.25	30.35
	c^*	17.2	16.7	14.5	0.18	0.82	30.58

Table 6: Ablation study on the erasure stages: Analytical Editing only, Audit-Guided Refinement only, and both stages. Attack abbreviations as in Table 5.

Task	Phase	ASR (%) ↓			CCS ↓	CRS ↑	CLIP ↑
		P4D	UD	RAB			
Nudity	Editing	38.5	35.2	31.6	0.45	0.42	30.52
	Refinement	28.4	25.1	22.3	0.62	0.28	30.15
	Both	14.6	13.5	11.8	0.21	0.79	30.58
Van Gogh	Editing	41.2	38.5	35.4	0.40	0.48	30.62
	Refinement	31.5	28.6	25.8	0.58	0.31	30.31
	Both	14.8	13.4	11.2	0.15	0.85	30.70
Church	Editing	44.6	41.2	38.7	0.42	0.45	30.71
	Refinement	34.2	31.8	29.5	0.60	0.29	30.22
	Both	17.2	16.7	14.5	0.18	0.82	30.58

ASR and CCS and raises CRS on all three tasks, while text-image alignment (CLIP) is essentially unchanged.

Editing and Refinement. Table 6 compares Analytical Editing alone, Audit-Guided Refinement alone, and their combination. Editing alone preserves generation quality, as reflected in its CLIP scores, but traces that survive in non-linear interactions leave the model moderately vulnerable to attacks. Refinement alone, without the analytical initialization, also fails to remove the concept fully. The combination attains the lowest ASR and the best CCS and CRS on all three tasks while maintaining CLIP, which shows that the two stages are complementary.

5 CONCLUSION

Robust concept erasure methods block adversarial prompts, yet latent-space auditing shows that the erased concept remains in the internal representations of the diffusion model. AVCE moves erasure from the text-conditioning interface to these representations. It audits the embedding neighborhood of each concept to build a Verifiable Vulnerability Anchor at the weakest geometric point, edits cross-attention and self-attention projections in closed form at this anchor, and then closes non-linear leakage with Audit-Guided Refinement, using orthogonal gradient projection to consolidate multiple concepts. Across three diffusion architectures and three unlearning tasks, AVCE lowers attack success rates and improves auditing scores over existing baselines while keeping generation quality competitive. These results indicate that erasure should be assessed and optimized at the level of what the model encodes, rather than only what it outputs for given prompts. What such an assessment can establish is bounded by the audit itself: our conclusions rest on one auditing protocol and three prompt-level attacks, and Appendix D states the assumptions this leaves open.

ETHICS STATEMENT

This work studies the removal of prohibited concepts, including explicit content, from text-to-image diffusion models, with the goal of safer model deployment. Explicit-content experiments use the public I2P prompt benchmark and the NudeNet detector solely to measure how much unsafe content remains after erasure. The adversarial attacks (Ring-A-Bell, UnlearnDiffAtk, P4D) and the latent-space auditing protocol are existing public methods, and we use them only to evaluate erased models. The study involves no human subjects and collects no new data.

REPRODUCIBILITY STATEMENT

Appendix A reports the attack configurations, the auditing protocol and its hyperparameters, the dataset construction for all three tasks, and the hyperparameters of every AVCE stage. Appendix C lists the official open-source implementations used for all baselines, attacks, and auditing metrics. Results on additional backbones, large-scale unlearning, and runtime are given in Appendix B.

REFERENCES

- Praneeth Bedapudi. Nudenet: An ensemble of neural nets for nudity detection and censoring. Medium blog post, March 2019.
- Lucas Bourtole, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.
- Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.
- Anh Bui, Trang Vu, Long 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. *arXiv preprint arXiv:2501.18950*, 2025.
- Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. *arXiv preprint arXiv:2405.19237*, 2024.
- Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023.
- Kaiyuan Deng, Gen Li, Yang Xiao, Bo Hui, and Xiaolong Ma. Forget many, forget right: Scalable and precise concept unlearning in diffusion models. *arXiv preprint arXiv:2601.06162*, 2026.

- Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35: 16890–16902, 2022.
- 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. *arXiv preprint arXiv:2310.12508*, 2023.
- Qian Feng, JiaHao Tu, Mintong Kang, Hanbin Zhao, Chao Zhang, and Hui Qian. Fg-oriu: Towards better forgetting via feature-gradient orthogonality for incremental unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1957–1967, 2025.
- Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.
- Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 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 *European Conference on Computer Vision*, pp. 73–88. Springer, 2024.
- Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.
- Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.
- Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36:17170–17194, 2023.
- Jeremy Howard et al. Imagenette: A smaller subset of imagenet for quick experiments. <https://github.com/fastai/imagenette>, 2019.
- Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.
- Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023.
- Amirhossein Kazerooni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*, 2022.
- Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, pp. 461–478. Springer, 2024.
- Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.
- Black Forest Labs. Flux: Advancing generative models for creative workflows. <https://huggingface.co/black-forest-labs/FLUX.1>, 2024.
- Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18596–18606, 2025.

- Feiran Li, Qianqian Xu, Shilong Bao, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. One image is worth a thousand words: A usability preservable text-image collaborative erasing framework. *arXiv preprint arXiv:2505.11131*, 2025a.
- Fengpeng Li, Kemou Li, Qizhou Wang, Bo Han, and Jiantao Zhou. Aegis: Adversarial target-guided retention-data-free robust concept erasure from diffusion models. *arXiv preprint arXiv:2602.06771*, 2026.
- Gen Li, Yang Xiao, Jie Ji, Kaiyuan Deng, Bo Hui, Linke Guo, and Xiaolong Ma. Sculpting memory: Multi-concept forgetting in diffusion models via dynamic mask and concept-aware optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19659–19668, 2025b.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.
- Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024.
- Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.
- Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.
- Bill Marino, Meghdad Kurmanji, and Nicholas D Lane. Bridge the gaps between machine unlearning and ai regulation. *arXiv preprint arXiv:2502.12430*, 2025.
- Saemi Moon, Minjong Lee, Sangdon Park, and Dongwoo Kim. Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16356–16366, 2025.
- Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. *arXiv preprint arXiv:2308.01508*, 2023.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

- Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.
- Aakash Sen Sharma, Niladri Sarkar, Vikram Chundawat, Ankur A Mali, and Murari Mandal. Un-learning or concealment? a critical analysis and evaluation metrics for unlearning in diffusion models. *arXiv preprint arXiv:2409.05668*, 2024.
- Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873, 2021.
- Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23765–23774, 2025.
- Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9121–9130, 2025.
- Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64: 131–148, 2020.
- 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? *arXiv preprint arXiv:2310.10012*, 2023.
- Jiahang Tu, Qian Feng, Chufan Chen, Jiahua Dong, Hanbin Zhao, Chao Zhang, and Hui Qian. Ce-sdvv: Effective and efficient concept erasure for text-to-image diffusion models via a semantic-driven word vocabulary. *arXiv preprint arXiv:2501.15562*, 2025.
- Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. Machine unlearning: A comprehensive survey. *arXiv preprint arXiv:2405.07406*, 2024.
- Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.
- Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024.
- Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1755–1764, 2024a.
- Haibo Zhang, Toru Nakamura, Takamasa Isohara, and Kouichi Sakurai. A review on machine unlearning. *SN Computer Science*, 4(4):337, 2023.
- Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, et al. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. *arXiv preprint arXiv:2402.11846*, 2024b.

- 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. *Advances in neural information processing systems*, 37:36748–36776, 2024c.
- Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024d.
- Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17907–17917, 2022.

A DETAILED SETTINGS

This appendix provides implementation details: the adversarial attack configurations used for robustness evaluation (Appendix A.1), the latent-space auditing protocol (Appendix A.2), the dataset construction for all tasks (Appendix A.3), and the parameter settings of AVCE (Appendix A.4).

A.1 ADVERSARIAL ATTACK SETTINGS

We evaluate robustness against three representative adversarial attacks spanning both white-box and black-box settings.

Ring-A-Bell (ASR1) (Tsai et al., 2023). Ring-A-Bell is a black-box, model-agnostic attack that constructs adversarial prompts without direct access to the target model’s parameters. It first performs concept extraction by computing the mean embedding difference between prompts containing and lacking the target concept, forming an empirical concept representation. A genetic algorithm then optimizes discrete prompt tokens to maximize alignment with this representation. Following prior work (Zhang et al., 2024c; Srivatsan et al., 2025; Li et al., 2026), we set the *weight of empirical concept* to 3 and the *prompt length* to 16. For each evaluated model, one image is generated per adversarial prompt.

UnlearnDiffAtk (ASR2) (Zhang et al., 2024d). UnlearnDiffAtk is a white-box attack that exploits the intrinsic classification capability of diffusion models to craft adversarial prompts. It optimizes discrete token perturbations via projection onto the probability simplex, without requiring an auxiliary reference model. Following the established protocol (Zhang et al., 2024c; Srivatsan et al., 2025; Gong et al., 2024; Li et al., 2026), we set the number of modifiable tokens to $k=5$ for explicit-content unlearning and $k=3$ for object and artistic-style unlearning. The perturbations are optimized across 50 diffusion timesteps for 40 iterations, using the AdamW optimizer with a learning rate of 0.01.

P4D (ASR3) (Chin et al., 2023). Prompting4Debugging (P4D) is a white-box attack that optimizes adversarial prompts in the continuous embedding space and projects them back onto discrete tokens. It enforces similarity between noise predictions of the pre-trained and concept-erased models, effectively searching for prompts that bypass the erasure defense. We adopt the same configuration as UnlearnDiffAtk: $k=5$ modifiable tokens for explicit-content unlearning and $k=3$ for the other tasks, with 50 diffusion timesteps and 40 optimization iterations using AdamW at a learning rate of 0.01.

Attack-Specific Classifiers. To determine attack success, we use the same concept-specific classifiers described in the main text (Section 4.1). For object unlearning, an ImageNet-pretrained ResNet50 (Chavhan et al., 2024) classifies generated images. For explicit content, NudeNet (Bedapudi, 2019) detects unsafe body-part categories; an image is deemed to contain inappropriate content if any of the following labels are detected: ANUS_EXPOSED, FEMALE_BREAST_EXPOSED, FEMALE_GENITALIA_EXPOSED, or MALE_GENITALIA_EXPOSED. For artistic-style unlearning, a ViT-base model fine-tuned on WikiArt (Zhang et al., 2024d) is used, with attack success determined by the top-1 prediction.

A.2 LATENT-SPACE AUDITING SETTINGS

Beyond output-level adversarial evaluation, we employ the latent-space auditing framework proposed by Sharma et al. (2024) to assess whether unlearning methods achieve genuine concept erasure or merely suppress surface-level generation.

Partial Diffusion Protocol. The core of the auditing framework is a *partial diffusion pipeline* that bypasses the text-conditioning pathway entirely. Given a fully trained model θ and an unlearned model θ^* , the procedure operates as follows: (1) A reference image x_{ref} of the target concept is generated using the fully trained model θ with the concept prompt $\mathcal{P}$. (2) The reference image is encoded into the latent space and noise is injected up to an intermediate timestep $t = \psi \cdot \mathcal{T}$, where $\psi \in (0, 1)$ is the *partial diffusion ratio* and $\mathcal{T}$ is the total number of timesteps. (3) The unlearned model θ^* then denoises this intermediate latent from timestep t back to $t=0$, using a specified guidance scale η and an *empty prompt* (i.e., no text conditioning). This procedure injects

concept-informative latent codes at intermediate denoising steps, probing whether the unlearned model’s deep parameters still retain the ability to reconstruct concept-related visual patterns.

Evaluation Metrics. Two complementary metrics are computed from the partial diffusion outputs:

- **Concept Retrieval Score (CRS)** measures the feature-level similarity between partially diffused outputs and the original concept domain. Specifically, CRS computes the cosine similarity in CLIP feature space between the output of the partial diffusion pipeline and the reference image. A higher CRS indicates that the concept’s visual features are more thoroughly removed from the model’s internal representations, meaning the unlearned model fails to reconstruct concept-related patterns even when provided with concept-informative latents. For genuinely unlearned models, CRS should be *high* (close to 1.0), reflecting that the model no longer encodes the target concept.
- **Concept Confidence Score (CCS)** quantifies the confidence of a domain classifier in assigning the partially diffused output to the pre-unlearning concept domain. A lower CCS indicates that the model’s internal representations no longer encode concept-discriminative features. For genuinely unlearned models, CCS should be *low* (close to 0.0).

Hyperparameters. Following the default settings in Sharma et al. (2024), we set the partial diffusion ratio to $\psi = 0.6$ and evaluate across a sweep of guidance scales $\eta \in \{1, 2, 4, 6, 8, 10\}$. We report the CCS and CRS values averaged over the guidance sweep to ensure robustness of the evaluation. For each target concept, 50 reference images are generated from the fully trained model using distinct random seeds, and the metrics are averaged across all reference images. The DDIM scheduler with 50 inference steps is used for both the forward noise injection and the reverse denoising process.

A.3 DATASET CONSTRUCTION

We describe the datasets used for the three unlearning tasks, together with the extended benchmark for large-scale multi-concept evaluation.

Object Unlearning. We evaluate multi-concept object unlearning on Imagenette (Howard et al., 2019), a ten-class subset of ImageNet containing easily recognizable categories: *tench*, *English springer*, *cassette player*, *chain saw*, *church*, *French horn*, *garbage truck*, *gas pump*, *golf ball*, and *parachute*. For each target class, we construct evaluation prompts using "a photo of a [class name]" and several variants such as "[class name] in a natural setting" and "a realistic image of [class name]". Following prior works (Gandikota et al., 2023; Chavhan et al., 2024; Deng et al., 2026), we generate 500 images per class with different random seeds to measure Unlearn Acc. Unlearn Acc is the top-1 accuracy of an ImageNet-pretrained ResNet50 on the images generated for the erased classes.

Explicit Content Unlearning. We evaluate explicit-content unlearning on the I2P dataset (Schramowski et al., 2023), which contains prompts likely to induce unsafe generations, including nudity, violence, self-harm, and shocking imagery. Following prior works (Zhang et al., 2024c; Srivatsan et al., 2025; Gong et al., 2024), we focus on erasing *nudity* as a representative sensitive concept. For each model, we generate one image per prompt, yielding 4,703 images in total. NudeNet is applied with a threshold of 0.6 under the default I2P setting. Generation quality is further evaluated on COCO-30K (Lin et al., 2014) using FID and CLIP score.

Artistic-Style Unlearning. For artistic-style unlearning, we follow the protocol of Chavhan et al. (2024); Deng et al. (2026); Gandikota et al. (2023) and erase five styles: *Van Gogh*, *Pablo Picasso*, *Rembrandt*, *Andy Warhol*, and *Claude Monet*. For each artist, 20 prompts are used to generate evaluation images, such as "A painting in the style of [artist name]" and "[subject] by [artist name]". We report three CLIP-based metrics: CLIP_a for alignment with the erased style, CLIP_c for generation quality on MS-COCO, and $\text{CLIP}_d = \text{CLIP}_c - \text{CLIP}_a$ as a measure of erasure selectivity averaged over the five target styles.

Extended Benchmark for Large-Scale Unlearning. To assess scalability beyond standard benchmarks, we additionally use the extended dataset of Deng et al. (2026), which expands the forgetting scope to 5, 10, 15, and 20 concepts from a broader pool of ImageNet classes and artistic styles. Each scale includes standardized prompts, evaluation splits, and concept-specific classifiers, enabling

controlled comparison as the number of erased concepts increases. We use this benchmark for the scalability study in Appendix B.1 (Fig. 6), which reports Unlearn Acc, ASR, CCS, and CRS against the number of forgotten concepts.

A.4 DETAILED SETTINGS OF AVCE

Verifiable Vulnerability Anchor. For each target concept, we run $K=10$ randomly initialized probe searches. Each search is optimized for 200 iterations with step size $\beta = 5 \times 10^{-3}$ using the Adam optimizer. The repulsion coefficient is set to $\tau = 0.5$, and the SA auditing strength to $\mu = 0.3$. Probe embeddings are initialized by adding Gaussian noise $\mathcal{N}(0, 0.1^2 I)$ to the original concept embedding $c_{o,i}$. Timesteps t are sampled uniformly from $\{1, \dots, 1000\}$ at each iteration, and noisy latents z_t are constructed from a fixed set of 5 reference latent codes. After the search, each probe is tested over 4 fresh draws of (z_t, t) : it is kept as successful if its noise-prediction error against the original-concept prediction is below half of the error obtained with the null prompt on the same draws; if no probe passes, the single best probe is kept. The principal direction $u_{i,1}$ is extracted via truncated SVD on the residual matrix Δ_i formed from the successful perturbations.

Cross-Attention Analytical Erasure. The Semantic Modulator uses $P=50$ calibration samples per concept, generated from the base model using diverse prompts. The Fisher stability score and Symmetric Uncertainty are computed with a temperature of $\sigma_{\text{temp}} = 5.0$ for the sigmoid mappings. The ridge coefficient is set to $\lambda = 0.1$ for the closed-form solution $W^* = Q_1 Q_2^{-1}$. The preservation set $\mathcal{R}$ contains the content-token and end-of-text embeddings of 50 neutral prompts describing everyday objects and scenes. The entire cross-attention editing is performed as a single one-shot matrix computation.

Self-Attention Analytical Shrinkage. The Visual Localizer extracts concept-specific features using a mid-process noisy latent at timestep $t=500$. The shrinkage coefficient is set to $\kappa = 1.0$. Shrinkage is applied to all SA key and value projection matrices across all UNet layers simultaneously.

Audit-Guided Refinement. The refinement phase alternates between CA and SA updates for 500 total iterations. We use the Adam optimizer with a learning rate of 1×10^{-5} and batch size of 1. The pathway-specific loss coefficients are set to $\lambda_{\text{ca}} = 0.5$ and $\lambda_{\text{sa}} = 0.5$, and the retain coefficient to $\lambda_r = 10$; each update draws one neutral prompt at random. Timesteps are uniformly sampled at each iteration. For multi-concept consolidation, we use an EMA decay of 0.99 for the difficulty scores $\tilde{\mathcal{L}}_i$ and sample $B=4$ concepts per mini-batch. Pairwise gradient conflicts are resolved via orthogonal projection following Eq. (14).

B MORE EXPERIMENTAL RESULTS

B.1 SCALABILITY OF AVCE

To evaluate scalability, we run object unlearning with an increasing number of forgotten concepts ($N \in \{5, 10, 15, 20\}$). Figure 6 shows that the baselines degrade as N grows: Unlearn Acc, Attack Success Rate (ASR), and Concept Confidence Score (CCS) rise, and Concept Retrieval Score (CRS) falls, which indicates interference between concepts and a loss of robust erasure at scale. AVCE degrades far more slowly. With the erasure grounded in the Verifiable Vulnerability Anchor and gradient conflicts resolved during multi-concept consolidation, it maintains low Unlearn Acc, ASR, and CCS and a high CRS even when 20 objects are erased simultaneously.

Computational overhead is a further concern in multi-concept unlearning. Sequential methods incur a training time that grows linearly with the number of target concepts ($\mathcal{O}(N)$). AVCE instead updates the model for all concepts concurrently through the consolidation strategy of Section 3.3. Table 7 reports the total unlearning time of AVCE for $N \in \{1, 3, 5, 10, 20\}$ on SD v1.5 and SDXL. Erasing a single concept takes approximately 6.02 minutes on SD v1.5 and 15.05 minutes on SDXL (measured on a single NVIDIA A100 GPU). Erasing 20 concepts simultaneously takes 6.74 and 16.86 minutes, respectively, an overhead of roughly 12%.

Table 7: Computational time (in minutes) of the proposed AVCE method across different numbers of concepts (N). Total time grows by approximately 12% from 1 to 20 concepts.

Model Architecture	Number of Forgotten Concepts (N)				
	$N = 1$	$N = 3$	$N = 5$	$N = 10$	$N = 20$
SD v1.5	6.02	6.18	6.32	6.55	6.74
SDXL	15.05	15.45	15.78	16.35	16.86

B.2 GENERALIZATION ACROSS ARCHITECTURES AND TASKS

We further evaluate AVCE on two other diffusion architectures, SD v1.5 (Table 8) and Flux 1.0 (Table 9 and Table 10).

Several baselines do not transfer across architectures. UCE, RECE, and SM achieve acceptable unlearning efficacy on SD v1.5 (Table 8; for example, 14.62% and 13.48% Unlearn Acc for UCE and RECE, respectively), but on the larger Flux 1.0 model their Unlearn Acc rises to around 75%–85% (Table 9). This suggests that existing closed-form and fine-tuning strategies have difficulty with the more entangled internal representations of newer, larger architectures.

AVCE performs consistently across architectures. On SD v1.5 it reaches an Unlearn Acc of 4.15%. On Flux 1.0, where every baseline keeps Unlearn Acc above 75%, AVCE lowers Unlearn Acc to 16.32% and keeps ASR below 19%. Table 10 shows the same pattern for explicit-content unlearning on Flux 1.0: AVCE keeps ASR between 13.4% and 16.5% under the three attacks, whereas every baseline exceeds 80%.

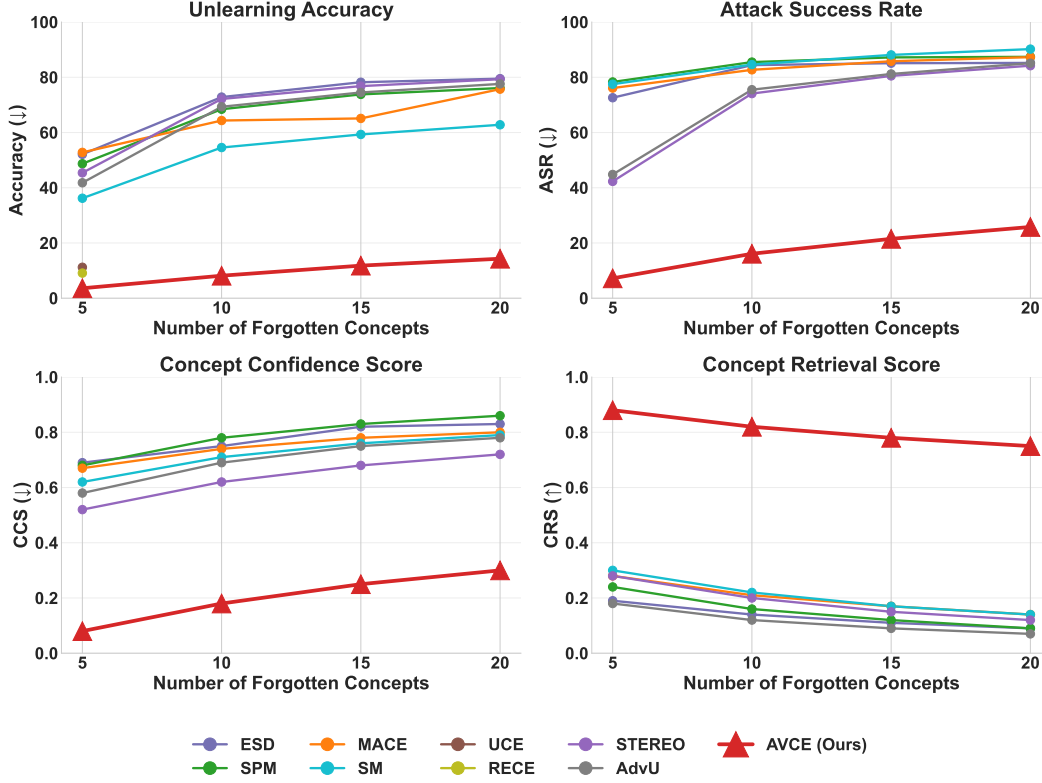


Figure 6: Scalability analysis of multi-concept object unlearning. We report Unlearn Acc (↓), Attack Success Rate (↓), Concept Confidence Score (↓), and Concept Retrieval Score (↑) across varying numbers of forgotten concepts ($N \in \{5, 10, 15, 20\}$). AVCE (red triangles) remains stable as N grows, whereas the baselines degrade.

Table 8: Quantitative results of multi-concept object unlearning on SD v1.5. Comparison of erasure accuracy, robustness, and generation quality across methods.

Metric	Method									
	SD v1.5	ESD	SPM	MACE	SM	UCE	RECE	STEREO	AdvU	AVCE
Unlearn Acc (↓)	87.23	77.83	68.43	64.34	30.45	14.62	13.48	42.15	49.32	4.15
CLIP (↑)	31.62	31.45	31.53	31.26	31.24	30.95	31.12	30.67	30.42	30.58
FID (↓)	11.92	13.87	14.14	14.53	14.76	15.82	15.46	15.02	15.39	14.87
ASR1 (↓)	99.2	84.7	81.2	78.5	76.8	74.5	71.2	68.2	69.7	13.5
ASR2 (↓)	97.6	88.3	85.6	82.1	81.5	79.8	77.1	73.8	75.6	16.7
ASR3 (↓)	98.2	92.2	89.7	87.4	86.4	85.2	82.5	80.4	81.3	17.2
CRS (↑)	–	0.14	0.16	0.21	0.38	0.35	0.42	0.20	0.32	0.82
CCS (↓)	–	0.85	0.78	0.74	0.61	0.54	0.47	0.62	0.59	0.18

Table 9: Quantitative results of multi-concept object unlearning on Flux 1.0. Comparison of erasure accuracy, robustness, and generation quality across methods.

Metric	Method									
	Flux	ESD	SPM	MACE	SM	UCE	RECE	STEREO	AdvU	AVCE
Unlearn Acc (↓)	89.45	86.24	83.51	79.82	81.25	77.43	75.68	82.35	80.54	16.32
CLIP (↑)	32.14	31.85	31.92	31.74	31.68	31.42	31.55	31.32	31.18	31.46
FID (↓)	11.05	12.95	13.24	13.58	13.72	14.85	14.52	14.15	14.36	13.92
ASR1 (↓)	99.5	92.1	89.4	85.2	87.6	84.3	81.7	79.5	80.1	15.2
ASR2 (↓)	98.7	94.5	91.8	88.6	90.1	87.9	85.4	83.2	84.8	17.4
ASR3 (↓)	99.1	96.3	94.2	92.1	93.4	91.5	88.6	87.4	88.2	18.6
CRS (↑)	–	0.22	0.24	0.29	0.31	0.19	0.23	0.27	0.18	0.67
CCS (↓)	–	0.76	0.72	0.68	0.65	0.74	0.70	0.63	0.67	0.29

Table 10: Quantitative results of explicit-content unlearning on Flux 1.0. Comparison of robustness (ASR1/2/3) and auditing scores (CRS/CCS) across methods.

Metric	Method									
	Flux	ESD	SPM	MACE	SM	UCE	RECE	STEREO	AdvU	AVCE
ASR1 (↓)	99.2	93.5	90.2	86.4	88.1	85.3	82.6	80.2	81.5	13.4
ASR2 (↓)	98.5	95.8	92.7	89.5	91.2	88.6	86.3	84.1	85.7	15.1
ASR3 (↓)	98.9	97.4	95.1	92.8	94.3	92.4	90.5	88.7	89.4	16.5
CRS (↑)	–	0.21	0.25	0.28	0.30	0.20	0.24	0.26	0.17	0.77
CCS (↓)	–	0.78	0.74	0.69	0.67	0.75	0.72	0.65	0.68	0.22

C CODE REFERENCE

To ensure fair and reproducible comparison, we evaluate our approach against the most comparable state-of-the-art unlearning and concept editing methods. Their official open-source implementations are listed below, and all baseline results in this paper are obtained using these repositories with recommended settings unless otherwise noted:

Unlearn Methods

- ESD: <https://github.com/rohitgandikota/erasing>
- SPM: <https://github.com/Con6924/SPM>
- MACE: <https://github.com/Shilin-LU/MACE>
- UCE: <https://github.com/rohitgandikota/unified-concept-editing>
- SM: <https://github.com/coulsonlee/Sculpting-Memory-ICCV-2025>
- RECE: <https://github.com/CharlesGong12/RECE>
- AdvUnlearn: <https://github.com/OPTML-Group/AdvUnlearn>
- STEREO: <https://github.com/koushiksrivats/robust-concept-erasing>

Attack & Auditing Methods

- CCS/CRS: <https://github.com/respailab/unlearning-or-concealment>
- P4D: <https://github.com/zhiyichin/P4D>
- Ring-A-Bell: <https://github.com/chiayi-hsu/Ring-A-Bell>
- UnlearnDiffAtk: <https://github.com/OPTML-Group/Diffusion-MU-Attack>

D LIMITATIONS

Our audit relies on a single protocol, the partial-diffusion audit of Sharma et al. (2024) with its CRS and CCS metrics; other probes of internal representations may expose traces that this protocol does not. Robustness is measured against three prompt-level attacks (Ring-A-Bell, UnlearnDiffAtk, and P4D). We do not evaluate attacks that adapt to the anchor construction, nor recovery of an erased concept by fine-tuning the edited model. The analytical edits act on attention projections and assume white-box access to the model weights.