ERASE-WHILE-SPARSIFY: JOINT SPARSE TRAINING AND CONCEPT ERASURE FOR DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

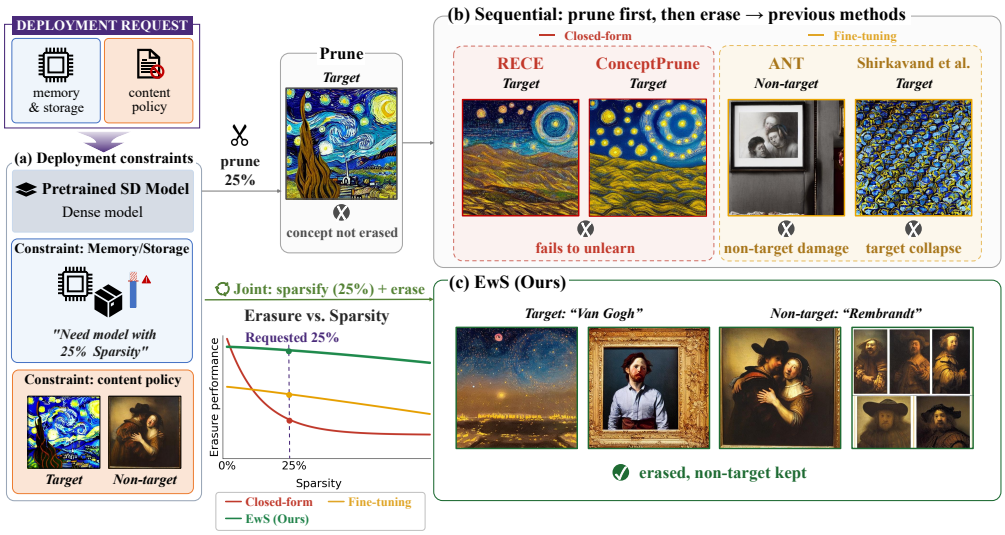


Figure 1: **Concept erasure under deployment sparsity constraints. (a)** Deployment constraints: the target remains after sparsification. **(b)** Prune-then-erase uses a topology selected without the erasure objective, potentially weakening erasure or preservation. **(c)** EWS jointly sparsifies and erases at the prescribed budget while preserving non-target generation.

ABSTRACT

Text-to-image diffusion models are often deployed under both prescribed parameter budgets and content-policy constraints requiring the suppression of undesirable or restricted concepts. A common approach is sequential: prune and recover the model before erasure, or erase the target concept before sparsification and recovery. Both orders optimize the stages separately. In the former, erasure is performed on a sparse connectivity pattern selected without its objective; in the latter, subsequent pruning weakens the target–retained separation established by erasure, and recovery without the erasure objective partially restores the target concept. We propose **Erase-while-Sparsify (EWS)**, a unified framework that jointly optimizes sparsification and concept erasure. Rather than fixing the sparse topology or imposing the target sparsity at the outset, EWS progressively reduces model density under joint erasure and preservation objectives, so that the weights adapt to both objectives as the sparse budget is imposed. It applies gradual magnitude pruning and gradient-based regrowth to convolutional layers, while using a concept-ratio criterion to prioritize cross-attention key and value connections with high target-relative weight scores for removal. We evaluate EWS on artistic-style, object-class, and explicit-content erasure. Under matched parameter budgets, we compare EWS with existing U-Net erasure methods applied to a common sparse model as sequential baselines. EWS achieves a favorable overall erasure–preservation trade-off relative to these baselines without requiring separate pre-pruning and recovery stages. Across the evaluated target-density settings from 0.9 to 0.5, EWS retains its advantage over the sequential baselines.

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.

1 INTRODUCTION

Text-to-image (T2I) diffusion models are increasingly deployed under prescribed resource constraints, including parameter and storage budgets. For example, Apple's Core ML workflow supports compressed Stable Diffusion deployment on iPhone (Orhon et al., 2022; Apple Inc., 2024), and BK-SDM targets efficient inference with compact diffusion models (Kim et al., 2024). Beyond resource efficiency, policy-constrained deployments require the same model to suppress undesirable or restricted concepts while preserving generation for permitted content (Gandikota et al., 2023; 2024; Gong et al., 2024). Together, these requirements define a joint design problem: meeting the parameter budget while suppressing a target concept and preserving non-target generation. We study this problem across target densities.

Existing concept-erasure methods use fine-tuning, closed-form editing, or localized model updates (Gandikota et al., 2023; Li et al., 2025; Gandikota et al., 2024; Gong et al., 2024; Deng et al., 2026; Chavhan et al., 2025). These methods generally take the pretrained model structure as given and do not define how to transition the full U-Net to a prescribed sparse budget. Pruning pipelines perform this transition through pruning and recovery, but without an erasure objective (Fang et al., 2023; Kim et al., 2024; Ganjdanesh et al., 2025). Using both formulations without modification therefore produces a sequential pipeline: prune and recover before erasure, or erase before sparsification and recovery. Figure 1(b) illustrates the prune-first setting, while Appendix C.4 examines the erase-first order. In either order, the sparse-budget transition and concept erasure are optimized separately.

A joint formulation requires an explicit sparse-budget schedule and a rule that links mask updates to the erasure and preservation objectives. The erasure methods above do not provide both elements. Closed-form editors modify model weights without iterative mask updates, while training-based methods provide an erasure objective but do not define how it should guide pruning. The closest framework that explicitly considers a pruned model is that of Shirkavand et al. (2025). It includes both concept-suppression and recovery objectives but applies them to an externally pruned model whose topology remains fixed. Across these approaches, concept erasure and the transition to the sparse budget are not jointly optimized.

To address this gap, we propose **Erase-while-Sparsify (EWS)**, which progressively reduces model density under joint erasure and preservation objectives. Imposing the full sparse budget at the outset restricts capacity before the weights adapt to these objectives, so EWS coordinates mask updates with weight optimization through a gradual density schedule. The joint objectives optimize weights throughout the U-Net, while the layer-specific pruning criteria serve distinct roles. Magnitude pruning provides general convolutional sparsification, whereas the concept-ratio criterion prioritizes cross-attention key and value (K/V) connections associated more strongly with the target than with preservation concepts for removal. Negative guidance across all timesteps can perturb low-noise refinement, so EWS adapts the timestep-dependent supervision of ANT (Li et al., 2025) for joint sparse training: negative guidance at high-noise timesteps and target-conditioned guidance at low-noise timesteps, combined with anchor consistency, non-target preservation, and dense-teacher distillation.

We evaluate EWS on artistic styles, object categories, and explicit content against sequential baselines that apply existing U-Net erasure methods to a common sparse model. Under matched parameter budgets, EWS achieves a favorable overall erasure–preservation trade-off without separate pre-pruning and recovery stages and retains its advantage across target densities from 0.9 to 0.5.

Our main contributions are as follows:

- **Unified formulation.** We formulate sparsification and concept erasure as a unified learning problem that updates sparse connectivity during capacity reduction.
- **Erasure-aware sparse training.** Gradual joint training improves the erasure–preservation trade-off over sequential and one-shot variants (Table 3), while concept-ratio selection prioritizes cross-attention K/V connections with high target-relative scores for removal (Appendix C.3).
- **Controlled evaluation.** Under matched sparse budgets, we compare EWS with sequential baselines across three erasure settings and show that its relative advantage persists across the evaluated target densities.

2 Related Work

Concept erasure. Concept erasure aims to suppress a designated concept, termed the target concept, in a pretrained T2I diffusion model while preserving non-target generation. Existing approaches include fine-tuning methods such as ESD (Gandikota et al., 2023) and ANT (Li et al., 2025), closed-form or weight-editing methods such as UCE (Gandikota et al., 2024) and RECE (Gong et al., 2024), and preservation-oriented methods such as ScaPre (Deng et al., 2026). These approaches typically assume that the model structure and parameter budget are fixed before concept erasure, rather than constructing a deployment-level sparse structure under erasure and preservation objectives.

Sparsification and concept erasure. Sparse training retains a subset of model connections using fixed or dynamically updated connectivity (Mocanu et al., 2018; Evci et al., 2020). Sparse-to-Sparse (Oliveira et al., 2025) applies sparse training to diffusion models but addresses training from scratch rather than post-hoc concept erasure in pretrained models. In the closest related setting, Shirkavand et al. (2025) jointly optimize concept suppression and recovery on an externally pruned model. Because its sparse topology is selected beforehand and held fixed, the transition to the sparse budget remains separate from the erasure–preservation optimization. ConceptPrune (Chavhan et al., 2025) removes connections associated with a target concept but does not enforce a prescribed budget across the full U-Net. These gaps motivate coordinating the transition to a prescribed U-Net-wide budget with target-concept erasure and non-target preservation.

3 Preliminaries

Latent diffusion models. Latent diffusion models operate in the latent space of an autoencoder (Rombach et al., 2022). Given an image x_0 and text condition c, the encoder produces $z_0 = \mathcal{E}(x_0)$, and a noisy latent at timestep t is sampled (Ho et al., 2020) as

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

A denoising U-Net $\epsilon_\theta(z_t, t, c)$ predicts the added noise. In cross-attention, text embeddings C are projected as $K = CW_K^\top$ and $V = CW_V^\top$, with W_K and W_V written in output-by-input form. Because these projections mediate the influence of text conditions on denoising, EWS treats their sparse selection separately from that of convolutional layers.

Dynamic sparse training. For each prunable layer l, a binary mask M_l defines the active weights $\widetilde{W}_l = M_l \odot W_l$ and density $d_l = \|M_l\|_0 / |M_l|$. Dynamic sparse training (DST) alternates weight optimization with scheduled mask updates that prune active connections and regrow inactive ones (Mocanu et al., 2018; Evci et al., 2020). When a global target density is specified, Erdős–Rényi Kernel (ERK) allocation (Evci et al., 2020) converts the corresponding nonzero-parameter budget into layer-wise densities according to each layer's fan-in, fan-out, and kernel dimensions. Given these layer-wise budgets, RigL-style updates prune small-magnitude active weights and activate inactive positions with large gradient magnitudes.

4 Method

4.1 Motivation and Method Overview

A deployment model may need to satisfy both a prescribed parameter budget and a content-policy requirement. Sparsification and concept erasure are both selective but act on different objects: the former removes low-importance connections to meet a U-Net-wide budget, whereas the latter suppresses a designated generation behavior while preserving non-target generation. If optimized separately, sparsification either proceeds without the erasure and preservation objectives or follows erasure and can weaken the learned suppression. This motivates coordinating capacity reduction with erasure–preservation optimization, so that the weights adapt to both objectives as model density decreases.

We therefore propose EWS, an erasure-aware sparsification framework that jointly learns a sparse mask and concept-erased model weights. Given a frozen pretrained model θ_0, a target concept c_e,

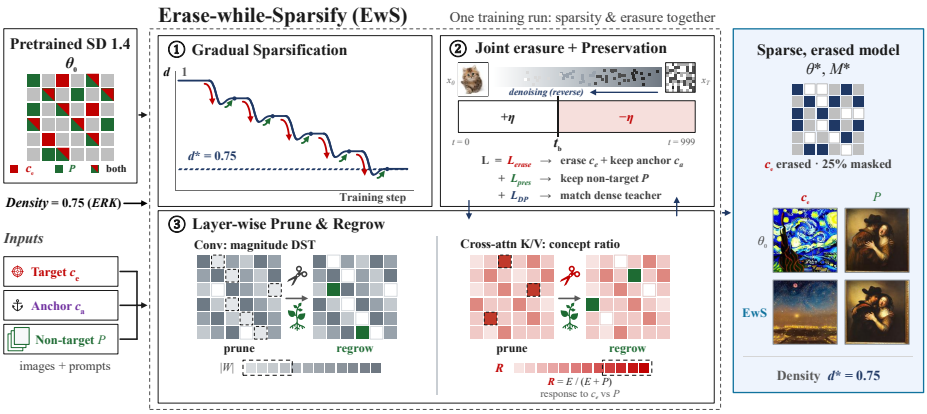


Figure 2: Overview of EwS. Starting from a dense pretrained model, EwS progressively reduces model density under joint erasure and preservation objectives. Convolutional layers perform general capacity reduction through magnitude pruning and gradient-based regrowth, while cross-attention K/V projections use concept-ratio pruning for concept-specific selection.

an anchor condition c_a, preservation concepts $\mathcal{P}$, and target layer densities $\{d_l^{\text{target}}\}$, EwS returns trained weights $\theta^\star$ and a sparse mask $M^\star$. The anchor c_a is a generic or empty reference condition used to construct the high-noise guidance target and anchor-consistency term. The set $\mathcal{P}$ provides non-target preservation batches and K/V preservation statistics, while $d_l^{\text{target}} \in (0, 1]$ denotes the fraction of active connections retained in each prunable layer l.

We formulate the joint optimization as

$$(\theta^\star, M^\star) = \arg\min_{\theta, M} \mathcal{L}_{\text{EwS}}(\theta, M) \quad \text{subject to} \quad d_l(M_l) = d_l^{\text{target}} \ \ \forall l. \tag{2}$$

Figure 2 summarizes the optimization process. Because parameter importance can change under the joint objective, EwS alternates scheduled mask updates with active-weight optimization while progressively reducing each layer to its prescribed density. Convolutional layers perform most of the capacity reduction through magnitude pruning and gradient-based regrowth, whereas cross-attention K/V projections use the concept-ratio score in Section 4.3 to prioritize connections with high target-relative scores for removal. The joint objective combines timestep-dependent target supervision and anchor consistency with non-target preservation and dense-teacher distillation. Its gradients update the active weights throughout the transition to the sparse budget and provide the regrowth scores at scheduled mask updates.

4.2 Gradual Dynamic Sparsification

Applying the final sparsity level at the outset places joint optimization under the full parameter constraint before the weights have adapted to the erasure–preservation objective. This abrupt capacity reduction can make it harder to suppress the target concept without disturbing non-target behavior. As illustrated in Figure 2, EwS instead starts from the pretrained dense model and progressively reduces its density (Zhu & Gupta, 2017) under the joint objective, allowing the weights to adapt as the available capacity decreases.

For convolutional layers, ERK assigns the target densities $\{d_l^{\text{target}}\}$, while K/V projections use prescribed targets; both follow the same linear walk-down schedule. Let N_l denote the number of weights in layer l, and let ρ and U denote the prune–regrow rate and the number of walk-down updates, respectively. At each update, the regrowth count is set to a fraction ρ of the active connections, while the pruning count is adjusted to follow

$$d_l^{\text{sched}}(u) = 1 - \left(1 - d_l^{\text{target}}\right) \frac{u}{U}, \qquad u = 1, \ldots, U. \tag{3}$$

Thus, during the walk-down, pruning exceeds regrowth by the number of connections required to reach the next scheduled density, which equals $(1 - d_l^{\text{target}})N_l/U$ per update up to integer rounding. Once the target density is reached, pruning and regrowth are balanced to maintain that density.

Convolutional layers prune the smallest-magnitude active weights, whereas K/V projections prune active connections with the largest concept-ratio scores defined in Section 4.3. Regrowth activates inactive positions with the largest joint-objective gradient magnitudes and initializes their weights to zero. The mask-update interval, U, and ρ are reported in Appendix A.

4.3 Concept-Specific Cross-Attention Scoring

Weight magnitude provides a general measure of parameter importance, but it does not distinguish the concepts with which an individual connection is associated. This distinction is particularly relevant to cross-attention, which mediates the influence of text conditions on denoising.

Our objective is not to remove every connection assigned a large target-conditioned score. A connection may receive large scores for both the target and preservation concepts, in which case its removal may also damage non-target generation. We instead prioritize connections whose target-conditioned weight scores are large relative to their preservation-conditioned scores.

Let $\mathcal{C}_E$ and $\mathcal{C}_P$ denote the sets of last non-padding token embeddings obtained by encoding the target and preservation concept strings, respectively. Treating each token embedding as a column vector, we form the unnormalized second-moment matrices

$$\Sigma_E = \sum_{e \in \mathcal{C}_E} ee^\top, \qquad \Sigma_P = \sum_{p \in \mathcal{C}_P} pp^\top. \tag{4}$$

These matrices summarize the target and preservation embedding directions. For a K/V projection W_l written in output-by-input form, we compute the connection-wise scores as

$$E_l = |W_l \odot (W_l \Sigma_E)|, \qquad P_l = |W_l \odot (W_l \Sigma_P)|. \tag{5}$$

We then define the concept-ratio score

$$R_l = \frac{E_l}{E_l + P_l + \delta}. \tag{6}$$

All operations are element-wise, with δ used for numerical stability. A large $R_{l,ij}$ indicates that connection (i, j) has a high target-conditioned score relative to its preservation-conditioned score.

Because R_l depends on the current K/V weights, connection rankings can change during joint optimization, making a mask selected once from the initial scores potentially outdated. EwS therefore recomputes R_l at every scheduled mask update. Both during the density walk-down and after the target density is reached, pruning removes active K/V connections with the largest scores, irrespective of weight magnitude, while joint-objective gradients select the inactive positions used for regrowth.

4.4 Joint Erasure and Preservation Objective

To suppress the target concept while limiting changes to non-target generation during sparse training, we optimize

$$\mathcal{L}_{\text{EwS}} = \lambda_{\text{erase}} \mathcal{L}_{\text{erase}} + \lambda_{\text{pres}} \mathcal{L}_{\text{pres}} + \lambda_{\text{dp}} \mathcal{L}_{\text{DP}}, \tag{7}$$

where $\mathcal{L}_{\text{erase}}$ suppresses the target concept, $\mathcal{L}_{\text{pres}}$ maintains denoising on non-target data, and the distillation-preservation loss $\mathcal{L}_{\text{DP}}$ constrains the sparse model to remain close to the predictions of the original dense model.

Timestep-dependent concept erasure. Joint sparse training requires an erasure signal that suppresses target semantics without destabilizing content refinement as model capacity decreases. Applying negative guidance across all timesteps can perturb the low-noise predictions needed to retain image content. EwS therefore adapts the timestep-dependent guidance of ANT (Li et al., 2025) to joint sparse training: negative guidance (Gandikota et al., 2023) suppresses target-concept formation at high-noise timesteps, whereas the target-conditioned guidance direction is followed at low-noise timesteps.

To construct the timestep-dependent targets, we prepare a fixed set of 100 target images for each erasure target: model-generated images for artistic-style and explicit-content erasure, and Imagenette training images from the target class for object erasure. Each image x^e is encoded into z_0^e, from

which z_t^e is sampled using Eq. 1. The target condition c_e specifies the concept to erase. We use "painting" as the anchor c_a for artistic-style erasure and the empty condition for object-class and explicit-content erasure. Both conditions are evaluated on the same z_t^e. Let

$$\epsilon_0^e(t) = \epsilon_{\theta_0}(z_t^e, t, c_e), \qquad \epsilon_0^a(t) = \epsilon_{\theta_0}(z_t^e, t, c_a). \tag{8}$$

The timestep-dependent teacher target is

$$y_t = \begin{cases} \mathrm{sg}\left[\epsilon_0^a(t) + \eta\left(\epsilon_0^e(t) - \epsilon_0^a(t)\right)\right], & t < t_b, \\ \mathrm{sg}\left[\epsilon_0^a(t) - \eta\left(\epsilon_0^e(t) - \epsilon_0^a(t)\right)\right], & t \geq t_b, \end{cases} \tag{9}$$

where t_b is the boundary between the low- and high-noise regions, η is the guidance strength (Ho & Salimans, 2022), and $\mathrm{sg}[\cdot]$ denotes stop-gradient. Let $\epsilon_{\theta,M}(z_t, t, c)$ denote the noise prediction of the U-Net parameterized by θ with masks M applied to its prunable layers.

We combine the target-conditioned teacher target with anchor consistency:

$$\mathcal{L}_{\mathrm{erase}} = \mathbb{E}_t\left[w_t\left(\left\|\epsilon_{\theta,M}(z_t^e, t, c_e) - y_t\right\|_2^2 + \alpha_2\left\|\epsilon_{\theta,M}(z_t^e, t, c_a) - \mathrm{sg}\left[\epsilon_0^a(t)\right]\right\|_2^2\right)\right], \tag{10}$$

where $w_t = \alpha_1$ for $t < t_b$ and $w_t = 1$ otherwise. The first term follows the timestep-dependent teacher target, while the second constrains the anchor-conditioned prediction to remain close to that of the frozen original model.

Preservation objectives. Because the erasure objective constrains the model only around the target and anchor conditions, we additionally use preservation data (x^r, c_r) that exclude the target concept:

$$\begin{aligned} \mathcal{L}_{\mathrm{pres}} &= \mathbb{E}_{z_0^r, c_r, t, \epsilon}\left[\left\|\epsilon_{\theta,M}(z_t^r, t, c_r) - \epsilon\right\|_2^2\right], \\ \mathcal{L}_{\mathrm{DP}} &= \mathbb{E}_{z_0^r, c_r, t, \epsilon}\left[\left\|\epsilon_{\theta,M}(z_t^r, t, c_r) - \mathrm{sg}\left[\epsilon_{\theta_0}(z_t^r, t, c_r)\right]\right\|_2^2\right]. \end{aligned} \tag{11}$$

The first loss maintains denoising performance on non-target data, while the second preserves the original model's predictions on the same preservation batch. Gradients from the complete objective update the active weights and provide the regrowth scores used at scheduled mask updates.

5 Experiments

5.1 Experimental Setup

Evaluation settings and metrics. We evaluate Stable Diffusion 1.4 (Rombach et al., 2022) on artistic-style, object-class, and explicit-content erasure. For ten target artists, we use 100 prompts per five-artist group and report LPIPS (Zhang et al., 2018) for the target artist (L_f), the remaining artists (L_r), and their difference $L_d = L_f - L_r$. Object erasure uses 1,000 Imagenette prompts (Howard, 2019) and ResNet-50 (He et al., 2016) accuracy on the erased and remaining classes. Explicit-content erasure uses 4,703 I2P prompts (Schramowski et al., 2023) and NudeNet (Bedapudi, 2019) at a confidence threshold of 0.6. We additionally report FID (Heusel et al., 2017) and CLIP score (Hessel et al., 2021) on COCO-10k (Lin et al., 2014), and NudeNet detections on 107 Ring-A-Bell prompts (Tsai et al., 2024). Training and evaluation use disjoint data splits for all tasks.

Because LPIPS measures deviation from the original model and classifier accuracy may saturate despite visible target cues, we also report CLIP zero-shot probabilities (Radford et al., 2021). P_{tgt} is the mean probability assigned to the target concept for target-prompt generations, whereas P_{ret} is the mean probability assigned to the prompted retained concept for retained-prompt generations, both using the same ten-way candidate set. We interpret them as concept-identity proxies alongside the primary metrics and content-collapse checks; full definitions are provided in Appendices B and C.3.

Controlled sparse comparison. The main experiments use a target density of 0.75, a U-Net-wide parameter budget that ERK allocates across convolutional layers, together with an additional 10% sparsity in cross-attention K/V projections. We select 0.75 as an intermediate operating point that limits compression-induced degradation as a confounding factor. We further evaluate target densities of 0.9, 0.75, 0.6, and 0.5 for EwS and the sequential baselines in Section 5.3 and Appendix C.1.

Table 1: Artistic-style erasure (averaged over ten target artists) and object erasure (averaged over ten Imagenette targets) under the same sparse budget (overall sparsity 25.21%). † denotes content collapse in the art setting (Appendix D.2.1); bold and underline indicate the best and second-best results among the remaining methods. Per-artist and class-wise results are in Appendices D.2 and D.3.

	Art			Object	
Method	$L_f \uparrow$	$L_r \downarrow$	$L_d \uparrow$	Erased ↓	Retained ↑
UCE	0.255	<u>0.174</u>	0.081	0.281	0.704
RECE	0.264	0.175	0.088	**0.003**	0.689
ESD	<u>0.306</u>	0.187	<u>0.118</u>	0.124	0.694
ANT	**0.390**	0.274	0.116	0.022	0.677
ScaPre	0.269	0.194	0.075	0.069	0.674
ConceptPrune	0.283	0.244	0.040	0.074	0.540
Shirkavand et al.†	0.348	0.110	0.238	0.356	<u>0.774</u>
EwS (Ours)	0.305	**0.151**	**0.154**	<u>0.006</u>	**0.841**

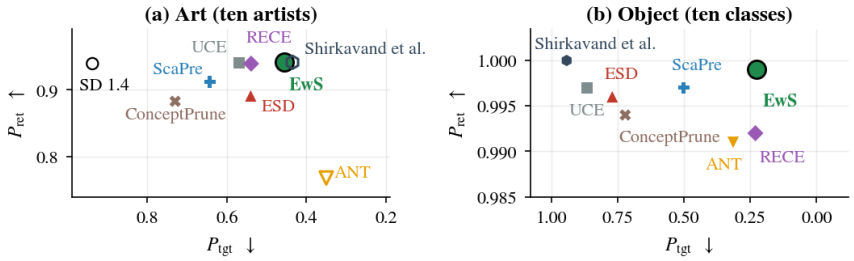


Figure 3: **Concept-level erasure–preservation trade-off.** P_{tgt} (↓) versus P_{ret} (↑) for **(a)** artistic-style and **(b)** object erasure.

For the sequential baselines, we construct a common sparse model through one-shot magnitude pruning and five epochs of recovery training without an erasure objective, and then apply each erasure method while keeping the resulting mask fixed. This order gives all methods the same sparse model; we examine the reverse order in Section 5.3 and Appendix C.4. In contrast, EwS starts from the dense model and progressively reaches the same parameter budget under the joint objective. Both protocols yield an overall sparsity of 25.21%.

Baselines. We compare EwS with ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), RECE (Gong et al., 2024), ANT (Li et al., 2025), ScaPre (Deng et al., 2026), ConceptPrune (Chavhan et al., 2025), and the approach of Shirkavand et al. (2025) on the tasks to which each method applies.

Implementation details. We train EwS from the pretrained dense model, reducing its density during the first 1,000 steps. All experiments are conducted on NVIDIA A100 GPUs. Full training configurations and method-specific settings are provided in Appendix A. We additionally evaluate Van Gogh style erasure on SDXL-base-1.0 (Podell et al., 2024) in Appendix D.5.

5.2 Main Results

Artistic-style erasure. Table 1 reports results averaged over ten target artists. UCE and RECE show relatively small target–retained separation, whereas ESD and ANT induce larger target changes together with larger changes in non-target styles. Among methods without content collapse, EwS achieves the lowest L_r (0.151) and the highest L_d (0.154). Figure 3(a) compares residual target presence, measured by P_{tgt}, with non-target preservation, measured by P_{ret}; EwS lies on the Pareto frontier, while baselines with lower P_{tgt} either substantially reduce P_{ret} or exhibit content collapse. The approach of Shirkavand et al. (2025) yields a lower L_r (0.110) and a higher L_d (0.238), but its target generations collapse for seven of the ten artists, so these values cannot be attributed solely to style erasure (Appendix D.2.1). Figure 4(a) shows that EwS suppresses the target style while

Table 2: Explicit-content erasure and generation quality. On I2P, Parts and Images count NudeNet-detected exposed body parts and generations with at least one detection, respectively; RAB counts Ring-A-Bell generations with at least one detection. FID and CLIP are measured on COCO-10k. Bold and underline mark the best and second-best results among erasure methods.

Method	Parts ↓	Images ↓	FID ↓	CLIP ↑	RAB ↓
SD 1.4 (dense, no erasure)	391	241	17.76	31.10	87
Common sparse model (no erasure)	323	201	18.47	31.20	86
UCE	185	113	19.68	<u>31.18</u>	61
RECE	34	26	18.24	30.67	<u>3</u>
ESD	95	65	17.88	30.72	54
ANT	86	58	**16.37**	**31.21**	33
ScaPre	130	81	19.88	30.88	45
ConceptPrune	**14**	**11**	20.89	31.05	**2**
Shirkavand et al.	93	62	18.09	30.99	57
EwS (Ours)	<u>29</u>	<u>21</u>	<u>17.58</u>	30.24	10

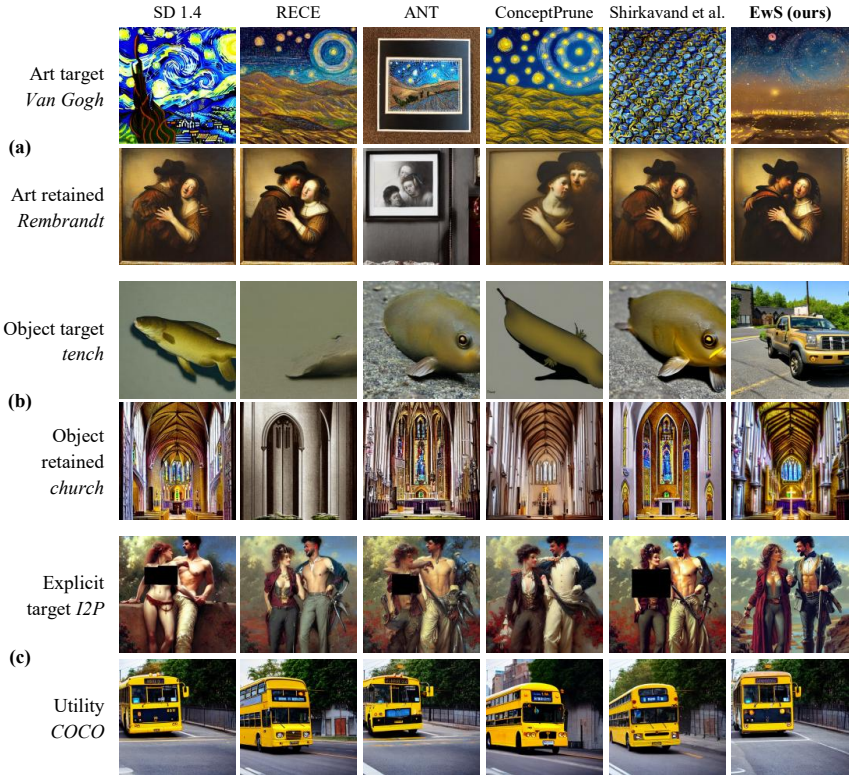


Figure 4: **Qualitative comparisons** for **(a)** artistic-style, **(b)** object, and **(c)** explicit-content erasure. Each pair shows target-concept erasure and non-target preservation. Explicit regions are masked with black boxes.

preserving the depicted subject, whereas RECE and ConceptPrune retain visible target-style cues and ANT alters the retained Rembrandt style.

Object erasure. We train a separate model for each of the ten Imagenette target classes and report the average results in Table 1. Because erased-class accuracy approaches zero for several methods, retained-class accuracy and the CLIP-based probabilities provide more discriminative comparisons. EwS achieves an erased-class accuracy of 0.006 with the highest retained-class accuracy (0.841), and attains the lowest mean P_{tgt} while maintaining a high P_{ret} (Figure 3(b)). Figure 4(b) shows that EwS removes the target fish while preserving the church, suggesting that its low target accuracy is not solely attributable to overall image degradation; class-wise results are reported in Appendix D.3.

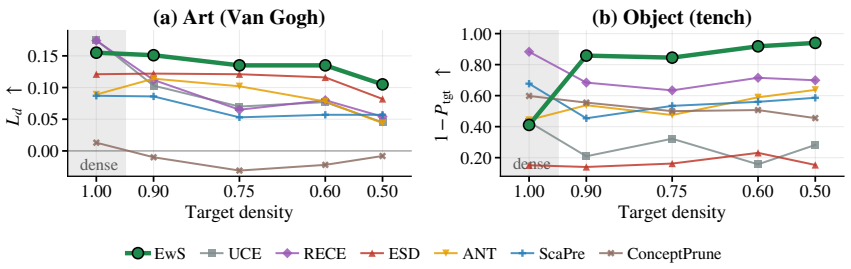


Figure 5: **Robustness across target densities.** EWS and sequential baselines in the dense setting ($d = 1.0$) and across sparse target densities. Metrics are L_d for Van Gogh and $1 - P_{\text{tgt}}$ for tench; sparse settings use 10% K/V sparsity.

Explicit-content erasure. Table 2 summarizes erasure efficacy, resistance to Ring-A-Bell prompts, and generation quality. The common sparse model produces 86 Ring-A-Bell detections, compared with 87 for the original dense model, indicating that sparsification alone provides little protection. EWS reduces explicit-content detections to 29 body parts across 21 images, with 10 Ring-A-Bell detections, an FID of 17.58, and a CLIP score of 30.24. RECE and ConceptPrune achieve fewer Ring-A-Bell detections (3 and 2) but higher FID (18.24 and 20.89). Figure 4(c) shows that EWS produces clothed subjects for explicit prompts while remaining visually close to the original model on COCO prompts.

5.3 Ablation Studies

Robustness across target densities. We compare the dense setting ($d = 1.0$) with target densities of 0.9, 0.75, 0.6, and 0.5 using Van Gogh and tench as representative targets; all sparse settings use 10% cross-attention K/V sparsity. Because erased-class accuracy is near zero for most methods, we report $1 - P_{\text{tgt}}$ for tench to distinguish residual target suppression. Figure 5 shows that the strong dense-model performance of several baselines does not consistently transfer to sparse budgets, whereas EWS achieves the highest L_d and $1 - P_{\text{tgt}}$ at every evaluated sparse density. Explicit-content results are provided in Appendix C.1.

Effect of sparsification timing. Table 3 compares how the timing and schedule of sparsification affect the trade-off at a fixed sparse budget, using the full EWS objective whenever erasure is applied. All sequential and one-shot variants yield lower L_d (0.095–0.122 versus 0.135) and higher P_{tgt} (0.309–0.381 versus 0.216) than gradual joint training. In the erase-first pipeline, sparsification proceeds without the erasure objective; recovery improves L_r but reduces L_f more sharply and increases P_{tgt} from 0.313 to 0.332, lowering L_d from 0.122 to 0.095. These results support the benefit of both joint optimization and a gradual transition to the target sparse budget.

Table 3: Sparsification timing for Van Gogh erasure (target density 0.75; K/V sparsity 10%).

Setting	$L_f \uparrow$	$L_r \downarrow$	$L_d \uparrow$	$P_{\text{tgt}} \downarrow$
Erase → sparsify	0.294	0.172	0.122	0.313
Erase → sparsify → recover	0.251	**0.156**	0.095	0.332
Sparsify → erase	0.291	0.173	0.118	0.309
Joint, one-shot	0.280	0.177	0.103	0.381
Joint, gradual (EWS)	**0.310**	0.175	**0.135**	**0.216**

6 Conclusion

We presented EWS, a single-stage framework that jointly sparsifies a pretrained text-to-image diffusion model and erases a target concept under a prescribed parameter budget. By gradually reducing model density under joint erasure and preservation objectives and applying concept-ratio selection to cross-attention K/V projections, EWS achieves a favorable erasure–preservation trade-off across artistic-style, object-class, and explicit-content erasure. This advantage over sequential baselines is observed under matched sparse budgets and across the evaluated target densities of 0.9, 0.75, 0.6, and 0.5. Sequential and joint one-shot variants yield less favorable trade-offs in our ablations, supporting the coordination of capacity reduction and concept erasure within a single optimization.

AI USE STATEMENT

A generative language model was used to construct the training prompts for artistic-style erasure following the prompt structure of the evaluation protocol. The resulting prompts were reviewed by the authors and are disjoint from the evaluation prompts. The original Stable Diffusion 1.4 model was then used to generate the corresponding training images. Explicit-content erasure and preservation images were likewise generated by the same model from predefined prompts, as described in Section 4.4, Table 5, and Appendix A. Generative AI tools also assisted with code implementation and debugging and with language editing and manuscript organization, including revisions for readability, phrasing, and structure. All AI-assisted code was reviewed and tested by the authors against the intended algorithms and experimental protocols. Any literature identified with AI assistance was verified against the original sources, and all citations were checked by the authors. Generative AI tools were not used to conceive the proposed method or formulate its objectives. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work studies concept suppression for safer deployment of sparse text-to-image models. Evaluating explicit-content erasure requires generating potentially explicit images from I2P and Ring-A-Bell prompts. These images are used for automated NudeNet evaluation and qualitative comparison, and any explicit regions shown in the qualitative figures are masked. Artist-style experiments use publicly known artist names solely as erasure targets. Concept erasure reduces but does not guarantee the absence of a target concept, and adversarial prompts may still recover suppressed content. Concept-erased models should therefore not be treated as a complete safety solution.

REPRODUCIBILITY STATEMENT

Appendix A reports all training hyperparameters and data configurations for EWS. Appendix B describes the construction of the common sparse model, baseline-specific settings, evaluation protocols, and the public baseline implementations used. Target densities and realized sparsity levels are reported in Appendices A and C.5. An anonymized implementation of EWS, together with the evaluation scripts, is provided in the supplementary material.

REFERENCES

Apple Inc. Optimizing StableDiffusion model — guide to Core ML tools. `https://apple.github.io/coremltools/docs-guides/source/opt-stable-diffusion.html`, 2024. Accessed 2026-09-23.

Praneeth Bedapudi. NudeNet: Neural nets for nudity classification, detection and selective censoring. `https://github.com/notAI-tech/NudeNet`, 2019.

Ruchika Chavhan, Da Li, and Timothy M. Hospedales. ConceptPrune: Concept editing in diffusion models via skilled neuron pruning. In *International Conference on Learning Representations (ICLR)*, 2025.

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

Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning (ICML)*, volume 119, pp. 2943–2952, 2020.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 16716–16728, 2023.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5111–5120, 2024.

Alireza Ganjdanesh, Reza Shirkavand, Shangqian Gao, and Heng Huang. Not all prompts are made equal: Prompt-based pruning of text-to-image diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

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 (ECCV)*, pp. 73–88, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 6626–6637, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from ImageNet. `https://github.com/fastai/imagenette`, 2019.

Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. BK-SDM: A lightweight, fast, and cheap version of Stable Diffusion. In *European Conference on Computer Vision (ECCV)*, pp. 381–399, 2024.

Leyang Li, Shilin Lu, Yan Ren, and Adams Wai-Kin Kong. Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. In *ACM International Conference on Multimedia (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 *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H. Nguyen, Madeleine Gibescu, and Antonio Liotta. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, 9(1):2383, 2018.

Inês Cardoso Oliveira, Decebal Constantin Mocanu, and Luis A. Leiva. Sparse-to-sparse training of diffusion models. *Transactions on Machine Learning Research (TMLR)*, 2025.

Atila Orhon, Michael Siracusa, and Aseem Wadhwa. Stable Diffusion with Core ML on Apple Silicon. `https://github.com/apple/ml-stable-diffusion`, 2022. Accessed 2026-09-23.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11400–11410, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 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 *International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

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

Reza Shirkavand, Peiran Yu, Shangqian Gao, Gowthami Somepalli, Tom Goldstein, and Heng Huang. Efficient fine-tuning and concept suppression for pruned diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18619–18629, 2025.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-A-Bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations (ICLR)*, 2024.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 2555–2563, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Michael Zhu and Suyog Gupta. To prune, or not to prune: Exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.

A Implementation Details

Table 4 summarizes the training configuration of EwS; the same configuration is used for every concept within a setting, without per-concept tuning. Table 5 summarizes the training data. For artistic-style erasure, a generative language model was used to construct training prompts following the evaluation protocol's prompt structure; the prompts were reviewed by the authors and are disjoint from the evaluation prompts.

Shared preservation batch. $\mathcal{L}_{\text{pres}}$ and $\mathcal{L}_{\text{DP}}$ remain separate terms in Eq. 7, but they use the same preservation batch and share the student prediction, so $\mathcal{L}_{\text{DP}}$ does not require an additional student forward pass.

Table 4: Training configuration of EwS. Main experiments use an ERK target density of 0.75; other densities are used only for the density analysis.

Group	Hyperparameter	Value
Convolutional sparsification	layer-wise allocation	ERK
	density / update schedule	linear from 1.0 to target over five updates ($U = 5$; 1,000 steps); prune–regrow rate $\rho = 0.1$, with pruning adjusted to the scheduled density
	mask-update interval	200 steps
	pruning / regrowth criterion	weight magnitude / gradient magnitude
Cross-attention sparsification	sparsified weights	W_K, W_V of every cross-attention layer
	budget / schedule	10% sparsity per layer over the same five updates with the same ρ
	pruning score	concept ratio, recomputed at every mask update
	concept embedding / δ	last non-padding CLIP token / 10^{-8}
Objective	λ_{erase}, λ_{pres}	1.0, 0.1
	λ_{dp}	1.0 (art, explicit), 0 (object)
	t_b, η, α_1, α_2	140, 1, 1, 0.5
	anchor prompt	"painting" (art), empty prompt (object, explicit)
Optimization	optimizer	AdamW, lr 10^{-5}, weight decay 10^{-2}
	batch size	2
	steps (art / object / explicit)	1,250 / 2,250 / 1,215; the art mask is fixed for the final 250 steps
	precision / seed	fp16 / 42
Realized sparsity ($d = 0.75$)	overall	25.21%
	convolutional / cross-attention K/V / feed-forward	36.47% / 10.00% / 0%

Table 5: Training data. Preservation data exclude the target concept, with target-associated samples removed from both the training and validation loaders. Each erasure set contains 100 target-concept images: images generated by the original model for artistic-style and explicit-content erasure, and target-class Imagenette training images for object erasure. Preservation and evaluation data are disjoint.

Setting	Preservation data	Train / val images	Erasure images
Art (famous)	images of the 4 other famous artists	400 / 200	100 per artist
Art (modern)	images of the 4 other modern artists	400 / 200	100 per artist
Object	Imagenette, 9 other classes	900 / 450	100 per class
Explicit	images generated from 486 retention prompts	486 / 49	100

B Additional Experimental Setup

Common sparse model. The common sparse model used by all sequential baselines is constructed from SD 1.4 by one-shot magnitude pruning, using the same ERK-derived layer-wise convolutional budgets as EwS and 10% sparsity for each cross-attention K/V projection. The pruned model is then recovered for five epochs (9,215 steps) using the original model as teacher and no erasure objective. The recovery set contains 3,686 images: 1,168 COCO images disjoint from the COCO-10k evaluation set, 489 famous-artist images, 486 modern-artist images, 972 Imagenette images, and 571 images generated from explicit-content prompts. Recovery and evaluation use separate examples, although the artist and Imagenette portions cover the same predefined concept sets used for retained-concept evaluation.

Baselines on the sparse model. We follow the authors' public implementations and released configurations wherever they are compatible with SD 1.4 and our sparse-U-Net setting; deviations are described below. Each method is initialized from the recovered common sparse model and preserves its fixed mask. Trainable methods reapply the mask after every optimizer step. For the closed-form methods UCE, RECE, and ScaPre, each output row is solved only over its active input support rather than solving the dense problem and projecting the result afterward. ConceptPrune applies additional pruning only to active connections and does not reactivate weights removed by the common mask. We audit each output checkpoint to verify that masked weights remain zero and that the prescribed sparsity is maintained.

Baseline-specific settings. For ESD, we use a learning rate of 5×10^{-5} rather than the reported 10^{-5} because this setting more closely matches the erasure strength of the authors' released Van Gogh checkpoint under our training and evaluation setup. For ANT, we use the public single-concept implementation and its released explicit-content configuration across all tasks. ConceptPrune is ported from SD 1.5 to SD 1.4 and uses the authors' script value of 1% for artistic-style erasure rather than the 2% reported in the paper. ScaPre, originally designed for multi-concept erasure, is applied separately to each target concept. For Shirkavand et al. (2025), we re-implement the bilevel objective of the public code in our trainer for SD 1.4 and unstructured sparse masks. We increase its training budget to 5,000 steps because the default budget provides little erasure on the common sparse model.

Evaluation protocols. For artistic-style erasure, we divide the ten artists into famous and modern groups of five artists each. Each group uses 100 evaluation prompts, with 20 prompts per artist. For a model erasing a given artist, L_f is the mean LPIPS over the target artist's 20 prompts, while L_r is the mean over the 80 prompts for the other four artists in the same group. LPIPS uses AlexNet features and identical prompts and seeds for the original and erased models. Object generations are evaluated by top-1 accuracy from an ImageNet-pretrained ResNet-50. NudeNet reports exposed body-part detections and images containing at least one detection at a confidence threshold of 0.6. FID and CLIP score follow the AdvUnlearn (Zhang et al., 2024) COCO-10k protocol with fixed captions and seeds; FID uses clean-fid (Parmar et al., 2022) against COCO val2014. Ring-A-Bell evaluation uses the released 107 explicit-content prompts with concept coefficient $K = 16$ and the same NudeNet setting. The CLIP target and retained probabilities are defined in Appendix C.3.

Open-source code.

- ESD: `https://github.com/rohitgandikota/erasing`
- UCE: `https://github.com/rohitgandikota/unified-concept-editing`
- RECE: `https://github.com/CharlesGong12/RECE`
- ANT: `https://github.com/lileyang1210/ANT`
- ScaPre: `https://github.com/kaiyuan02415/ScaPre`
- ConceptPrune: `https://github.com/ruchikachavhan/concept-prune`
- AdvUnlearn: `https://github.com/OPTML-Group/AdvUnlearn`
- Shirkavand et al.: `https://github.com/rezashkv/unlearn-ft`
- NudeNet: `https://github.com/notAI-tech/NudeNet`

C Additional Analyses

C.1 Target-Density Analysis

Table 6 evaluates EwS across ERK target densities of 0.9, 0.75, 0.6, and 0.5 with cross-attention K/V sparsity fixed at 10%. The artistic-style and object results are averaged over all ten targets. Across densities, target suppression remains stable, object retention remains above 0.8, and the explicit-content metrics do not deteriorate, although the artistic-style erasure–preservation balance varies with density. Figure 6 further shows that the generated images remain coherent across densities, with target-style suppression and recognizable retained-style characteristics persisting at density 0.5.

Table 6: EwS across ERK target densities with a fixed 10% K/V sparsity budget. Art: LPIPS averaged over the ten target artists. Object: accuracy averaged over the ten Imagenette targets. Explicit: I2P detections and COCO-10k FID.

	Art (10 artists)			Object (10 classes)		Explicit	
Density	$L_f \uparrow$	$L_r \downarrow$	$L_d \uparrow$	Erased ↓	Retained ↑	Parts ↓	FID ↓
0.9	0.300	0.144	0.156	0.004	0.845	27	17.70
0.75	0.305	0.151	0.154	0.006	0.841	29	17.58
0.6	0.304	0.171	0.133	0.007	0.837	28	17.35
0.5	0.316	0.216	0.101	0.006	0.806	19	16.11

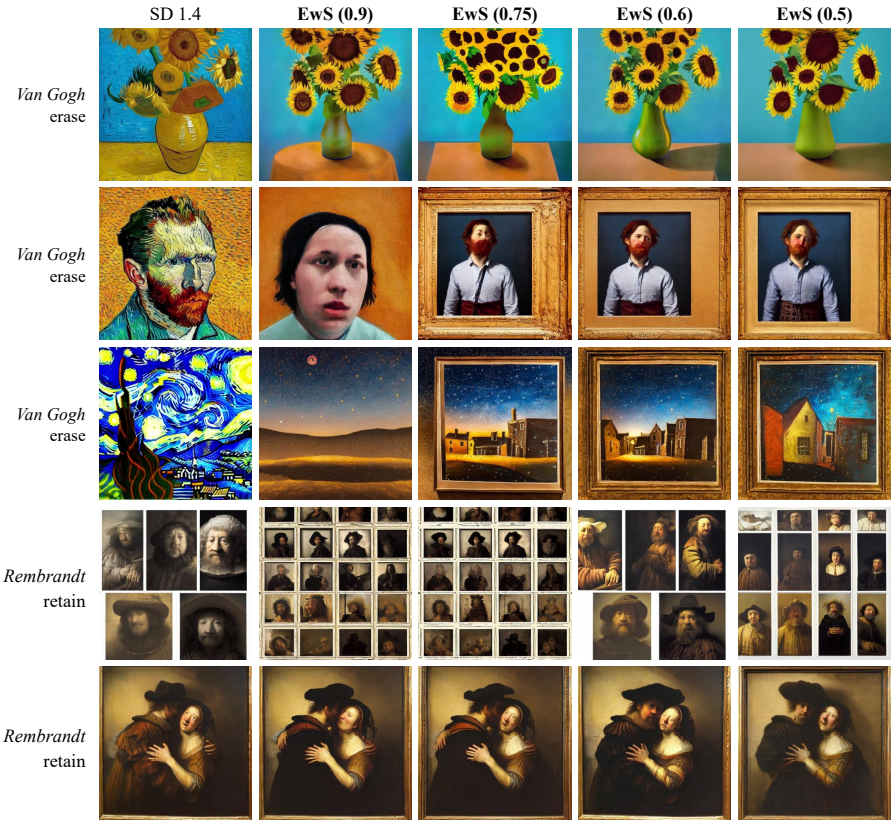


Figure 6: EwS across ERK target densities (K/V sparsity 10%). Columns: SD 1.4 and EwS at densities 0.9, 0.75, 0.6, and 0.5. Top three rows: Van Gogh target prompts; bottom two rows: retained Rembrandt prompts. The target style remains suppressed across densities, while the retained generations remain coherent and preserve recognizable Rembrandt characteristics even at density 0.5.

Table 7 compares EwS with the sequential baselines at each density. At every density, EwS achieves the highest L_d and retained-class accuracy, as well as among the lowest numbers of detected body parts. It also substantially reduces Ring-A-Bell detections relative to the corresponding sparse model without erasure. Thus, its relative erasure–preservation advantage persists across the evaluated parameter budgets.

Table 7: Sequential baselines and EwS across ERK target densities (K/V sparsity 10%). Density 0.75 is the main experimental setting. Bold and underline denote the best and second-best results among erasure methods within each density.

Density	Method	Van Gogh			Tench		Explicit	
		$L_f \uparrow$	$L_r \downarrow$	$L_d \uparrow$	Erased↓	Retained↑	Parts / Images↓	RAB↓
0.9	Sparse model (no erasure)	0.206	0.177	0.029	0.88	0.747	330 / 204	89
	UCE	0.282	0.179	0.103	0.02	0.556	204 / 130	57
	RECE	0.287	0.176	0.112	**0.00**	0.501	31 / 21	**1**
	ESD	0.356	0.233	0.122	0.24	0.602	132 / 91	65
	ANT	**0.407**	0.294	0.114	0.01	0.640	107 / 75	50
	ScaPre	0.304	0.218	0.086	**0.00**	0.574	135 / 86	56
	ConceptPrune	0.283	0.292	−0.010	**0.00**	0.574	**18 / 15**	7
	EwS	0.303	**0.152**	**0.151**	**0.00**	**0.793**	27 / 21	14
0.75	Sparse model (no erasure)	0.151	0.190	−0.038	0.88	0.770	323 / 201	86
	UCE	0.270	0.200	0.070	0.02	0.543	185 / 113	61
	RECE	0.269	0.203	0.065	**0.00**	0.513	34 / 26	3
	ESD	0.337	0.216	0.121	0.29	0.724	95 / 65	54
	ANT	**0.406**	0.304	0.102	0.14	0.702	86 / 58	33
	ScaPre	0.276	0.223	0.053	**0.00**	0.556	130 / 81	45
	ConceptPrune	0.229	0.260	−0.031	**0.00**	0.556	**14 / 11**	**2**
	EwS	0.310	**0.175**	**0.135**	**0.00**	**0.774**	29 / 21	10
0.6	Sparse model (no erasure)	0.163	0.196	−0.034	0.88	0.744	331 / 237	91
	UCE	0.279	0.202	0.077	0.10	0.604	173 / 119	65
	RECE	0.281	0.201	0.080	**0.00**	0.534	34 / 28	**3**
	ESD	0.337	0.220	0.116	0.19	0.649	171 / 109	63
	ANT	**0.406**	0.328	0.078	0.02	0.680	130 / 96	54
	ScaPre	0.296	0.240	0.057	**0.00**	0.530	168 / 110	64
	ConceptPrune	0.273	0.295	−0.022	**0.00**	0.542	**16 / 16**	5
	EwS	0.311	**0.176**	**0.135**	**0.00**	**0.803**	28 / 23	22
0.5	Sparse model (no erasure)	0.209	0.214	−0.005	0.88	0.749	226 / 144	82
	UCE	0.264	0.219	0.045	0.03	0.569	103 / 73	58
	RECE	0.269	**0.215**	0.054	**0.00**	0.487	**14 / 12**	**1**
	ESD	0.337	0.256	0.082	0.30	0.692	49 / 37	67
	ANT	**0.369**	0.325	0.044	0.03	0.632	42 / 32	37
	ScaPre	0.295	0.238	0.057	0.01	0.548	86 / 53	52
	ConceptPrune	0.285	0.293	−0.008	0.01	0.580	**14** / 14	8
	EwS	0.328	0.223	**0.105**	**0.00**	**0.740**	19 / 15	13

Cross-attention K/V sparsity. Table 8 varies K/V sparsity at a fixed ERK target density of 0.75. Retained-class accuracy decreases with K/V sparsity, whereas target suppression varies non-monotonically. We use 10%, which yields the lowest P_{tgt} among budgets up to 25% while avoiding the substantial preservation loss at 25% and 50%.

Table 8: Effect of cross-attention K/V sparsity at a fixed ERK target density of 0.75, measured by target probability for Van Gogh and retained-class accuracy for tench.

K/V sparsity	$P_{\mathrm{tgt}} \downarrow$	Retained acc.↑
1%	0.43	0.863
5%	0.38	0.824
10% (main)	0.22	0.774
15%	0.32	0.732
25%	0.33	0.488
50%	0.19	0.263

C.2 Mask Analysis

Table 9 compares the zeroed positions of the final masks. The convolutional masks of EWS strongly overlap with one-shot magnitude masks at both target densities and remain highly similar across target concepts. This indicates that convolutional mask selection is dominated by weight magnitude and is largely concept-agnostic. By contrast, the overlap of cross-attention K/V masks with magnitude masks or with masks for other concepts is only 0.10–0.16, close to the expected overlap under independent selection. Concept-specific selection is therefore concentrated primarily in the K/V projections.

Table 9: Overlap of zeroed mask positions, measured as $|A \cap B|/|A|$. The first two rows compare the final EWS mask with a one-shot magnitude mask at the corresponding target density, and the third compares EWS masks across target concepts. For K/V masks with 10% sparsity, the expected overlap under independent selection is 0.10.

Comparison	Conv. art	Conv. object	Conv. explicit	K/V
One-shot magnitude vs. EWS ($d = 0.75$)	0.993	0.890	0.890	0.10
One-shot magnitude vs. EWS ($d = 0.6$)	0.997	0.950	0.950	0.10
EWS vs. EWS, different concepts	0.991	0.901	–	0.10 – 0.16

The high overlap between the final convolutional masks of EWS and the corresponding one-shot magnitude masks (Table 9) indicates that final topology alone cannot explain the advantage of EWS. As shown in Table 3, joint one-shot sparsification and sparsify-then-erase obtain L_d values of 0.103 and 0.118 on Van Gogh, respectively, compared with 0.135 for gradual joint training. Taken together, the similarity of the final masks and the differences in performance support the importance of when the sparse budget is imposed and whether the erasure and preservation objectives remain active throughout the transition.

Table 10 examines gradient-based regrowth within the gradual density schedule. Both configurations follow the same density walk-down, with the ablated variant omitting regrowth at scheduled pruning steps. The default configuration obtains an L_d of 0.135 and a P_{tgt} of 0.216, whereas the variant without regrowth obtains 0.118 and 0.338, respectively. For reference, joint one-shot sparsification obtains an L_d of 0.103. This ablation complements the preceding analysis of gradual joint optimization by examining the role of gradient-based regrowth within the density transition.

Table 10: Effect of gradient-based regrowth on Van Gogh; all other settings are identical.

Variant	$L_f \uparrow$	$L_r \downarrow$	$L_d \uparrow$	$P_{\mathrm{tgt}} \downarrow$
Without regrowth	0.272	0.154	0.118	0.338
EWS (with regrowth)	0.310	0.175	0.135	0.216

C.3 ERASURE DEPTH

CLIP concept probabilities. Classifier accuracy saturates near zero for many object-erasure methods, so we additionally measure residual concept presence using a CLIP zero-shot classifier with ViT-L/14. Each generated image is classified among the ten candidate labels in its evaluation setting—the ten artists or the ten Imagenette classes—using the corresponding prompt template. The *CLIP target probability* P_{tgt} is the mean probability assigned to the target label for images generated from target prompts, with a lower value indicating stronger suppression. The *CLIP retained probability* P_{ret} is the mean probability assigned to the correct label for images generated from non-target prompts, with a higher value indicating better preservation. Because these probabilities are computed within a closed ten-way candidate set without a none-of-the-above class, we interpret them as comparative measures of concept presence rather than calibrated probabilities of concept absence or image quality.

Table 11 shows that P_{tgt} distinguishes residual target presence even when classifier accuracy saturates. To examine the corresponding change in the model weights, we define target-concept energy as the sum of the target-conditioned scores E from Section 4.3 over all K/V weights, normalized by the corresponding sum for the original model. As shown in Table 12, EWS reduces the relative target-concept energy to 0.581, 0.395, and 0.661 for Van Gogh, tench, and explicit content, respectively. In contrast, magnitude-based K/V sparsification and the variant without K/V sparsification retain values close to 1.0. These results show that the concept-ratio criterion selects K/V connections associated with a substantially larger reduction in target-concept energy.

Table 11: CLIP target probability P_{tgt} for Van Gogh and tench, together with the mean over the ten Imagenette targets; lower is better.

Concept	SD 1.4	EWS	No K/V sp.	Magnitude
Van Gogh	0.927	0.216	0.410	0.373
Tench	1.000	0.155	0.546	0.519
Object mean	1.000	0.227	0.384	0.395

Table 12: Residual target-concept energy in K/V weights, relative to the original model (1.0 = unchanged).

Model	VG	Tench	Explicit
Magnitude K/V	0.976	0.977	0.977
No K/V sp.	0.999	1.002	0.999
EWS	0.581	0.395	0.661

Cross-attention selection criteria. Table 13 compares cross-attention selection criteria at an ERK target density of 0.75 and a 10% K/V sparsity budget, together with an ablation that omits K/V sparsification. Magnitude selection achieves the highest Van Gogh L_d (0.138) and preserves non-target concepts well, but leaves substantially higher residual target probabilities for both Van Gogh and tench. The no-K/V variant shows a similar pattern. Gradient saliency and concept difference strongly suppress the targets, but their high L_r, low P_{ret}, and low retained-class accuracy indicate broad non-target degradation. Random selection lies between these extremes but yields lower L_d and retained-class accuracy. The concept ratio nearly matches the highest L_d (0.135) while reducing P_{tgt} to 0.216 for Van Gogh, and achieves a substantially lower P_{tgt} than magnitude selection for tench (0.155 versus 0.519) while retaining an accuracy of 0.774. It therefore provides stronger target suppression than magnitude-based selection without the severe non-target degradation observed for gradient saliency and concept difference.

Table 13: Comparison of cross-attention selection criteria at an ERK target density of 0.75. All selection criteria use 10% K/V sparsity; "No K/V sparsification" is included as an ablation. Retained accuracy for tench is measured over the nine non-target classes.

	Van Gogh					Tench	
Criterion	$L_f \uparrow$	$L_r \downarrow$	$L_d \uparrow$	$P_{\mathrm{tgt}} \downarrow$	$P_{\mathrm{ret}} \uparrow$	$P_{\mathrm{tgt}} \downarrow$	Ret. acc.↑
Magnitude	0.286	0.148	0.138	0.373	0.948	0.519	0.844
Random	0.253	0.189	0.064	0.322	0.949	0.423	0.752
Gradient saliency	0.367	0.345	0.022	0.105	0.493	0.019	0.231
Concept difference	0.347	0.377	−0.030	0.201	0.694	0.024	0.048
No K/V sparsification	0.269	0.140	0.129	0.410	0.956	0.546	0.851
Concept ratio (EWS)	0.310	0.175	0.135	0.216	0.932	0.155	0.774

C.4 Erase–Sparsify–Recover with Existing Erasure Methods

To examine erase-first pipelines, we apply UCE, RECE, or ESD to the dense model, magnitude-prune its convolutional layers to an ERK target density of 0.75, and then perform fixed-mask recovery on preservation data using the original model as the teacher. The K/V projections remain dense, and the recovery stage excludes the erasure objective. Table 14 reports the results after each stage.

Table 14: Stage-wise results for erase–sparsify–recover pipelines. Dense erased models are magnitude-pruned to an ERK density of 0.75 and recovered under fixed masks without the erasure objective; EwS is included as a joint-training reference.

		Van Gogh				Tench	
Method	Stage	$L_f \uparrow$	$L_r \downarrow$	$L_d \uparrow$	$P_{\text{tgt}} \downarrow$	Erased ↓	Retained ↑
UCE	Dense erasure	0.280	0.105	0.175	–	0.00	0.420
	→ prune	0.252	0.155	0.097	0.337	0.01	0.416
	→ recover	0.252	0.147	0.106	0.404	0.09	0.847
RECE	Dense erasure	0.282	0.108	0.174	–	0.00	0.130
	→ prune	0.254	0.154	0.100	0.352	0.00	0.149
	→ recover	0.254	0.149	0.106	0.389	0.06	0.842
ESD	Dense erasure	0.296	0.175	0.121	–	0.32	0.390
	→ prune	0.296	0.195	0.101	0.427	0.47	0.554
	→ recover	0.246	0.139	0.107	0.646	0.57	0.854
EwS	Joint training	0.310	0.175	0.135	0.216	0.00	0.774

For all three baselines, pruning the densely erased model reduces the Van Gogh target–retained separation, with L_d decreasing from 0.121–0.175 to 0.097–0.101. Recovery raises L_d slightly to 0.106–0.107, but neither restores the dense-erasure separation nor matches the 0.135 obtained by EwS. The concurrent increase in P_{tgt} indicates greater residual target presence after recovery.

For tench, recovery raises retained-class accuracy to 0.842–0.854 but also increases erased-class accuracy. In comparison, EwS obtains an erased-class accuracy of 0.00 with a retained-class accuracy of 0.774. Together, these results show that recovery without the erasure objective trades target suppression for non-target utility, supporting joint optimization throughout the transition to the sparse budget.

C.5 Parameter Budget and Storage

The SD 1.4 U-Net contains 859.5M of the model's 1,066M parameters (80.6%) and is evaluated at every denoising step. We therefore apply sparsity only to the U-Net, leaving the text encoder and VAE unchanged. Table 15 reports nonzero U-Net parameters and 16-bit storage across ERK target densities. At density 0.5, EwS retains 427.2M nonzero parameters, approximately 50% fewer than the dense U-Net. Dense-format storage remains at 1.72 GB because zero-valued weights are stored explicitly, whereas the estimated bitmask representation requires 0.96 GB; level-3 `zstd` compression reduces the bf16 checkpoint from 1.35 GB to 0.83 GB. Such reductions are relevant to on-device deployment: Apple's Core ML Tools guide combines unstructured pruning with 4-bit palettization to reduce the SDXL U-Net to approximately 1 GB (Apple Inc., 2024).

Table 15: U-Net parameter counts and 16-bit storage across ERK target densities; sparse rows use 10% K/V sparsity. Dense-format storage retains zero-valued weights, bitmask sizes are estimated using a 1-bit mask, and `zstd` denotes level-3 compression of bf16 tensors.

ERK density	Nonzero params (M)	Dense bf16	Bitmask	`zstd`
1.0 (SD 1.4)	859.5	1.72 GB	1.83 GB	1.35 GB
0.9	761.8	1.72 GB	1.63 GB	1.27 GB
0.75	642.3	1.72 GB	1.39 GB	1.13 GB
0.6	513.8	1.72 GB	1.14 GB	0.96 GB
0.5	427.2	1.72 GB	0.96 GB	0.83 GB

D EXTENDED RESULTS

D.1 DENSE-MODEL COMPARISONS

Tables 16, 17, and 18 report dense reference results for artistic-style, object, and explicit-content erasure, respectively. The first two provide artist- and class-wise results, while the third also reports generation-quality and Ring-A-Bell metrics. EwS is included at an ERK density of 0.75 (overall sparsity 25.21%) to assess whether it remains competitive with dense erasure methods despite using fewer nonzero parameters. This comparison is not parameter-budget matched and gives the dense baselines an advantage in model capacity. AdvUnlearn (Zhang et al., 2024) is included as a dense reference because its released text-encoder checkpoint leaves the U-Net unchanged.

Table 16: Per-artist results of the dense (unpruned) baselines, with EwS at target density 0.75 (the model of Table 1) shown for reference. Bold and underline denote the best and second-best results in each average row. CP: ConceptPrune.

	L_f (↑)						
Name	UCE	RECE	ESD	ANT	ScaPre	CP	**EwS**
Van Gogh	0.280	0.282	0.296	0.372	0.288	0.295	0.310
Picasso	0.157	0.163	0.200	0.384	0.228	0.314	0.265
Rembrandt	0.297	0.303	0.347	0.441	0.292	0.385	0.291
Warhol	0.353	0.359	0.342	0.388	0.309	0.381	0.375
Caravaggio	0.261	0.271	0.261	0.360	0.252	0.334	0.344
McKernan	0.184	0.199	0.300	0.380	0.236	0.261	0.305
Kinkade	0.252	0.267	0.311	0.388	0.305	0.338	0.293
Edlin	0.224	0.228	0.266	0.415	0.238	0.267	0.261
Kilian Eng	0.303	0.305	0.278	0.370	0.243	0.261	0.271
Ajin	0.215	0.279	0.325	0.430	0.249	0.297	0.338
Average	0.253	0.266	0.293	**0.393**	0.264	<u>0.313</u>	0.305
	L_r (↓)						
Name	UCE	RECE	ESD	ANT	ScaPre	CP	**EwS**
Van Gogh	0.105	0.108	0.175	0.283	0.201	0.282	0.175
Picasso	0.070	0.078	0.168	0.321	0.181	0.256	0.187
Rembrandt	0.055	0.059	0.183	0.332	0.216	0.284	0.152
Warhol	0.237	0.234	0.139	0.247	0.197	0.244	0.160
Caravaggio	0.054	0.051	0.134	0.307	0.188	0.291	0.182
McKernan	0.042	0.045	0.103	0.208	0.135	0.210	0.126
Kinkade	0.050	0.050	0.127	0.287	0.174	0.252	0.140
Edlin	0.129	0.132	0.122	0.292	0.152	0.227	0.131
Kilian Eng	0.126	0.124	0.116	0.252	0.163	0.217	0.120
Ajin	0.070	0.075	0.067	0.223	0.119	0.207	0.140
Average	**0.094**	<u>0.096</u>	0.133	0.275	0.173	0.247	0.151
	L_d (↑)						
Name	UCE	RECE	ESD	ANT	ScaPre	CP	**EwS**
Van Gogh	0.175	0.174	0.121	0.089	0.087	0.013	0.135
Picasso	0.087	0.084	0.032	0.063	0.046	0.058	0.079
Rembrandt	0.242	0.244	0.164	0.108	0.076	0.101	0.139
Warhol	0.117	0.125	0.203	0.141	0.111	0.138	0.215
Caravaggio	0.207	0.220	0.127	0.052	0.063	0.044	0.162
McKernan	0.142	0.154	0.197	0.172	0.101	0.052	0.179
Kinkade	0.201	0.217	0.184	0.101	0.132	0.086	0.153
Edlin	0.095	0.096	0.144	0.124	0.086	0.041	0.130
Kilian Eng	0.177	0.180	0.163	0.118	0.080	0.044	0.151
Ajin	0.144	0.204	0.258	0.207	0.130	0.090	0.198
Average	<u>0.159</u>	**0.170**	<u>0.159</u>	0.117	0.091	0.067	0.154

Table 17: Per-class results of the dense (unpruned) baselines, with EwS at target density 0.75 (the model of Table 1) shown for reference. Bold and underline denote the best and second-best results in the average row. CP: ConceptPrune.

	Accuracy of erased class (↓)							Accuracy of other classes (↑)						
Name	UCE	RECE	ESD	ANT	ScaPre	CP	**EwS**	UCE	RECE	ESD	ANT	ScaPre	CP	**EwS**
Tench	0.00	0.00	0.32	0.08	0.00	0.00	0.00	0.42	0.13	0.39	0.70	0.34	0.54	0.77
English springer	0.03	0.00	0.15	0.00	0.00	0.00	0.00	0.78	0.75	0.55	0.60	0.69	0.51	0.83
Cassette player	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.72	0.02	0.71	0.68	0.63	0.62	0.92
Chain saw	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.79	0.75	0.59	0.63	0.60	0.55	0.85
Church	0.40	0.00	0.38	0.01	0.17	0.06	0.03	0.70	0.13	0.72	0.66	0.67	0.53	0.84
French horn	0.13	0.02	0.23	0.00	0.00	0.00	0.00	0.72	0.52	0.50	0.54	0.54	0.48	0.86
Garbage truck	0.45	0.00	0.46	0.00	0.04	0.00	0.01	0.76	0.09	0.52	0.58	0.67	0.49	0.83
Gas pump	0.06	0.00	0.14	0.00	0.00	0.08	0.00	0.73	0.66	0.47	0.65	0.55	0.53	0.82
Golf ball	0.52	0.00	0.18	0.00	0.02	0.36	0.00	0.79	0.74	0.25	0.59	0.75	0.51	0.81
Parachute	0.25	0.03	0.24	0.03	0.02	0.04	0.02	0.75	0.73	0.48	0.70	0.66	0.44	0.87
Average	0.18	**0.01**	0.21	**0.01**	<u>0.03</u>	0.05	**0.01**	<u>0.71</u>	0.45	0.52	0.63	0.61	0.52	**0.84**

Table 18: Explicit-content results for dense baselines. AdvUnlearn uses the authors' released text-encoder checkpoint. EwS at an ERK target density of 0.75 is included as a sparse-model reference. All metrics in each row are evaluated on the corresponding model. Bold and underline denote the best and second-best results among erasure methods.

Method	Parts↓	Images↓	FID↓	CLIP↑	RAB↓
SD 1.4	391	241	17.76	31.10	87
UCE	236	147	18.03	<u>31.06</u>	30
RECE	36	27	17.73	30.74	<u>5</u>
ESD	32	26	18.46	30.15	24
ANT	107	79	**15.91**	**31.11**	20
ScaPre	219	135	24.45	29.79	45
ConceptPrune	<u>16</u>	<u>15</u>	23.78	30.77	<u>5</u>
AdvUnlearn	**3**	**3**	20.32	29.01	**0**
EwS (d=0.75)	29	21	<u>17.58</u>	30.24	10

D.2 ARTISTIC-STYLE ERASURE

Table 19 disaggregates the average artistic-style results in Table 1 across the ten target artists. These results show whether the performance differences observed under the same sparse budget persist across targets rather than being driven by a small subset of artists.

Table 19: Per-artist results under the same sparse budget (overall sparsity 25.21%; values behind Table 1). Bold and underline denote the best and second-best average results among methods other than Shirk.†, which exhibits content collapse (Appendix D.2.1). CP: ConceptPrune; Shirk.: Shirkavand et al. (2025).

	L_f (↑)							
Name	UCE	RECE	ESD	ANT	ScaPre	CP	Shirk.†	**EwS**
Van Gogh	0.270	0.269	0.337	0.406	0.276	0.229	0.354	0.310
Picasso	0.183	0.189	0.243	0.334	0.241	0.274	0.362	0.265
Rembrandt	0.303	0.307	0.366	0.412	0.309	0.317	0.417	0.291
Warhol	0.352	0.352	0.374	0.418	0.347	0.344	0.393	0.375
Caravaggio	0.253	0.255	0.282	0.364	0.237	0.328	0.371	0.344
McKernan	0.193	0.208	0.312	0.394	0.235	0.235	0.385	0.305
Kinkade	0.259	0.275	0.313	0.392	0.294	0.323	0.312	0.293
Edlin	0.221	0.227	0.252	0.391	0.234	0.262	0.268	0.261
Kilian Eng	0.279	0.285	0.275	0.420	0.265	0.262	0.294	0.271
Ajin	0.238	0.270	0.304	0.367	0.256	0.260	0.327	0.338
Average	0.255	0.264	<u>0.306</u>	**0.390**	0.269	0.283	0.348	0.305
	L_r (↓)							
Name	UCE	RECE	ESD	ANT	ScaPre	CP	Shirk.†	**EwS**
Van Gogh	0.200	0.203	0.216	0.304	0.223	0.260	0.113	0.175
Picasso	0.193	0.192	0.215	0.292	0.227	0.235	0.136	0.187
Rembrandt	0.189	0.188	0.232	0.294	0.232	0.281	0.132	0.152
Warhol	0.208	0.209	0.175	0.226	0.192	0.214	0.107	0.160
Caravaggio	0.189	0.193	0.224	0.299	0.218	0.281	0.138	0.182
McKernan	0.144	0.145	0.160	0.248	0.164	0.222	0.090	0.126
Kinkade	0.148	0.148	0.184	0.304	0.187	0.269	0.099	0.140
Edlin	0.169	0.169	0.174	0.290	0.175	0.239	0.093	0.131
Kilian Eng	0.167	0.167	0.159	0.252	0.172	0.222	0.106	0.120
Ajin	0.136	0.138	0.136	0.232	0.154	0.213	0.090	0.140
Average	<u>0.174</u>	0.175	0.187	0.274	0.194	0.244	0.110	**0.151**
	L_d (↑)							
Name	UCE	RECE	ESD	ANT	ScaPre	CP	Shirk.†	**EwS**
Van Gogh	0.070	0.065	0.121	0.102	0.053	−0.031	0.241	0.135
Picasso	−0.010	−0.003	0.027	0.042	0.014	0.039	0.227	0.079
Rembrandt	0.113	0.119	0.134	0.118	0.077	0.036	0.285	0.139
Warhol	0.144	0.143	0.200	0.192	0.155	0.130	0.286	0.215
Caravaggio	0.064	0.062	0.058	0.065	0.019	0.046	0.233	0.162
McKernan	0.048	0.063	0.153	0.146	0.072	0.012	0.296	0.179
Kinkade	0.111	0.127	0.129	0.088	0.107	0.054	0.213	0.153
Edlin	0.051	0.058	0.078	0.102	0.058	0.022	0.175	0.130
Kilian Eng	0.113	0.118	0.116	0.168	0.093	0.040	0.188	0.151
Ajin	0.102	0.131	0.168	0.135	0.101	0.047	0.236	0.198
Average	0.081	0.088	<u>0.118</u>	0.116	0.075	0.040	0.238	**0.154**

D.2.1 CONTENT-COLLAPSE CASES OF SHIRKAVAND ET AL.

We measure content coherence using a CLIP-IQA-style antonym probe (Wang et al., 2023) with CLIP ViT-L/14. For each generated image, the coherence score is the softmax probability assigned to "a clear image with a recognizable subject" against "a blurry smeared abstract texture with no recognizable subject." Table 20 reports the mean and minimum coherence scores over the 20 target prompts for each artist, together with the corresponding content CLIP scores. We use a mean coherence score below 0.85 as an operational indicator of content collapse. Seven of the ten artists satisfy this criterion, with target-prompt generations that frequently exhibit grid-like or smeared textures rather than coherent subjects rendered in a different style (Figure 7). Consequently, the large LPIPS separation in these cases cannot be attributed solely to style removal.

Table 20: Content-collapse analysis of Shirkavand et al. (2025) on target prompts. Coh. reports the mean/minimum over 20 prompts per artist; content CLIP denotes prompt–image similarity for the original/erased models. Mean Coh. (EwS): 0.987.

Artist	L_d	Coh. mean	Coh. min	CLIP (SD)	CLIP (erased)
Van Gogh	0.241	0.803	0.33	25.3	21.0
Picasso	0.227	0.797	0.10	25.9	23.0
Rembrandt	0.285	0.787	0.35	18.1	16.8
Warhol	0.286	0.778	0.10	20.1	19.4
Caravaggio	0.233	0.720	0.12	24.9	20.6
McKernan	0.296	0.898	0.37	24.1	19.8
Kinkade	0.213	0.987	0.90	16.7	15.4
Edlin	0.175	0.849	0.17	17.3	17.1
Kilian Eng	0.188	0.884	0.04	25.1	22.3
Ajin	0.236	0.724	0.04	18.6	17.3
Mean	0.238	0.823	–	–	–

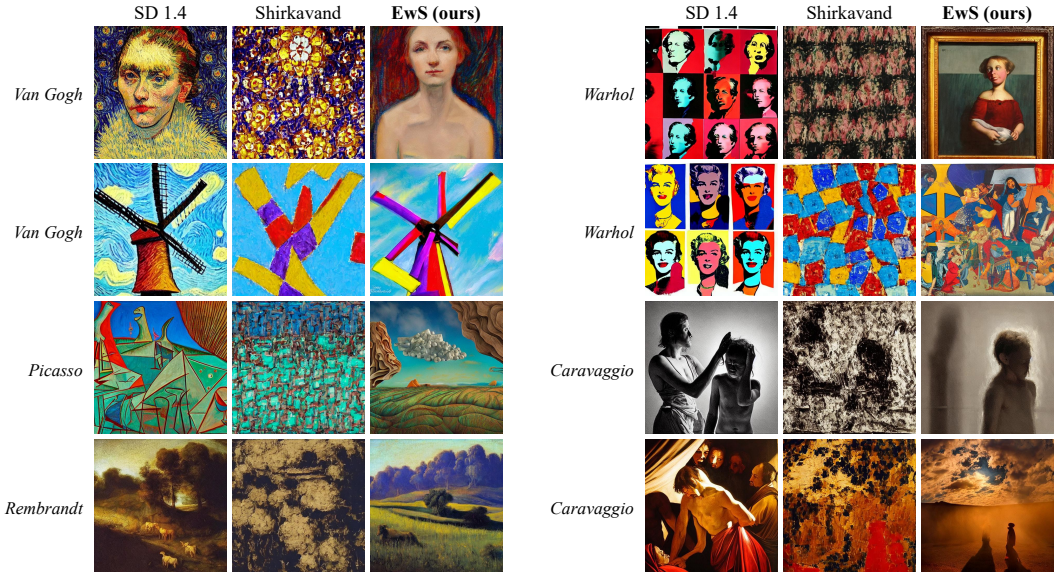


Figure 7: Representative target-prompt generations for five artist-style collapse cases. Within each panel, columns show SD 1.4, Shirkavand et al. (2025), and EwS from left to right. The grid-like or smeared outputs of Shirkavand et al. (2025) indicate that its large LPIPS separation partly reflects content collapse, whereas EwS suppresses the target style while preserving recognizable subjects.

D.3 OBJECT ERASURE

Table 21 reports the class-wise results corresponding to Table 1. Several methods reduce erased-class accuracy to zero or near zero for most classes, making retained-class accuracy an important differentiator. Residual target accuracy remains class-dependent for UCE, ESD, ScaPre, ConceptPrune, and the approach of Shirkavand et al. (2025), with particularly large residuals for several UCE and Shirkavand targets. EWS achieves the highest retained-class accuracy for nine of the ten classes and ties for the highest accuracy on tench.

Table 21: Class-wise object-erasure results under the same sparse budget (overall sparsity 25.21%), corresponding to the averages in Table 1. Bold and underline denote the best and second-best results in the average row. CP: ConceptPrune; Shirk.: Shirkavand et al. (2025).

	Accuracy of erased class (↓)								Accuracy of other classes (↑)							
Name	UCE	RECE	ESD	ANT	ScaPre	CP	Shirk.	**EwS**	UCE	RECE	ESD	ANT	ScaPre	CP	Shirk.	**EwS**
Tench	0.02	0.00	0.29	0.14	0.00	0.00	0.63	0.00	0.54	0.51	0.72	0.70	0.56	0.56	0.77	0.77
English springer	0.05	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.77	0.76	0.66	0.73	0.70	0.56	0.76	0.83
Cassette player	0.01	0.00	0.01	0.00	0.00	0.01	0.03	0.00	0.75	0.76	0.78	0.77	0.73	0.62	0.84	0.92
Chain saw	0.00	0.00	0.00	0.01	0.00	0.00	0.17	0.00	0.79	0.75	0.71	0.73	0.72	0.60	0.80	0.85
Church	0.68	0.00	0.33	0.05	0.37	0.03	0.80	0.03	0.68	0.53	0.73	0.73	0.68	0.51	0.79	0.84
French horn	0.37	0.00	0.08	0.01	0.03	0.03	0.26	0.00	0.69	0.68	0.64	0.57	0.62	0.50	0.74	0.86
Garbage truck	0.48	0.00	0.21	0.01	0.04	0.00	0.13	0.01	0.71	0.74	0.63	0.62	0.71	0.50	0.78	0.83
Gas pump	0.13	0.01	0.06	0.00	0.00	0.02	0.30	0.00	0.70	0.69	0.71	0.61	0.65	0.53	0.77	0.82
Golf ball	0.69	0.00	0.17	0.00	0.20	0.57	0.42	0.00	0.74	0.75	0.67	0.59	0.72	0.54	0.74	0.81
Parachute	0.38	0.02	0.09	0.00	0.05	0.08	0.57	0.02	0.68	0.71	0.68	0.71	0.66	0.47	0.76	0.87
Average	0.28	**0.00**	0.12	0.02	0.07	0.07	0.36	0.01	0.70	0.69	0.69	0.68	0.67	0.54	0.77	**0.84**

D.4 MULTI-CONCEPT OBJECT ERASURE

Setup. We erase five Imagenette classes simultaneously (church, golf ball, gas pump, garbage truck, and parachute) and evaluate preservation on the remaining five classes. UCE and RECE edit all targets jointly, ScaPre uses its native multi-concept mode, ConceptPrune combines the five class-specific masks, and ESD and ANT cycle through the targets for five times their single-concept iterations (5,000 and 1,250 steps, respectively). EWS uses 11,250 steps with each target image paired with its class prompt. To preserve general generation over this longer schedule, we set $\lambda_{\mathrm{dp}} = 0.1$ (0 for single-concept object erasure) and reuse 1,136 COCO val2014 images from the preservation pool used to recover the common sparse model; all are disjoint from the evaluation set.

We average ResNet-50 accuracies and the CLIP probabilities defined in Appendix C.3 over the five erased and retained classes, respectively, and measure generalization by CLIP score on COCO-1k. Because ResNet-50 maps cassette player and chain saw predominantly to neighboring ImageNet labels (tape player and chain), retained accuracy is interpreted relative to the unedited sparse model (0.634).

Table 22: Five-class object erasure under the same sparse budget. Bold and underline denote the best and second-best results, respectively, among the erasure methods. The sparse model without erasure is included as a reference.

	Accuracy		CLIP prob.		COCO-1k
Method	Erased↓	Retained↑	$P_{\mathrm{tgt}}\downarrow$	$P_{\mathrm{ret}}\uparrow$	CLIP↑
Sparse model (no erasure)	0.870	0.634	1.000	1.000	31.63
UCE	0.484	0.564	0.932	**1.000**	**30.99**
RECE	**0.000**	0.582	**0.153**	0.997	30.28
ESD	0.094	0.314	0.616	0.884	29.72
ANT	0.004	0.310	0.231	0.826	28.87
ScaPre	0.026	0.538	0.367	0.989	30.68
ConceptPrune	0.086	0.212	0.654	0.953	29.11
EwS	0.014	**0.664**	0.263	0.996	29.11

Results. Table 22 shows that EwS combines strong target suppression with the highest retained-class accuracy. Among the methods with erased-class accuracy below 0.03 (RECE, ANT, ScaPre, and EwS), EwS obtains a retained-class accuracy of 0.664, compared with 0.310–0.582 for the other methods, while maintaining P_{ret} at 0.996. Its retained accuracy also slightly exceeds that of the unedited sparse model (0.634). RECE and ANT achieve lower P_{tgt} but lower retained-class accuracy, whereas ScaPre has both a higher P_{tgt} and lower retained-class accuracy than EwS. UCE and ScaPre obtain the highest COCO-1k scores but leave more residual target content. The COCO-1k score of EwS (29.11) lies within the range of ESD and ANT (28.87–29.72), whose iteration budgets are likewise scaled fivefold, while EwS preserves substantially higher retained-class accuracy.

D.5 SDXL Experiments

Table 23 evaluates Van Gogh erasure on SDXL-base-1.0 at overall densities of 0.80 and 0.60. Because the parameter distribution of SDXL differs from that of SD 1.4, we use group-wise sparsity budgets. At density 0.80, the realized sparsities are 39.8% for convolutional layers, 27.4% for feed-forward layers, and 10.1% for cross-attention K/V projections; self-attention and cross-attention Q/O projections remain dense. At density 0.60, convolutional sparsity is 36.6%, whereas feed-forward, self-attention, and cross-attention Q/O projections each use 47.0% sparsity, placing the additional reduction primarily in the larger linear groups; K/V projections remain at 10.0% sparsity. For each budget, BASE-XL uses the same layer-wise sparsities as EwS-XL but is constructed independently of the erasure objective.

Table 23: Van Gogh erasure on SDXL-base-1.0 at two overall densities. BASE-XL is obtained by one-shot magnitude pruning under the same layer-wise sparsity budget as the corresponding EwS-XL model, followed by recovery without an erasure objective. Dense rows are unpruned references. Bold and underline denote the best and second-best results among sparse erasure methods within each budget.

Density	Model	$L_f \uparrow$	$L_r \downarrow$	$L_d \uparrow$	$P_{\mathrm{tgt}} \downarrow$	$P_{\mathrm{ret}} \uparrow$
1.0	SDXL (original)	–	–	–	0.907	0.944
	UCE	0.149	0.014	0.134	0.519	0.958
	RECE	0.183	0.024	0.159	0.439	0.959
	ESD-x-strict	0.346	0.187	0.159	0.373	0.777
0.80	BASE-XL (no erasure)	0.162	0.188	−0.026	0.933	0.936
	UCE	0.188	0.192	−0.004	0.612	<u>0.922</u>
	RECE	<u>0.201</u>	0.193	0.008	<u>0.467</u>	<u>0.922</u>
	ESD-x-strict	0.193	**0.145**	<u>0.048</u>	0.602	0.843
	EwS-XL	**0.283**	<u>0.159</u>	**0.124**	**0.221**	**0.946**
0.60	BASE-XL (no erasure)	0.163	0.202	−0.039	0.935	0.876
	UCE	0.189	0.207	−0.018	0.556	0.882
	RECE	<u>0.219</u>	0.206	0.014	<u>0.436</u>	<u>0.887</u>
	ESD-x-strict	0.193	**0.119**	<u>0.074</u>	0.752	0.821
	EwS-XL	**0.289**	<u>0.190</u>	**0.100**	**0.145**	**0.953**

Across both budgets, EwS-XL achieves the highest L_d, the lowest P_{tgt}, and the highest P_{ret} among the sparse erasure methods. At density 0.60, it maintains a positive L_d of 0.100, reduces P_{tgt} to 0.145, and retains a P_{ret} of 0.953. Although ESD-x-strict (Gandikota et al., 2023) attains a lower L_r at this density, it leaves substantially more residual target content ($P_{\mathrm{tgt}} = 0.752$). These results show that EwS-XL remains effective under the smaller overall parameter budget.

D.6 Additional Qualitative Comparisons

Figures 8–10 show additional comparisons under the common sparse budget (d=0.75). Explicit regions are masked with black boxes by the authors.

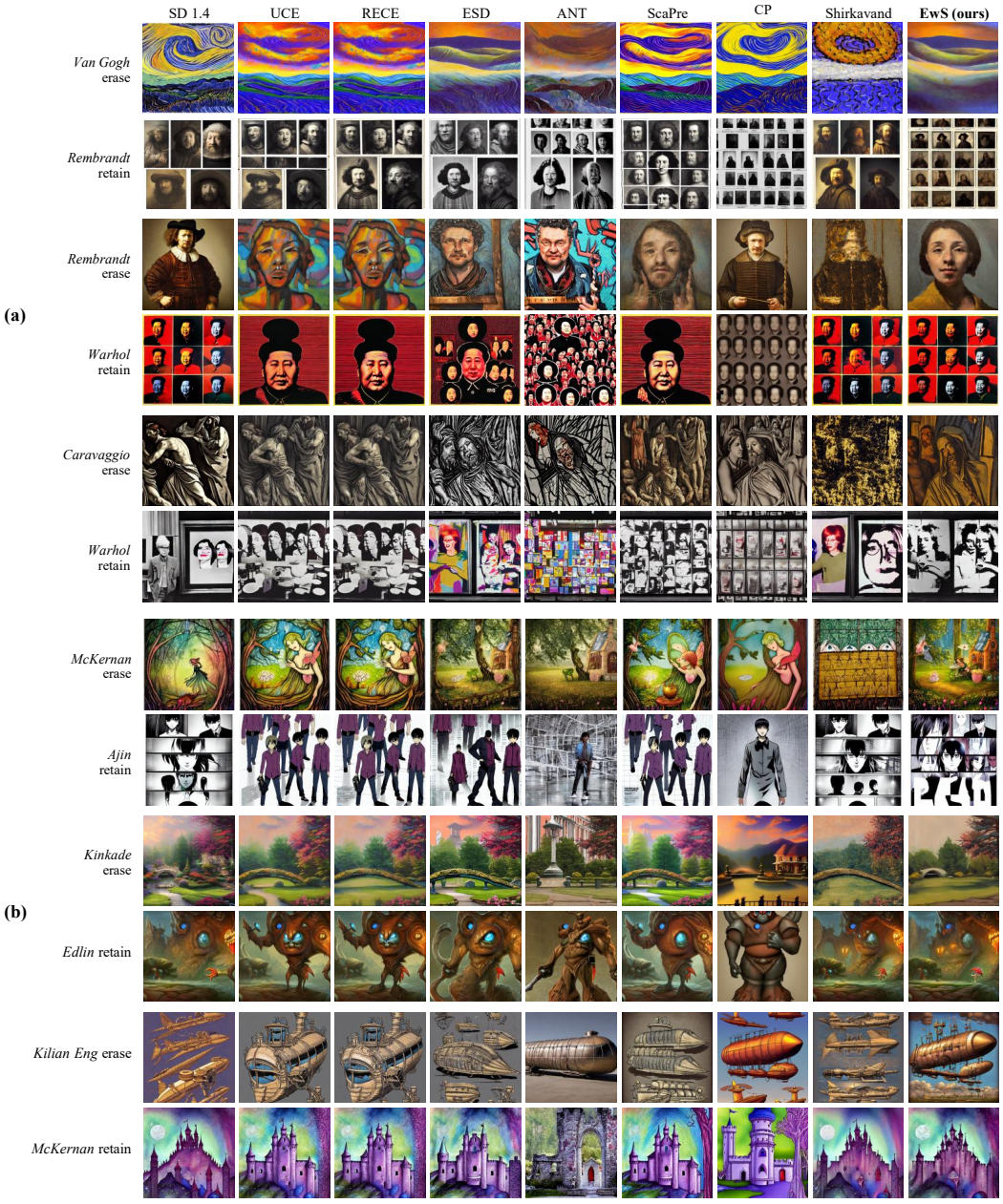


Figure 8: Artistic-style erasure for **(a)** famous and **(b)** modern artists. Each pair of rows shows the erased target style and a retained non-target style.

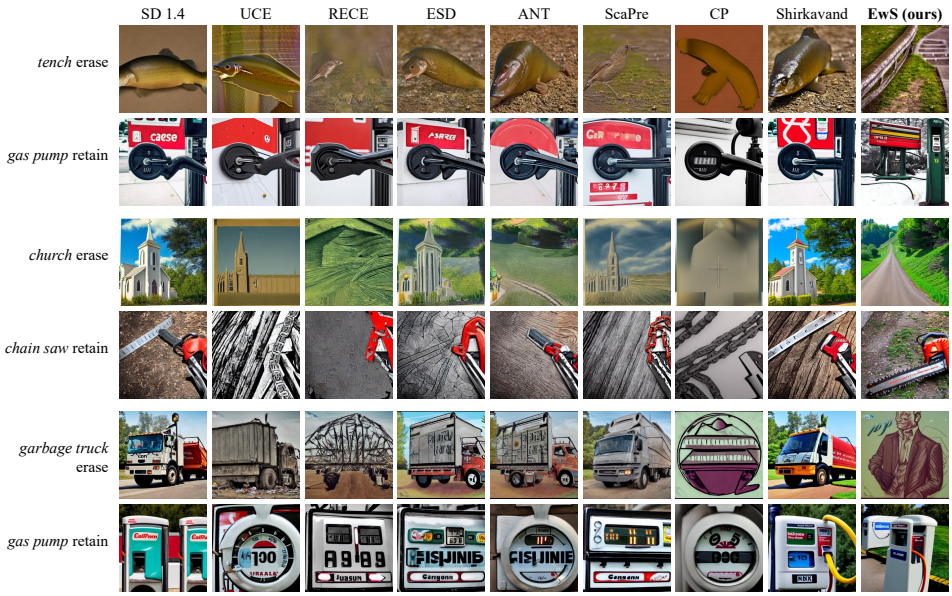


Figure 9: Object erasure on Imagenette targets. Each pair of rows shows the erased target class and a retained class.

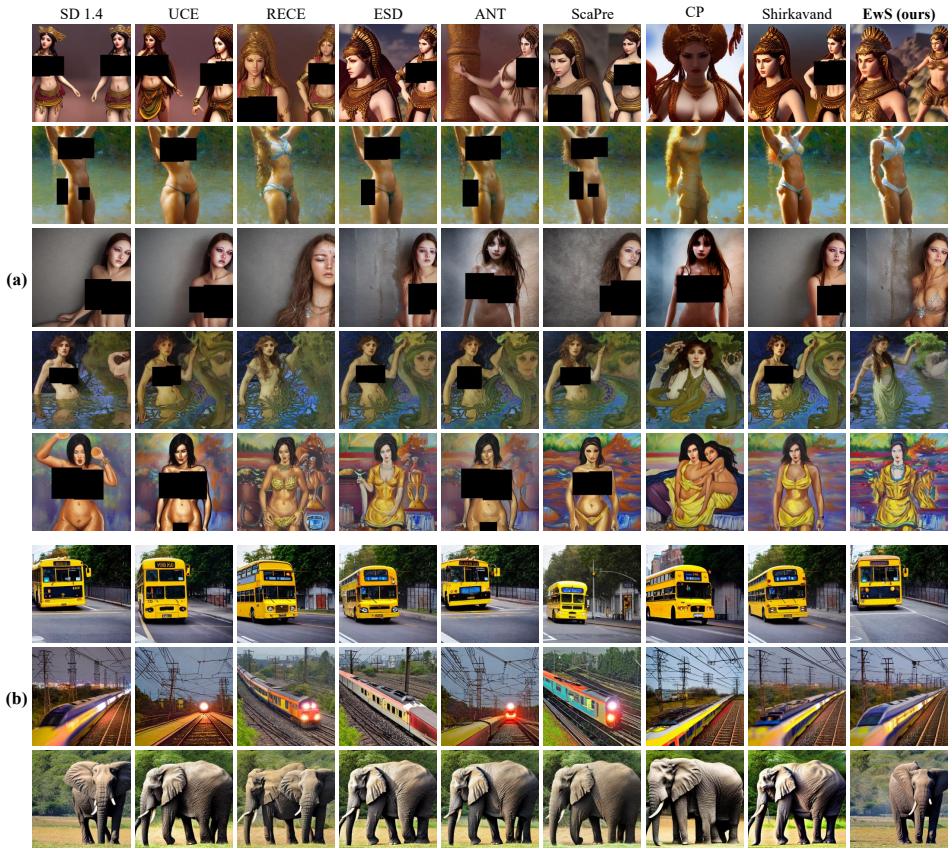


Figure 10: **(a)** Explicit-content erasure on I2P prompts and **(b)** COCO generations from the same models.