SPECTRAL SALIENCY-GUIDED CONCEPT ERASURE VIA ORDERED GRADIENT PROJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept erasure aims to suppress a target concept in a generative model while preserving unrelated behavior. Existing methods commonly restrict either the coordinates or the directions of an update, although these operations need not commute. We propose Spectral Saliency-Guided Erasure (SSGE), which estimates low-rank forget and retain gradient subspaces once at initialization (the retain subspace is a denoising-gradient proxy), derives a continuous coordinate-wise gate from their loadings, and projects the gated forget gradient onto the orthogonal complement of the sampled retain subspace, without restricting updates to a single forget-derived adapter subspace. For an orthonormal retain basis, this ordering makes the resulting forget component orthogonal to the sampled basis; reversing the operator order or subsequently applying adaptive preconditioning need not preserve this property. On Stable Diffusion 1.5, SSGE reduces Van Gogh target CLIP from 0.3160 to 0.2279 across three training seeds, compared with 0.2678 for a protocol-matched deterministic UCE baseline, but produces larger changes on retain prompts. Across five Imagenette objects, suppression transfers to held-out synonyms and compositional prompts, while all adjacent-class prompt sets also lose alignment. SSGE therefore provides a well-defined local gradient constraint and consistent target suppression, but does not guarantee broad functional preservation.

1 INTRODUCTION

Latent diffusion models generate high-quality images from text (Ho et al., 2020; Rombach et al., 2022), but their ability to reproduce artistic styles and object concepts creates a need for targeted model editing. An effective concept-erasure method should suppress the designated behavior without substantially changing nearby or unrelated concepts. This objective is difficult because target and non-target behaviors share parameters and internal representations.

Existing approaches modify the training objective, analytically edit model weights, or restrict the coordinates that may change (Gandikota et al., 2023; 2024; Fan et al., 2024). Coordinate selection answers where an update is applied, whereas subspace projection constrains the directions it may follow. These decisions are not interchangeable: a diagonal coordinate gate generally does not commute with a subspace projector, and applying the gate after projection can reintroduce a protected component.

We address this interaction with Spectral Saliency-Guided Erasure (SSGE). SSGE estimates separate forget and retain subspaces from normalized gradient samples, converts their coordinate loadings into a continuous gate, and then projects the gated forget gradient away from the retain subspace. Unlike gradient-adapter methods that fit a single forget-derived basis on frozen weights, SSGE estimates both bases in the flattened editable-parameter space and gates coordinates continuously. This order makes the constructed forget component lie in the sampled retain null space; we also characterize why reversing the order and using an adaptive optimizer can break that identity.

We evaluate SSGE on Stable Diffusion 1.5 using one artistic style and five object concepts. Three training seeds measure run-to-run variation and component effects, held-out prompts test transfer, and retain-basis controls probe the mechanism. The method consistently suppresses target concepts, but also reduces alignment for every adjacent-class prompt set, distinguishing the local geometric guarantee from behavioral preservation.

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.

A constrained gradient need not determine edited-model behavior: the retain basis is a finite-rank estimate, the optimizer can rotate the applied displacement, and shared representations are reused; we therefore treat orthogonality as local and test preservation directly with unrelated and adjacent prompts.

Our main contributions are:

- an init-time forget–retain leverage-score gate in which U_f supplies coordinate weights and a frozen-at-init rank-16 retain subspace supplies the projection subspace, rather than a paired dual-basis spectral method;
- an operator-ordering remark: gating before projection keeps the constructed forget gradient orthogonal to the sampled basis, while reversed ordering or adaptive preconditioning leaks the protected component back; and
- an evaluation discipline of paired prompt–generation-seed controls, three training seeds as the replication unit, held-out synonym/compositional/adjacent prompt banks, and an explicit refusal to pool images as independent replicates.

2 Related Work

2.1 Inference-Time Intervention

Inference-time interventions control generation without changing diffusion weights: Safe Latent Diffusion steers generation away from unsafe directions (Schramowski et al., 2023), Prompt-to-Prompt controls cross-attention during sampling (Hertz et al., 2023), and SAeUron intervenes on sparse activation features (Cywiński & Deja, 2025). These methods avoid fine-tuning but must remain active for every generation, whereas SSGE constructs a persistent parameter update.

2.2 Tuning-based Erasure

Machine unlearning removes the influence of designated training data without full retraining (Cao & Yang, 2015; Bourtoule et al., 2021); exact and certified formulations provide formal guarantees (Guo et al., 2020; Neel et al., 2021), while selective scrubbing and deep unlearning approximate this objective (Golatkar et al., 2020a;b; Kurmanji et al., 2023). Sparsity, synaptic dampening, remain-geometry, and influence functions further motivate localized parameter updates (Jia et al., 2023; Foster et al., 2024; Huang et al., 2024b; Koh & Liang, 2017). Our setting edits prompt-conditioned behavior and does not claim certified removal.

Diffusion concept erasure includes iterative, analytical, and localized interventions. ESD optimizes an anti-guidance objective (Gandikota et al., 2023), anchor-based ablation remaps unwanted concepts (Kumari et al., 2023), and Forget-Me-Not applies a lightweight attention update (Zhang et al., 2024a). UCE analytically edits cross-attention projections (Gandikota et al., 2024), and MACE extends low-rank editing to multiple concepts (Lu et al., 2024). Other approaches select weights with large forget-gradient magnitudes (Fan et al., 2024), prune concept-specific neurons (Chavhan et al., 2025), or dynamically select parameters (Li et al., 2025).

2.3 Preservation on Unerased Concepts

Preservation has been addressed through continual-learning objectives (Heng & Soh, 2023), auxiliary erasers (Huang et al., 2024a), separable multi-concept updates (Zhao et al., 2024), and value-space interventions (Wang et al., 2025). Related studies consider undesirable influence (Wu et al., 2025a), adversarial recovery (Zhang et al., 2024b), and concept resurgence after subsequent fine-tuning (Suriyakumar et al., 2024). DuMo preserves the original U-Net path and modulates an auxiliary eraser across timesteps and layers (Han et al., 2025), while DoCo removes components that conflict with a concept-preserving gradient (Wu et al., 2025b). These methods address retention through auxiliary or decomposed updates; SSGE instead encodes retention into the forget update itself through the retain-basis projection, so preservation does not require a separate module or objective beyond the distillation term already present.

Recent evaluations broaden preservation beyond generic prompt fidelity by testing model integrity, collateral ripple effects, residual knowledge, and rapid relearning (Schioppa et al., 2024; Amara et al., 2025; Xuan & Li, 2025; Gao et al., 2024), motivating our separate retain, adjacent, synonym, and compositional prompt banks.

The proposed construction is also related to low-dimensional adaptation (Hu et al., 2022; Aghajanyan et al., 2021), continual-learning protection (Kirkpatrick et al., 2017; Saha et al., 2021; Farajtabar et al., 2020), and gradient surgery for task interference (Yu et al., 2020). When task gradients conflict, PCGrad projects a task gradient onto the normal plane of the other gradient; SSGE instead uses a fixed init-time retain subspace and a diagonal gate applied before projection, a composition of two linear operators that does not adapt the protected directions online. Closest in appearance is SEMU (Sendera et al., 2025), which freezes the base weights, projects the accumulated unlearning gradient orthogonally to the current weights, SVDs that projected gradient, and fine-tunes an appended low-rank adapter in the per-layer matrix outer-product space: a single forget-derived basis with no retain basis and no forget–retain tradeoff gate. SSGE instead works across the flattened editable parameters: U_f enters only through its coordinate leverage scores that drive the diagonal gate, and U_r supplies the fixed subspace onto which the gated forget gradient is projected in place, with no adapter. The two methods therefore define different operators: SEMU restricts each update to the range of a forget-derived adapter, whereas SSGE applies a coordinate-wise gate followed by a retain-null-space projection, $g \mapsto P_\rho W g$ with $P_\rho = I - \rho U_r U_r^\top$, whose output generally does not lie in $\mathrm{span}(U_f)$. Null-space calibration also constrains updates through protected directions (Chen et al., 2024). Our matched evaluation includes UCE for style and ESD-X, UCE, and the analytical SPEED editor (Li et al., 2026) for objects. We do not report a validated SEMU baseline; Appendix L documents our unsuccessful SD1.5 adaptation as a diagnostic rather than a comparison. SalUn and MACE remain outside our empirical coverage.

3 METHODOLOGY

SSGE separates coordinate selection from directional protection. Phase 1 estimates compact forget and retain gradient subspaces and converts their coordinate loadings into a fixed continuous gate. Phase 2 gates each forget gradient, projects it away from the captured retain subspace, and adds a complementary retain gradient. Figure 1 summarizes the two phases and the required operator order.

3.1 GRADIENT SUBSPACE CONSTRUCTION

Let $\epsilon_{\theta_0}(x_t, t, c)$ be the frozen model's conditional noise prediction and $\epsilon_\theta(x_t, t, c)$ the editable model. For a forget condition c_f, the canonical anti-guidance target is

$$\epsilon_f^- = \epsilon_{\theta_0}(x_t, t, \varnothing) - \eta\big(\epsilon_{\theta_0}(x_t, t, c_f) - \epsilon_{\theta_0}(x_t, t, \varnothing)\big). \tag{1}$$

The forget loss matches the editable prediction to ϵ_f^-. For a retain condition c_r, the retain loss distills the frozen conditional prediction:

$$L_f = \|\epsilon_\theta(x_t, t, c_f) - \epsilon_f^-\|_2^2, \tag{2}$$

$$L_r = \|\epsilon_\theta(x_t, t, c_r) - \epsilon_{\theta_0}(x_t, t, c_r)\|_2^2. \tag{3}$$

We sample the forget latents before evaluating the loss so that the subspace estimator and the subsequent gradient computation use the same latent state.

At initialization, L_r is nearly zero because the editable and teacher models coincide. We therefore estimate retain geometry in Phase 1 with the standard denoising objective

$$L_{\mathrm{den}} = \|\epsilon_\theta(x_t, t, c_r) - \varepsilon\|_2^2, \tag{4}$$

where x_t is formed using sampled noise ε. Let $g_f^{(i)} = \nabla_\theta L_f$ and $h_r^{(i)} = \nabla_\theta L_{\mathrm{den}}$ denote independent Phase-1 samples. After normalizing each sample to unit ℓ_2 norm, we form

$$\bar{g}_f^{(i)} = \frac{g_f^{(i)}}{\|g_f^{(i)}\|_2}, \qquad G_f = [\bar{g}_f^{(1)}, \ldots, \bar{g}_f^{(N_f)}], \tag{5}$$

and define G_r analogously from the normalized retain gradients. Phase 2 instead uses the unnormalized forget gradient $g_f = \nabla_\theta L_f$ and the teacher-distillation gradient $g_r = \nabla_\theta L_r$ at each optimizer step.

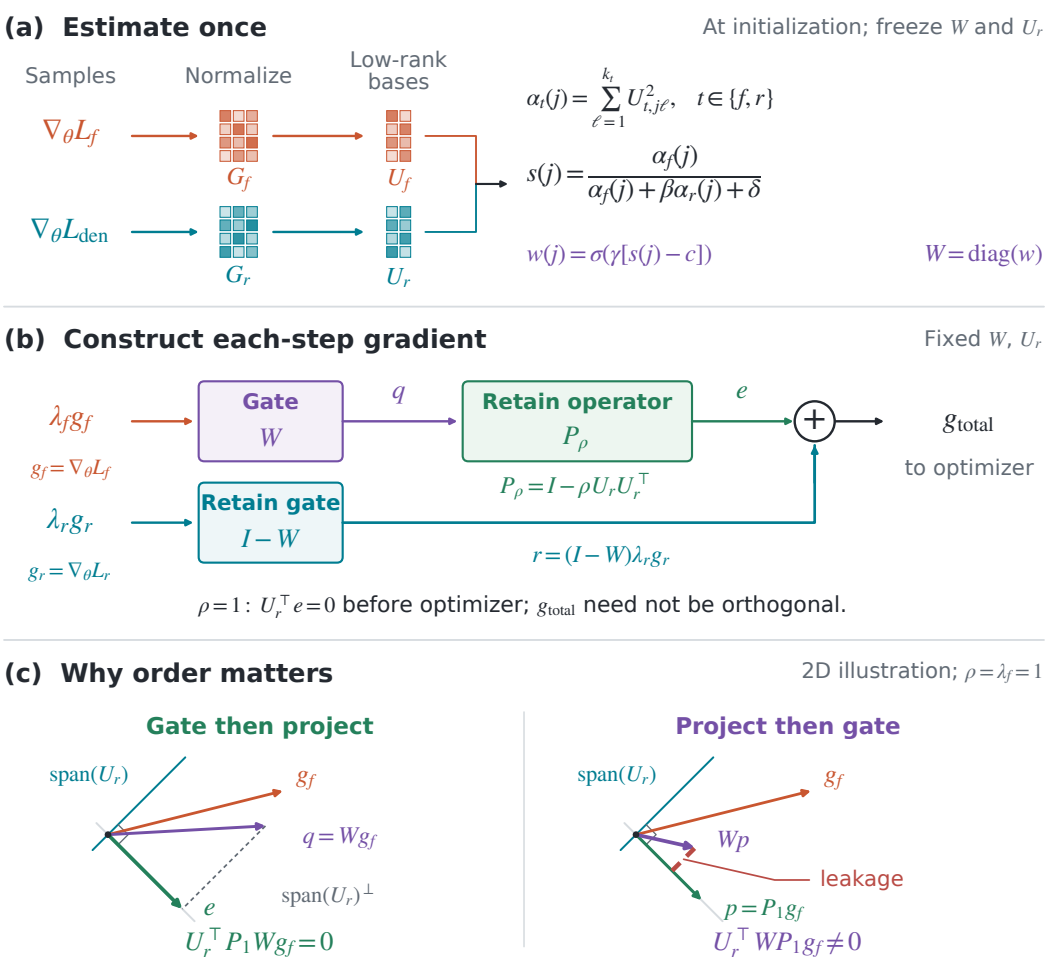


Figure 1: Overview of SSGE. At initialization, normalized forget-loss and retain-denoising gradients yield fixed orthonormal bases U_f, U_r and the coordinate gate W. At each step, the forget branch is gated and then projected by $P_\rho = I - \rho U_r U_r^\top$; the complementary retain branch is added afterward. For $\rho = 1$, $U_r^\top P_1 W \lambda_f g_f = 0$ for the constructed forget component before optimizer transformation. The same two-dimensional example illustrates both operator orders; reversing them can reintroduce a component in $\text{span}(U_r)$.

3.2 Spectral Saliency Gating

For either matrix G, we eigendecompose the smaller Gram matrix $G^\top G$ and recover the leading left singular directions. Eigenvalues below $\max(10^{-12}, r_{\text{tol}}\lambda_{\max})$ are discarded. The resulting forget and retain bases, U_f and U_r, are re-orthogonalized in float64 with a Cholesky correction, and an independent float64 computation verifies $\|U^\top U - I\|_{\max}$ before training proceeds; Appendix B provides the numerical details.

Separate factorizations preserve the roles of the two gradient sets. Combining them before factorization would not indicate whether a coordinate supports forgetting or retention.

For coordinate j, the canonical loading-based leverage scores are

$$\alpha_f(j) = \sum_{\ell=1}^{k_f} U_{f,j\ell}^2, \qquad \alpha_r(j) = \sum_{\ell=1}^{k_r} U_{r,j\ell}^2. \tag{6}$$

They define a continuous gate

$$s(j) = \frac{\alpha_f(j)}{\alpha_f(j) + \beta\alpha_r(j) + \delta}, \qquad w(j) = \sigma(\gamma(s(j) - c)). \tag{7}$$

The leverage score is the diagonal of the corresponding orthogonal projector: for a rank-k basis U,

$$\alpha(j) = e_j^\top U U^\top e_j, \qquad 0 \le \alpha(j) \le 1, \qquad \sum_{j=1}^{D} \alpha(j) = k. \tag{8}$$

It is therefore invariant to rotations of the recovered basis and depends only on the captured subspace. This property is important because individual singular vectors need not be stable when singular values are close, whereas their projector can remain stable.

The relative score $s(j)$ is high when a coordinate is prominent in the forget subspace but weak in the retain subspace; β controls the retain-loading penalty, c the gate center, and γ the transition sharpness. Coordinates with large loadings in both subspaces remain contested rather than forget-specific, and unlike a binary mask the continuous gate preserves graded evidence.

Column normalization makes these scores reflect recurring directions rather than singular-value-weighted gradient energy; otherwise, a few large-norm samples could dominate both the estimated span and its coordinate loadings. The Phase-2 coefficients λ_f and λ_r control update magnitude separately. The bases and gate remain fixed throughout each training run.

For the canonical seed-42 Phase-1 checkpoint, the fixed gate has a pronounced bimodal distribution: 70.7% of editable coordinates have $w > 0.9$, 10.4% have $w < 0.1$, and 35.0% satisfy $|w-1| > 0.05$ (Appendix C). Thus, most coordinates receive near-unit forget-gradient weighting, while a smaller fraction is strongly attenuated. These two extremes together cover 81.1% of editable coordinates; the remaining 18.9% lie in the transition band of the gate (Appendix C).

Given a forget gradient g_f and retain gradient g_r, SSGE constructs

$$\widetilde{g}_f = w \odot \lambda_f g_f, \tag{9}$$

$$g_f^{\text{final}} = (I - \rho U_r U_r^\top)\widetilde{g}_f, \tag{10}$$

$$g_{\text{total}} = g_f^{\text{final}} + (1 - w) \odot \lambda_r g_r. \tag{11}$$

The forget basis enters only through the diagonal leverage score $\alpha_f(j)$ that shapes $w(j)$; the gated forget update is not constrained to lie in $\text{span}(U_f)$. Consequently, coordinates with near-unit weights pass $\lambda_f g_f$ essentially unchanged, while coordinates below 0.1 are excluded from the forget term and instead shape the retain contribution $(1 - w) \odot \lambda_r g_r$ (Equations 9–11).

3.3 Ordered Retain-Subspace Projection

Projecting g_{total} would also remove from the retain-distillation gradient any component lying in $\text{span}(U_r)$, precisely the directions Phase 1 identifies as retain-relevant. SSGE therefore projects only $W\lambda_f g_f$ and adds $(I - W)\lambda_r g_r$ afterward. The construction prevents the forget component from entering the captured span while leaving the retain component available for distillation. Consequently, $U_r^\top g_{\text{total}}$ need not be zero, and the formal claim is component-specific. The retain basis U_r is estimated from denoising gradients, whereas the retained component is the teacher-distillation gradient g_r; we treat the denoising gradient as a natural parameter-space geometry proxy and accept this mismatch as a finite-rank estimate rather than an exact match.

Let $P_\rho = I - \rho U_r U_r^\top$ and $W = \text{diag}(w)$. For an orthonormal captured basis,

$$U_r^\top g_f^{\text{final}} = (1 - \rho) U_r^\top W \lambda_f g_f. \tag{12}$$

Fact 1 (Local orthogonality of the constructed forget component)**.** *For an orthonormal basis U_r of the captured retain subspace and full projection ($\rho = 1$), the constructed forget component $g_f^{\text{final}} = P_1 W \lambda_f g_f$ satisfies $U_r^\top g_f^{\text{final}} = 0$ exactly; this holds before optimizer transformation and does not constrain the added retain component or the resulting parameter displacement. This identity constrains only the sampled basis (rank 16 in the main experiments); it does not bound the retain loss, which depends on the uncaptured portion of the retain gradient field.*

Proof. $U_r^\top P_\rho W \lambda_f g_f = (1-\rho) U_r^\top W \lambda_f g_f$ by Equation (12), which vanishes at $\rho = 1$. □

Setting $W = I$ reduces this to the trivial pure-projection identity $U_r^\top P_1 g_f = 0$; the non-trivial content is the commutator $[W, P]$ below, which governs when a coordinate gate can preserve that identity.

The ordering is essential: $U_r^\top P_1 W g_f = 0$, whereas projecting before gating gives $U_r^\top W P_1 g_f$, which is generally nonzero because W and P_1 need not commute.

The reversed-order leakage is governed exactly by this non-commutativity. With $P = P_1$ and commutator $[W, P] = WP - PW$,

$$U_r^\top W P g_f = U_r^\top [W, P] g_f, \tag{13}$$

$$\|U_r^\top W P g_f\|_2 \le \|[W, P]\|_2 \, \|g_f\|_2. \tag{14}$$

The leakage vanishes when the gate preserves both the retain subspace and its orthogonal complement, but a coordinate-dependent gate has no reason to satisfy this condition. Thus the required ordering follows from the combination of coordinate selectivity and directional protection; it is not merely an implementation convention.

Even the correct order does not make the optimizer displacement orthogonal. Let H denote a diagonal coordinate-wise preconditioner and $P = P_1$. Since $U_r^\top P = 0$,

$$U_r^\top H P W g_f = U_r^\top (HP - PH) W g_f. \tag{15}$$

Thus a nonzero commutator $[H, P]$ reintroduces a captured retain component, and Adam's momentum mixes gradients further. The guarantee applies only to the constructed forget gradient, not the retain component, optimizer displacement, uncaptured directions, or generated behavior.

The two operators impose distinct restrictions: $W g_f$ localizes the edit to high-leverage coordinates but leaves $U_r^\top W g_f$ generally nonzero, whereas $P_1 g_f$ obtains directional orthogonality but cannot separate forget-specific from contested coordinates; SSGE composes them. The projection strength ρ scales the protected component by $1-\rho$ (Equation (12)), giving the exact identity at full projection and a controlled residual at partial.

Computational cost. Storing the Phase-1 gradient columns costs $O(DN)$ CPU memory, forming the two small Gram matrices costs $O(D(N_f^2 + N_r^2))$ time, and recovering rank-k_f, k_r bases costs $O(D(N_f k_f + N_r k_r))$. After Phase 1 the full sample matrices are released; each Phase-2 projection adds $O(Dk_r)$ work and stores only the active basis and coordinate gate on the accelerator.

4 Experiments

4.1 Experimental Setup

Tasks and baselines. All experiments use Stable Diffusion 1.5. The style task erases Van Gogh, and the object task erases garbage truck, golf ball, church, parachute, and cassette player from Imagenette (Howard, 2019). Main SSGE results and the component/order controls use training seeds 42, 123, and 3407. UCE (Gandikota et al., 2024) is evaluated on the same Van Gogh prompts, generation seeds, and metrics. Because UCE is deterministic, it supplies one matched operating point rather than a training-seed distribution. Object baselines use one seed-42 ESD-X training run and one fixed UCE/SPEED edit per concept, evaluated on the full matched prompt streams. The current-vector and basis-refresh variants use only training seed 42; additional gate-allocation and random-basis controls use three seeds. Hyperparameters were selected with a seed-42 pilot and fixed before the reported SSGE runs. The canonical configuration uses gate parameters $\gamma = 10$, $\beta = 0.5$, and $c = 0.5$, projection strength $\rho = 1$, anti-guidance scale $\eta = 3$, loss weights $\lambda_f = 3$ and $\lambda_r = 1$, and learning rate 5×10^{-5}; Appendix C provides the full configuration. Phase-1 concept latents use 50 denoising steps at guidance 3; generation uses DDIM with 50 steps, guidance 7.5, and 512×512 resolution. Training disables TF32, enables deterministic algorithms, and uses a gradient-norm cap of 5.

Table 1: **Van Gogh style erasure.** Van Gogh CLIP results under the matched protocol. Values are positive drops from the original model. SSGE reports mean (sample SD) across three training seeds; UCE is deterministic.

SSGE reports mean (sample SD) across three training seeds; UCE is a fixed edit with no training-seed variance. FID is COCO-10k against the shared original-model FID; no matched UCE FID is available.

Method	Target CLIP ↑	Retain CLIP ↓	Forget LPIPS	Retain LPIPS ↓	COCO FID (10k)
SSGE (3 seeds)	**0.0881** (0.020)	0.0277 (0.004)	0.7220 (0.037)	0.5769 (0.023)	**18.0892** (0.9961)
UCE (fixed)	0.0482	**0.0017**	0.6045	**0.3807**	–
Original SD1.5	–	–	–	–	18.9724

Evaluation protocol. Standard forget and retain prompts come from the construction banks, while synonym, adjacent-class, and compositional prompts are held out; synonyms and compositions measure transfer beyond the construction strings and adjacent prompts a stricter locality test. Original and edited models use identical prompt–generation-seed pairs, and training runs, not images, are the replication unit. The canonical configuration edits all 16 U-Net cross-attention blocks, estimates 16-dimensional bases from 32 forget and 32 retain samples, and runs 200 AdamW steps. The main Van Gogh bank uses 50 concept and 70 retain prompts, each with generation seeds 42, 123, and 3407 (150 and 210 pairs per checkpoint). The component and sensitivity scans use a separate 250-image concept stream and 30 generic retain images. Each object has 20 standard, 12 synonym, 12 compositional, 15 retain, and 8 adjacent prompts, each with three generation seeds. Prompt texts, seeds, and settings are in the appendix and the separately shipped evidence package. Within each protocol, every checkpoint is evaluated on the same complete prompt–generation-seed set, so the reported run-level variation reflects training seeds only.

Metrics and statistics. Target suppression is measured with target-text CLIP and alignment with full-prompt CLIP (OpenAI ViT-B/32 (Radford et al., 2021)); lower target and higher full-prompt CLIP are better. LPIPS with the AlexNet backbone (Zhang et al., 2018) measures paired perceptual change, and Clean-FID (Parmar et al., 2022) is reported on 10,000 COCO train2017 images (Lin et al., 2014) for the original model and three SSGE style checkpoints, with matched captions, generation seeds and real-image reference. This descriptive comparison uses one shared original-model generation stream. Paired measurements are averaged within each run and summarized as mean and sample SD across the three run-level values ($n = 3$ for SSGE); these are descriptive, not confidence intervals, and image pairs are not pooled as independent replicates. Single-edit baselines (UCE, SPEED) and the seed-42 exploratory controls carry no training-seed uncertainty estimate at all. These metrics are behavioral proxies, not evidence of semantic deletion or preservation.

Reproducibility controls. All fifteen object checkpoints are checked for canonical loading, recovered rank, basis orthogonality, and logs; original-image streams are byte-matched before paired metrics. Phase-1 counts, prompt banks, and tolerances are frozen with the protocol (details in the appendix), and seeds are recorded per manifest, preventing sample-level replication or post-hoc selection.

4.2 Quantitative Analysis

Style erasure. SSGE reduces Van Gogh target CLIP from 0.3160 to 0.2279 (0.0881 ± 0.020 across seeds; full-precision values in Table 3) and retain-prompt CLIP by 0.0277 ± 0.004, with forget and retain LPIPS of 0.7220 ± 0.037 and 0.5769 ± 0.023; UCE reduces target CLIP by 0.0482 and retain CLIP by 0.0017 with LPIPS 0.6045 and 0.3807. SSGE suppresses target-text CLIP more strongly at this operating point while UCE preserves retain prompts better; with one deterministic baseline and one concept this does not establish overall superiority, and we have not estimated a Pareto frontier.

Object erasure.

Macro baseline (five objects; Table 8). ESD-X: one seed-42 run; UCE/SPEED: fixed edits.

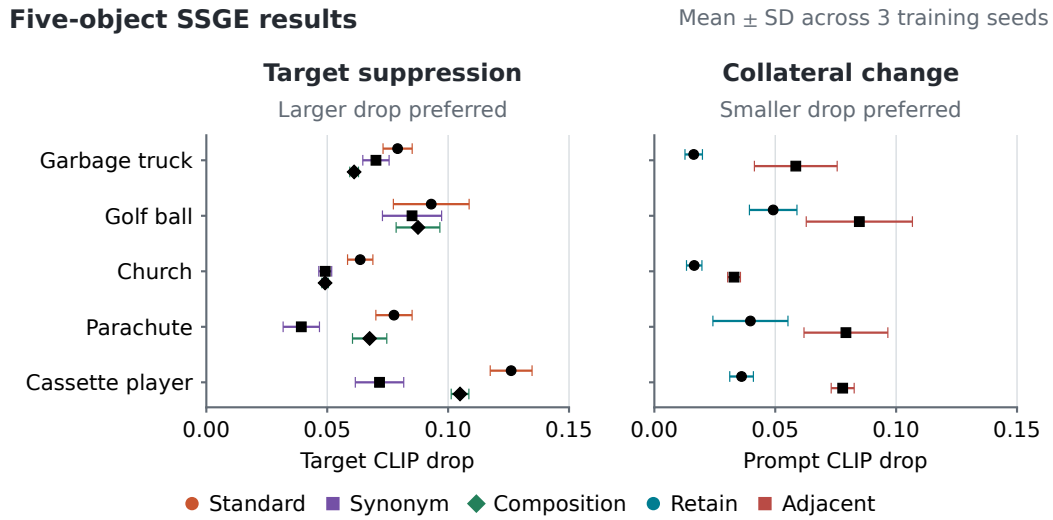


Figure 2: Five-object SSGE results. Points and error bars show the mean and sample SD across three training seeds, not confidence intervals. Target suppression transfers to held-out synonyms and compositions, while every adjacent set loses more prompt alignment than its retain set. Numerical values are in Table 11.

Metric	SSGE	ESD-X	UCE	SPEED
Mean target drop ↑	**0.0879**	0.0813	0.0761	0.0298
Mean retain drop ↓	0.0315	0.0200	0.1110	**0.0025**
Mean adjacent drop ↓	0.0667	0.0477	0.1078	**0.0021**

Across the five objects, target CLIP decreases by 0.0636–0.1261 on standard, 0.0393–0.0851 on held-out synonyms, and 0.0491–0.1049 on held-out compositions, transferring beyond construction strings for every concept. Preservation is weaker: prompt CLIP drops by 0.0163–0.0492 on retain and 0.0330–0.0848 on adjacent prompts, with every adjacent set dropping more than its retain set, and adjacent LPIPS ranges 0.5525–0.7379. Across concepts, cassette player shows the strongest standard-prompt suppression (0.1261 ± 0.0087) with the second-lowest retain drop, while church shows the smallest target drop (0.0636 ± 0.0052) alongside the smallest adjacent drop (0.0330 ± 0.0026). Per concept, ESD-X standard target drops range 0.0203–0.1228, UCE 0.0614–0.0867, and SPEED 0.0122–0.0705; ESD-X and UCE each reach at least one value above 0.08 (0.1228 and 0.0867), while SPEED stays below 0.08 everywhere (Table 8).

Macro-averaged over concepts, the standard target CLIP drop is 0.0879, with synonym/composition transfer at 0.0631/0.0740 (ratios 0.72/0.84, not semantic removal rates). On matched streams SSGE gives the strongest average suppression but more retain change than ESD-X or SPEED, while UCE changes retain prompts more (above table; Table 8). Pointwise, ESD-X leads garbage truck by only 0.0025, below SSGE's own seed SD 0.0060 with ESD-X variance unestimated; ESD-X leads golf ball and UCE leads church nominally, so no method dominates. SSGE mean adjacent drop 0.0667 trails ESD-X 0.0477 and is about 30× SPEED's 0.0021, more than twice its own retain drop. Style COCO Clean-FID is 18.9724 for the original model versus 18.0892 for SSGE (Table 4); this matched comparison is descriptive and does not establish per-image quality or preservation.

A fixed ResNet-50 also shows target recognition falling from 73.7% to 1.7% on standard prompts, 48.3% to 2.8% on synonyms, and 42.8% to 1.1% on compositions after SSGE editing (Table 9). Classifier recognition is an additional proxy, not a semantic-deletion certificate. Among baselines, SPEED leaves the highest residual recognition on standard prompts for four of the five concepts (41.7–75.0%, up to 75% for golf ball), whereas UCE reaches 0% on 14 of the 15 object prompt sets (Table 9).

Table 2: **Three-seed component and order controls.** Means (sample SDs) over training seeds 42, 123, and 3407 on the main style evaluation stream. Projection only removes the gate; gate only removes retain-subspace projection; reversed order gates after projection. Lower retain drop and LPIPS indicate less collateral change.

Variant	Target drop ↑	Retain drop ↓	Retain LPIPS ↓
Full SSGE	<u>0.0881</u> (0.0198)	0.0277 (0.0037)	0.5769 (0.0225)
Projection only	0.0870 (0.0116)	<u>0.0248</u> (0.0015)	**0.5691** (0.0388)
Gate only	**0.0922** (0.0195)	0.0355 (0.0163)	0.5902 (0.0667)
Reversed order	0.0853 (0.0164)	**0.0244** (0.0029)	<u>0.5696</u> (0.0398)

4.3 Qualitative Comparison

Appendix Figure 3 illustrates target suppression and its cost: SSGE changes the target painting toward a photograph, but also changes the non-target Monet style. The matched object row exposes different edited outcomes, including severe degradation for this UCE sample; a single example does not rank methods. Appendix M provides all three SSGE seeds, all five object concepts, and additional matched baseline rows. Aggregate conclusions rest on Table 1 and Figure 2. Each displayed stream uses its first registered prompt at generation seed 42, with no metric or visual-outcome filtering (Appendix M). Among the matched object panels, SPEED shows the least visual change from the original stream, consistent with its residual classifier recognition above (Table 9).

4.4 Ablation Studies

Table 2 tests components and operator order over three training seeds on the main style stream. Removing the gate changes mean target drop by only 0.0011 while improving both retain metrics, and reversing the order weakens suppression by 0.0028 despite losing the geometric identity. The paired full-minus-variant target differences have sample SDs of 0.0093, 0.0068, and 0.0055. These mean differences are small relative to the observed training-run variation and do not establish a consistent advantage for gating or ordering. Variant-level target-drop SDs (0.0116, 0.0195, 0.0164) also fall within the full model's 0.0198, so no variant mean separates from run-to-run variation. Seed-42 basis pilots likewise trade suppression against retention.

The separate three-seed frozen-Phase-1 controls (Table 5) show stronger suppression with learned than constant or permuted gates, mixed retention effects, and no consistent dominance over random retain bases. Exploratory scans (Tables 6 and 7) do not reset the training RNG, confounding parameter effects with training draws; they do not isolate individual hyperparameter effects and are reported descriptively. We therefore report an ordered procedure without claiming consistent superiority of every component. Held-out synonym, compositional, and adjacent evaluations expose the same suppression–retention trade-off; they do not establish immunity to prompt overfitting.

5 Conclusion and Limitations

SSGE combines a loading-based diagonal gate with an ordered retain-subspace projection, $g \mapsto P_\rho W g$, so the constructed forget gradient is orthogonal to a sampled retain basis. Three-seed results show target suppression on Van Gogh (target CLIP drop 0.0881 ± 0.020) and on five object concepts, including held-out synonym and compositional prompts, alongside alignment loss on every adjacent-class set. Our attempted SD1.5 adaptation of SEMU degraded both target and retain images; we report it as an unsuccessful fixed-configuration run, not as evidence that SEMU is inferior to SSGE (Appendix L).

We bound the claim explicitly. The ordered-projection identity constrains only the sampled 16-dimensional subspace and does not bound the retain loss, which depends on the uncaptured gradient field; the retain basis is estimated from denoising gradients while the update acts through a teacher-distillation pathway. Every quality number is a behavioral proxy—CLIP distance, paired LPIPS, FID, and a fixed classifier—and we claim no certified deletion, adversarial robustness, or cross-model generalization. Because adjacent-class drops remain roughly twice the retain drops, deployment should report adjacent and retain together, not retain alone.

Reproducibility Statement

The supplement provides per-image measurements, prompts, seeds, configurations, and scripts to recompute table means and sample SDs; it is an evidence package, not an end-to-end training distribution. The appendix records methods and limitations.

Ethics and Broader Impact

Concept-erasure methods can erase a style or object but can over-erase adjacent concepts, weaken provenance, or be reversed; we claim no certified deletion or safety. Erasing an artist's style bears on attribution, so deployment should add audit records, provenance disclosure, and human review.

AI Use Statement

AI assistants supported coding, debugging, methodology review, auditing, mathematical exposition, and manuscript revision; measurements come from model executions. The authors are responsible for the evidence and final content.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328, 2021.

Ibtihel Amara, Ahmed Imtiaz Humayun, Ivana Kajic, Zarana Parekh, Natalie Harris, Sarah Young, Chirag Nagpal, Najoung Kim, Junfeng He, Cristina Nader Vasconcelos, Deepak Ramachandran, Golnoosh Farnadi, Katherine Heller, Mohammad Havaei, and Negar Rostamzadeh. Erasing more than intended? how concept erasure degrades the generation of non-target concepts. *arXiv preprint arXiv:2501.09833*, abs/2501.09833, 2025.

Lucas Bourtoule, 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*, pp. 141–159, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015.

Ruchika Chavhan, Da Li, and Timothy M Hospedales. ConceptPrune: Concept editing in diffusion models via skilled neuron pruning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kSdWcw5mkp.

Huiqiang Chen, Tianqing Zhu, Xin Yu, and Wanlei Zhou. Machine unlearning via null space calibration. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 358–366, 2024.

Bartosz Cywiński and Kamil Deja. SAeUron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11738–11775, 2025.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gn0mIhQGNM.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 3762–3773, 2020.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12043–12051, 2024.

Rohit Gandikota, Joanna Materzyńska, 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.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*, abs/2410.12777, 2024.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020a.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *Computer Vision – ECCV 2020*, pp. 383–398. Springer, 2020b.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3832–3842, 2020.

Feng Han, Kai Chen, Chao Gong, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. DuMo: Dual encoder modulation network for precise concept erasure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3320–3328, 2025. doi: 10.1609/aaai.v39i3.32343.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17170–17194. Curran Associates, Inc., 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

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

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. OpenReview.net, 2022.

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

Zhehao Huang, Xinwen Cheng, Jinghao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. Unified gradient-based machine unlearning with remain geometry enhancement. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26377–26414. Curran Associates, Inc., 2024b.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51584–51605. Curran Associates, Inc., 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1885–1894, 2017.

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.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1957–1987. Curran Associates, Inc., 2023.

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 ICCV*, pp. 19659–19668, 2025.

Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. SPEED: Scalable, precise, and efficient concept erasure for diffusion models. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=aoEtzdRkGh`.

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*, 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.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

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 *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

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.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*. OpenReview.net, 2021.

Andrea Schioppa, Emiel Hoogeboom, and Jonathan Heek. Model integrity when unlearning with T2I diffusion models. *arXiv preprint arXiv:2411.02068*, abs/2411.02068, 2024.

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

Marcin Sendera, Łukasz Struski, Kamil Książek, Kryspin Musiol, Jacek Tabor, and Dawid Damian Rymarczyk. SEMU: Singular value decomposition for efficient machine unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53843–53866, 2025.

Vinith M. Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C. Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. *arXiv preprint arXiv:2410.08074*, 2024. URL https://arxiv.org/abs/2410.08074.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28759–28768, 2025.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28263–28273, 2025a.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8496–8504. Association for the Advancement of Artificial Intelligence, 2025b.

Hao Xuan and Xingyu Li. RUB: Evaluating residual knowledge in unlearned models. *arXiv preprint arXiv:2504.14798*, abs/2504.14798, 2025.

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, pp. 5824–5836. Curran Associates, Inc., 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 Workshops*, pp. 1755–1764, 2024a. doi: 10.1109/CVPRW63382.2024.00182.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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*, volume 37, pp. 36748–36776, 2024b.

Mengnan Zhao, Lihe Zhang, Tianhang Zheng, Yuqiu Kong, and Baocai Yin. Separable multi-concept erasure from diffusion models. *arXiv preprint arXiv:2402.05947*, abs/2402.05947, 2024.

A Scope of This Appendix

This appendix provides additional method details, the complete implementation configuration, full-precision results, retain-basis analyses, and evaluation integrity checks. Three-seed main and frozen-Phase-1 control aggregates are reported separately from the single-seed mechanism pilots.

B Additional Method Details

The main paper defines the forget loss L_f and the teacher-distillation retain loss L_r. At initialization, L_r is nearly zero because the editable model coincides with the frozen teacher. We therefore use the standard denoising objective in Phase 1:

$$L_{\text{den}} = \|\epsilon_\theta(x_t, t, c_r) - \varepsilon\|_2^2. \tag{16}$$

The retain columns are $h_r^{(i)} = \nabla_\theta L_{\text{den}}$, sampled independently from the forget columns $g_f^{(i)} = \nabla_\theta L_f$. In Phase 2, we use $g_r = \nabla_\theta L_r$, so the retain update distills toward the frozen model rather than fitting sampled noise.

For either normalized gradient matrix G, we symmetrize the small Gram matrix $G^\top G$, clamp negative eigenvalues caused by numerical roundoff to zero, and retain eigenvalues above $\max(10^{-12}, r_{\text{tol}}\lambda_{\max})$. We recover each left singular direction as $Gv_i/\sqrt{\lambda_i}$ and re-orthogonalize the basis with a float64 Cholesky correction. An independent float64 computation verifies $\|U^\top U - I\|_{\max}$ before training proceeds.

For recovered forget and retain bases U_f and U_r, define coordinate loadings

$$\alpha_f(j) = \|U_f[j,:]\|_2^2, \qquad \alpha_r(j) = \|U_r[j,:]\|_2^2. \tag{17}$$

The saliency and continuous gate are

$$s(j) = \frac{\alpha_f(j)}{\alpha_f(j) + \beta\alpha_r(j) + \delta}, \qquad w(j) = \sigma(\gamma(s(j) - c)). \tag{18}$$

The gate is applied before projection:

$$g_f^{\text{final}} = (I - \rho U_r U_r^\top)(w \odot \lambda_f g_f). \tag{19}$$

The optimizer receives

$$g_{\text{total}} = g_f^{\text{final}} + (1 - w) \odot \lambda_r g_r. \tag{20}$$

When $\rho = 1$, orthogonality holds for the constructed forget gradient relative to the captured basis. It does not extend to the AdamW parameter displacement, global functional invariance, or certified deletion.

Algorithm 1 (gate-then-project, Phase 2 per optimizer step). *Input:* forget gradient g_f, retain gradient g_r, frozen bases U_f, U_r, gate w, scalars $\lambda_f, \lambda_r, \rho$. (1) $\tilde{g}_f \leftarrow w \odot \lambda_f g_f$; (2) $g_f^{\text{final}} \leftarrow (I - \rho U_r U_r^\top)\tilde{g}_f$; (3) $g_{\text{total}} \leftarrow g_f^{\text{final}} + (1 - w) \odot \lambda_r g_r$; (4) update θ with g_{total} via AdamW. Steps (1)–(2) apply the gate before projection; reversing them reintroduces a sampled retain component.

C Implementation Details

The main experiments use one NVIDIA GeForce RTX 4090 with 24 GB memory on 64-bit Windows 11 Pro; additional sensitivity and allocation-control runs also use a second RTX 4090 host running Windows 10. Both runtimes used Python 3.10.15, PyTorch 2.7.1+cu118, Diffusers 0.37.0, Transformers 4.52.4, and CUDA 11.8. Training disabled TF32, enabled deterministic algorithms, and used the cuBLAS workspace setting `:4096:8` for the canonical runs. Training and generation seeds, model identities, and scoring resources are recorded in the manifests. The exploratory scans have the RNG limitation stated in their captions.

The configuration is as follows: all 16 UNet `attn2` blocks (`to_q`, `to_k`, `to_v`, and `to_out.0`; 44.0M parameters), with no block ranking and all other model components frozen; 32 independently ℓ_2-normalized forget columns and 32 retain columns; $k_f = k_r = 16$; $r_{\text{tol}} = 10^{-6}$; $\gamma = 10$; $\beta = 0.5$; gate center $c = 0.5$; denominator stabilizer $\delta = 10^{-12}$; projection $\rho = 1$; anti-guidance $\eta = 3$; $\lambda_f = 3$; $\lambda_r = 1$; 200 AdamW steps at learning rate 5×10^{-5}, batch size one, FP16, zero weight decay, and gradient-norm cap 5. Phase-1 concept latents use 50 denoising steps and guidance 3. Image generation uses DDIM, 50 steps, guidance 7.5, 512×512 resolution, and seeds 42, 123, and 3407.

These settings were selected after a seed-42 pilot and fixed before the three reported training runs. Their outcomes were not used to retune the method.

From the frozen Phase-1 checkpoint of the canonical run (seed 42, $\gamma = 10$), the coordinate gate w over 43,962,560 editable parameters has mean 0.802, std 0.331, min 0.0067, max 0.9933, and median 0.984 (5th/95th percentiles 0.019/0.993); 70.7% of coordinates exceed 0.9 and 10.4% fall below 0.1, so the gate is bimodal, and 35.0% of coordinates deviate from 1 by more than 0.05 (its effective support). The forget leverage scores sum to $k_f = 16$ as required.

D STATISTICAL AGGREGATION

For each training run, we first compute one metric mean. If m_t is the within-run mean for training seed t, the reported point estimate and sample standard deviation are

$$\bar{m} = \frac{1}{n}\sum_{t=1}^{n} m_t, \qquad s_m = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(m_t - \bar{m})^2}. \tag{21}$$

For SSGE training-seed aggregates, $n = 3$; single-edit baselines and exploratory controls have no training-seed uncertainty estimate. Image pairs are not pooled as independent replicates for training-seed uncertainty. We compute directional CLIP changes within each original–edited pair, average them within a training run, and then aggregate across training seeds.

Table 3: **Full-precision style results.** Full-precision Van Gogh results under the matched protocol. SSGE values are mean (sample SD) across three training seeds; UCE is deterministic.

Metric	SSGE (3 seeds)	UCE (det.)
Target CLIP (edited)	**0.227908** (0.019771)	0.267807
Target CLIP change	**-0.088104** (0.019771)	−0.048205
Forget LPIPS	0.722041 (0.037122)	**0.604536**
Retain CLIP (edited)	0.296667 (0.003663)	**0.322615**
Retain CLIP change	−0.027652 (0.003663)	**-0.001704**
Retain LPIPS	0.576910 (0.022469)	**0.380660**

Table 4: **Matched COCO-10k Clean-FID.** Original and edited models use the same 10,000 captions, per-image generation seeds and real-image reference. Differences subtract the single shared original-model FID. SD is across the three edited training checkpoints, not across independent COCO samples.

Model / training seed	FID	Difference
Original SD1.5	18.972405	–
SSGE / 42	17.134462	-1.837943
SSGE / 123	19.122069	+0.149664
SSGE / 3407	18.011067	-0.961338
SSGE mean (SD)	18.089199 (0.996104)	-0.883205

Original-model COCO Clean-FID is 18.9724; the three Van Gogh SSGE checkpoints yield 18.0892 ± 0.9961, a mean edited-minus-original difference of -0.8832 on the shared reference and generation inputs.

The matched COCO comparison uses DDIM-50, CFG 7.5, FP16 and batch one at 512×512, without a negative prompt or safety checker. Two Windows RTX 4090 runtimes generate disjoint sets of 5,000 original-model images; each first reproduces two archived edited images exactly. All real, original and three edited streams use explicit, unique 10,000-image manifests. Clean-FID 0.1.35, clean mode, InceptionV3W, batch 32 and NumPy covariance are fixed across streams; the three re-extracted edited scores must match the archived scores within 10^{-4}. FID measures distributional discrepancy, not target removal, per-image quality or certified preservation. The shared original stream and three training checkpoints do not provide independent COCO-sample uncertainty estimates.

The SSGE full-prompt CLIP score on forget images, distinct from target-text CLIP, is 0.285755 (0.015099 across training seeds). The original image streams are matched across methods; UCE remains a point estimate rather than a replicated uncertainty estimate.

E FROZEN-PHASE-1 ALLOCATION CONTROLS

This separately identified Van Gogh mechanism suite loads each training seed's canonical Phase-1 artifact and uses the main ordinary-AdamW hyperparameters. We compare the learned gate and

retain basis with a constant gate equal to the learned gate mean, a permutation preserving the gate distribution (control seed 1729), and a rank-matched random orthonormal retain basis (control seed 2718). The first two controls replace the gate in both the forget and complementary retain channels; the random-basis arm leaves the learned gate fixed. All three controls use the original gate-before-projection order. The same private control seeds are used across training seeds.

All four arms are trained for seeds 42, 123, and 3407 with identical Phase-2 RNG resets within each seed. Frozen loading does not replay the historical Phase-1 random draws, so the full arm is retrained and this suite is not pooled with canonical main checkpoints. Training uses 200 AdamW steps; neither realized displacement nor target suppression strength is matched. Two Windows RTX 4090 runtimes use matching package versions, with training and generation hosts held fixed within each seed's four-arm block. Seed 42 is trained and generated on runtime A; seed 123 on B; and seed 3407 is trained on A and generated on B. Thus the three-seed variation also includes runtime variation between blocks.

Evaluation uses the complete canonical 50-prompt held-out forget stream and 70-prompt retain stream, with three generation seeds per prompt. The twelve models yield 4,320 paired comparisons. All arms share the same original images and DDIM/50-step/guidance-7.5 configuration. Two original-model replay probes per cell must have mean absolute pixel error at most 1 on the $[0, 255]$ scale; the supplement reports the observed non-bitwise differences. Table 5 reports means and sample SDs over three training seeds and paired control-minus-full differences. These summaries treat the training seed, rather than an image pair, as the replication unit and do not imply significance. The random-basis and permutation controls each use one fixed control seed; these experiments do not estimate variation over random control constructions.

Table 5: **Frozen-Phase-1 allocation controls.** New ordinary-AdamW runs with the main hyperparameters and each seed's canonical Phase-1 artifact. Values are mean (sample SD) across three training seeds. Larger target drop means stronger suppression; smaller retain drop and LPIPS mean less collateral change. The Phase-2 RNG is reset for all four arms, so these checkpoints are separate from the canonical main results. Training budget, rather than realized displacement, is matched.

Variant	Target drop ↑	Retain drop ↓	Retain LPIPS ↓
Learned gate + learned basis	0.0839 (0.0161)	0.0294 (0.0081)	0.6011 (0.0831)
Constant-mean gate	0.0781 (0.0199)	0.0254 (0.0059)	0.5581 (0.0496)
Permuted gate	0.0799 (0.0167)	0.0307 (0.0078)	0.5818 (0.0481)
Random retain basis	0.0808 (0.0191)	0.0250 (0.0062)	0.5524 (0.0319)
Paired control minus full: mean (sample SD) over training seeds			
Constant-mean gate	-0.0057 (0.0038)	-0.0039 (0.0023)	-0.0431 (0.0337)
Permuted gate	-0.0040 (0.0020)	+0.0013 (0.0064)	-0.0194 (0.0479)
Random retain basis	-0.0030 (0.0048)	-0.0043 (0.0025)	-0.0488 (0.0514)

The full arm has a target CLIP drop of 0.0839 ± 0.0161, compared with 0.0781 ± 0.0199 for the constant gate, 0.0799 ± 0.0167 for the permuted gate, and 0.0808 ± 0.0191 for the random retain basis. Both gate controls weaken target suppression in each of the three training-seed blocks. The constant gate also reduces retain CLIP drop and LPIPS in every block; its paired changes average -0.0057 target drop, -0.0039 retain drop, and -0.0431 retain LPIPS. Thus the learned gate buys additional suppression at the expense of additional collateral change relative to uniform allocation. Permutation has mixed retention effects: its mean retain CLIP drop is 0.0013 larger, but its mean retain LPIPS is 0.0194 smaller than the full arm.

The random-basis control lowers retain CLIP drop in all three blocks and retain LPIPS in two. Its target suppression is weaker in two blocks and slightly stronger in the third, with a mean target-drop difference of -0.0030 ± 0.0048. These results do not establish that the learned retain geometry improves the suppression–preservation tradeoff. They show that gate allocation affects behavior under this budget, while leaving its advantage at matched suppression or optimizer displacement unresolved.

F RETAIN-BASIS ANALYSIS

These controls use the archived seed-42 fixed-basis pilot as reference, distinct from the canonical main-table checkpoint. Forget deltas average 150 aligned prompt–generation-seed pairs; retain deltas average 210 aligned pairs. They are control minus fixed subspace, so a negative target-CLIP delta indicates stronger erasure, a positive retain prompt-CLIP delta indicates better alignment, and a negative retain LPIPS delta indicates less perceptual displacement. Capture averages all 200 Phase-2 optimizer steps. These paired results characterize the effect of the basis choice but provide no training-seed uncertainty.

Table 6: **Component ablation.** Exploratory seed-42 Van Gogh protocol, separate from Table 1: 50 concept prompts with five generations each, 30 generic and 30 matched retain prompts, and 20 adjacent prompts. Target CLIP is scored against "a painting in the style of Van Gogh"; retain and adjacent rows use prompt-matched text. Rows remove the retain-subspace projection, the continuous modulation, or both from the full SSGE update. (Original 0.3112 uses five sequential draws per concept prompt; main-table original 0.3160 uses three explicitly paired generation seeds.) Phase 1 uses 96 gradient samples rather than the main experiment's 32. Seed 42 is recorded, but the launcher does not reset the training RNG between variants; these runs do not isolate component effects and are not pooled with the main experiment.

Variant	Target CLIP ↓	Retain CLIP ↑	Retain matched ↑	Adjacent CLIP ↑	Retain LPIPS ↓
Original	0.3112	0.3083	0.3286	0.3316	–
SSGE (full)	0.2192	0.3014	0.2946	0.3045	0.3720
w/o projection	0.2359	0.3052	0.2906	0.3068	0.3659
w/o modulation	0.2248	0.3020	0.2925	0.3076	0.3580
Binary mask	0.2400	0.3020	0.2875	0.2996	0.3700
Plain joint	0.2365	0.2980	0.3009	0.3107	0.3655

G FIXED OBJECT-RECOGNITION EVALUATION

To complement text–image similarity, we score all saved standard, synonym, and compositional object images using a fixed ResNet-50 (He et al., 2016), with torchvision's ImageNet-1K V2 weights and prescribed resize-to-232, center-crop-to-224, and ImageNet normalization. Inference uses FP32 evaluation mode. Target class IDs are 569 (garbage truck), 574 (golf ball), 497 (church), 701 (parachute), and 482 (cassette player). The classifier, preprocessing, concepts and complete evaluation streams are fixed before this additional scoring operation; no outcome-dependent crop, threshold or image selection is used.

The 90 paired streams contain 3,960 comparisons and reuse 660 original images. Table 9 reports target top-1 recognition rates; the evidence package also provides recognition loss conditioned on an original positive, including the denominator. SSGE uncertainty is computed across three training-run means. ESD-X, UCE and SPEED remain single-edit point estimates. UCE has the lowest recognition rates here, while the CLIP/LPIPS results show substantial collateral changes at its reported configuration. Original recognition is only 8.3–15.0% for cassette player, so its near-zero edited recognition offers limited evidence on its own. A generated image may fool a classifier or preserve recognizable target features despite a wrong label, so classification failure cannot establish semantic deletion or retention quality.

H FULL OBJECT RESULTS

Table 10 reports the paired CLIP and LPIPS results for every object concept and prompt set.

I EVALUATION INTEGRITY

We validate checkpoint loading, recovered rank, basis orthogonality, and training logs for all fifteen object models. The five prompt sets for each model produce 3,015 paired comparisons and 6,030

Table 7: **Exploratory sensitivity scans.** All completed Van Gogh runs, separated by host. The configurations record seed 42, but the launchers do not reset the training RNG for each row; only evaluation explicitly reseeds. Unless varied, $k = 16$, $\gamma = 10$, and $N = 96$; recovered rank is at most $\min(k, N)$. Each row uses 50 concept prompts with five draws and 30 generic retain images. Differences confound hyperparameters with training draws and are not pooled across hosts.

Setting	Target CLIP ↓	Retain CLIP ↑	Retain LPIPS ↓
Host A: 14 runs			
$\gamma = 1$	0.2311	0.2994	0.4341
$\gamma = 2$	**0.2072**	0.2998	0.3617
$\gamma = 5$	0.2585	<u>0.3033</u>	**0.3279**
$\gamma = 10$	0.2452	**0.3064**	0.3723
$\gamma = 20$	0.2238	0.3006	0.3465
$\gamma = 50$	0.2328	0.3015	<u>0.3450</u>
$k = 4$	<u>0.2131</u>	0.3006	0.3914
$k = 8$	0.2320	0.2956	0.3904
$k = 16$	0.2323	0.3008	0.3774
$k = 32$	0.2259	0.2947	0.4206
$N = 8$	0.2160	0.3018	0.3566
$N = 16$	0.2215	0.3013	0.4158
$N = 32$	0.2315	0.2954	0.4455
$N = 64$	0.2425	0.2975	0.3863
Host B: 9 runs			
$\gamma = 10$	0.2408	0.3008	0.3635
$\gamma = 20$	0.2186	**0.3072**	<u>0.3334</u>
$\gamma = 50$	0.2208	<u>0.3044</u>	**0.3055**
$k = 4$	0.2506	0.3030	0.3633
$k = 8$	**0.2123**	<u>0.3044</u>	0.3534
$k = 16$	0.2354	0.3020	0.3433
$k = 32$	0.2311	0.3023	0.3411
$N = 32$	<u>0.2173</u>	0.2991	0.3801
$N = 64$	0.2315	0.3009	0.3533

Table 8: **Matched-stream object baseline comparison.** Standard target CLIP drops use the same original images, prompts, and three generation seeds for all methods. SSGE reports mean (sample SD) over three training runs; ESD-X uses one seed-42 training run and UCE/SPEED one fixed edit per concept, without training-seed uncertainty. The final three rows are macro-averages across concepts. Evaluation matching does not equate training objectives or budgets.

Concept	SSGE	ESD-X	UCE	SPEED
Garbage truck	<u>0.0791</u> (0.0060)	**0.0816**	0.0735	0.0122
Golf ball	<u>0.0930</u> (0.0157)	**0.1228**	0.0777	0.0223
Church	<u>0.0636</u> (0.0052)	0.0203	**0.0813**	0.0205
Parachute	**0.0776** (0.0075)	<u>0.0674</u>	0.0614	0.0237
Cassette player	**0.1261** (0.0087)	<u>0.1141</u>	0.0867	0.0705
Mean target drop ↑	**0.0879**	<u>0.0813</u>	0.0761	0.0298
Mean retain drop ↓	0.0315	<u>0.0200</u>	0.1110	**0.0025**
Mean adjacent drop ↓	0.0667	<u>0.0477</u>	0.1078	**0.0021**

Table 9: **Independent object-recognition proxy.** Percent of generated images assigned the target ImageNet class by a fixed ResNet-50 (ImageNet-1K V2). Lower edited-model rates indicate less classifier recognition, not certified deletion. SSGE reports mean (sample SD) across three training runs; baselines are single-edit point estimates. All saved images are scored with the prescribed preprocessing. Original recognition rates expose failures already present before editing.

Concept	Prompt set	Original	SSGE	ESD-X	UCE	SPEED
Garbage truck	Standard	91.7	0.6 (1.0)	1.7	0.0	50.0
	Synonym	77.8	0.0 (0.0)	5.6	0.0	47.2
	Composition	30.6	0.0 (0.0)	5.6	0.0	11.1
Golf ball	Standard	95.0	1.7 (1.7)	0.0	1.7	75.0
	Synonym	66.7	2.8 (2.8)	0.0	0.0	58.3
	Composition	86.1	3.7 (4.2)	2.8	0.0	61.1
Church	Standard	88.3	6.1 (1.9)	43.3	0.0	56.7
	Synonym	72.2	11.1 (8.3)	27.8	0.0	44.4
	Composition	38.9	0.0 (0.0)	11.1	0.0	11.1
Parachute	Standard	78.3	0.0 (0.0)	1.7	0.0	41.7
	Synonym	16.7	0.0 (0.0)	5.6	0.0	16.7
	Composition	50.0	1.9 (3.2)	2.8	0.0	22.2
Cassette player	Standard	15.0	0.0 (0.0)	0.0	0.0	1.7
	Synonym	8.3	0.0 (0.0)	0.0	0.0	5.6
	Composition	8.3	0.0 (0.0)	0.0	0.0	0.0
Macro-average	Standard	73.7	1.7 (0.0)	9.3	0.3	45.0
Macro-average	Synonym	48.3	2.8 (1.1)	7.8	0.0	34.4
Macro-average	Composition	42.8	1.1 (1.0)	4.4	0.0	21.1

images in total. Metric summaries are recomputed from per-pair records, and the original images are verified to match across edited models for every shared prompt and generation seed.

The Van Gogh and garbage-truck examples preserve the historical prompt/seed selection, which used original-model prompt alignment and a symmetric corruption screen on the then-displayed models. Added baseline columns use those same prompts and seeds without a new screen. For golf ball, church, parachute, and cassette player, each displayed stream uses its first registered prompt at generation seed 42, with no metric or visual outcome filtering. Neither rule selects on target reduction, retain change, adjacent-class change, or LPIPS. The new figure provenance records prompt indices, generation and training seeds, source paths, and protocol/checkpoint identities. All source frames are displayed without cropping or retouching.

Post-generation checks preserve the original protocol and checkpoint identities. The evidence package includes per-image measurements and a table-recomputation script; independent image re-scoring is recorded separately from training.

These checks ensure internal consistency but do not expand the method's scope. A rank-16 basis cannot capture all gradient variation, and the fixed retain basis captures only a fraction of later retain gradients. The adjacent-class changes in Table 10 remain evidence against broad functional invariance.

J Additional Limitations

The evaluation does not include replicated comparisons with other recent erasure methods, larger diffusion architectures, safety benchmarks, or multi-concept editing. The basis-refresh/current-vector pilots use a single training seed; the three-seed allocation suite tests one fixed random basis and one gate permutation. Consequently, the experiments support neither overall method superiority nor claims about safety, cross-architecture generalization, or a complete causal account of preservation.

Table 10: **Full object results.** Full object results across three SSGE training seeds. Target-text CLIP is used for standard, synonym, and compositional prompts; full-prompt CLIP is used for retain and adjacent prompts. Deltas are edited minus original, and parentheses report the sample SD across training seeds. LPIPS is computed on paired images.

Prompt set	Original CLIP	Edited CLIP	Δ (seed SD)	Paired LPIPS (seed SD)
Garbage truck				
Standard forget	0.292340	0.213231	−0.079110 (0.006027)	0.570940 (0.010075)
Held-out synonym	0.286639	0.216487	−0.070152 (0.005440)	0.577617 (0.011296)
Retain	0.299077	0.282798	−0.016279 (0.003612)	0.473942 (0.001731)
Held-out adjacent	0.291386	0.232887	−0.058499 (0.017118)	0.592187 (0.003378)
Held-out compositional	0.265223	0.204098	−0.061125 (0.001804)	0.568793 (0.011375)
Golf ball				
Standard forget	0.314323	0.221304	−0.093018 (0.015667)	0.757882 (0.052757)
Held-out synonym	0.303968	0.218894	−0.085073 (0.012237)	0.703690 (0.067093)
Retain	0.316646	0.267493	−0.049152 (0.009831)	0.596071 (0.032763)
Held-out adjacent	0.305274	0.220495	−0.084779 (0.021956)	0.737895 (0.072921)
Held-out compositional	0.293975	0.206432	−0.087543 (0.009046)	0.703596 (0.044307)
Church				
Standard forget	0.288514	0.224869	−0.063644 (0.005237)	0.676090 (0.013818)
Held-out synonym	0.281173	0.231993	−0.049179 (0.002675)	0.634622 (0.018926)
Retain	0.297706	0.281218	−0.016488 (0.003191)	0.487975 (0.008912)
Held-out adjacent	0.304695	0.271732	−0.032963 (0.002649)	0.566555 (0.000455)
Held-out compositional	0.252708	0.203593	−0.049115 (0.001364)	0.541884 (0.011078)
Parachute				
Standard forget	0.305022	0.227415	−0.077607 (0.007495)	0.758719 (0.068547)
Held-out synonym	0.241669	0.202395	−0.039274 (0.007514)	0.632098 (0.037522)
Retain	0.325580	0.285841	−0.039738 (0.015523)	0.480280 (0.056468)
Held-out adjacent	0.334853	0.255620	−0.079234 (0.017341)	0.552456 (0.055784)
Held-out compositional	0.281638	0.214095	−0.067543 (0.007090)	0.621982 (0.063618)
Cassette player				
Standard forget	0.312695	0.186631	−0.126065 (0.008653)	0.684681 (0.036059)
Held-out synonym	0.267981	0.196354	−0.071627 (0.010014)	0.600674 (0.037552)
Retain	0.308119	0.272045	−0.036074 (0.004891)	0.553689 (0.020616)
Held-out adjacent	0.309241	0.231346	−0.077895 (0.004752)	0.624397 (0.012019)
Held-out compositional	0.284714	0.179785	−0.104929 (0.003660)	0.651402 (0.032747)

K NUMERICAL VALUES FOR THE FIVE-OBJECT SUMMARY

Table 11: **Five-object erasure.** Five-object CLIP drops as mean (sample SD) across three SSGE training seeds. Larger target drops indicate stronger suppression, while smaller retain and adjacent drops indicate better preservation. Synonym, compositional, and adjacent prompt sets are held out.

Concept	Target CLIP drop ↑			Collateral CLIP drop ↓	
	Standard	Synonym	Compositional	Retain	Adjacent
Garbage truck	0.0791 (0.0060)	0.0702 (0.0054)	0.0611 (0.0018)	0.0163 (0.0036)	0.0585 (0.0171)
Golf ball	0.0930 (0.0157)	0.0851 (0.0122)	0.0875 (0.0090)	0.0492 (0.0098)	0.0848 (0.0220)
Church	0.0636 (0.0052)	0.0492 (0.0027)	0.0491 (0.0014)	0.0165 (0.0032)	0.0330 (0.0026)
Parachute	0.0776 (0.0075)	0.0393 (0.0075)	0.0675 (0.0071)	0.0397 (0.0155)	0.0792 (0.0173)
Cassette player	0.1261 (0.0087)	0.0716 (0.0100)	0.1049 (0.0037)	0.0361 (0.0049)	0.0779 (0.0048)

L SEMU ADAPTATION DIAGNOSTIC

We attempted an SD1.5 adaptation of SEMU (Sendera et al., 2025) to test whether its frozen-base, low-rank-adapter update transfers to our setting. We do not report a validated SEMU baseline; this section documents our unsuccessful adaptation. Each concept uses one fixed configuration (training seed 42, all matched evaluation prompts, paired original and edited images by prompt and generation seed), and the protocol was fixed before the runs: no hyperparameters or checkpoints were selected after viewing the outcomes. The failure is specific to this implementation and configuration, not a judgment about the original method.

Table 12: **SEMU SD1.5 adaptation.** One fixed training run per object (seed 42) with all matched evaluation prompts; positive CLIP drops and LPIPS are reported. This documented adaptation is not a verified reproduction of the authors' SD1.4 results.

	Target CLIP drop			Retain		Adjacent	
Concept	Standard	Synonym	Comp.	CLIP	LPIPS	CLIP	LPIPS
Garbage truck	0.0679	0.0594	0.0379	0.0869	0.9639	0.0787	0.9733
Golf ball	0.0690	0.0617	0.0497	0.1015	0.8844	0.0662	0.8336
Church	0.0691	0.0572	0.0311	0.0637	0.7424	0.1045	0.8303
Parachute	0.0743	0.0202	0.0526	0.1312	0.9171	0.1529	0.9034
Cassette player	0.1010	0.0547	0.0718	0.1075	0.9154	0.1084	0.9168
Macro mean	0.0762	0.0506	0.0486	0.0981	0.8846	0.1022	0.8915

Interpretation. Macro-averaged over the five concepts, the target CLIP decrease is 0.0762 on standard, 0.0506 on synonym, and 0.0486 on compositional prompts, while the retain CLIP decrease is 0.0981 and the adjacent CLIP decrease 0.1022: the collateral CLIP-score decreases exceed the target decrease. Paired LPIPS values (retain 0.8846, adjacent 0.8915) and edited target-recognition rates of 0.0% on every prompt set indicate severe image change, so the low target recognition cannot be read as selective erasure. On the same matched prompt-generation-seed streams, the SSGE three-seed macro means are target 0.0879 (standard condition), retain 0.0315, and adjacent 0.0667 (Table 8). We report these diagnostics to document the failure mode, not as a performance comparison: the single-run SEMU numbers carry no training-seed uncertainty, LPIPS is on a different scale from CLIP, and the failure is specific to our adaptation.

M ADDITIONAL EFFECT RESULTS

The grids separate method comparisons from training-run variation and show target, adjacent, and unrelated prompts at a common scale. Each comparison uses the same prompt and generation seed, with byte-verified original images across methods. The selection rules are described in Evaluation Integrity. These examples illustrate observed changes; they do not estimate success rates.

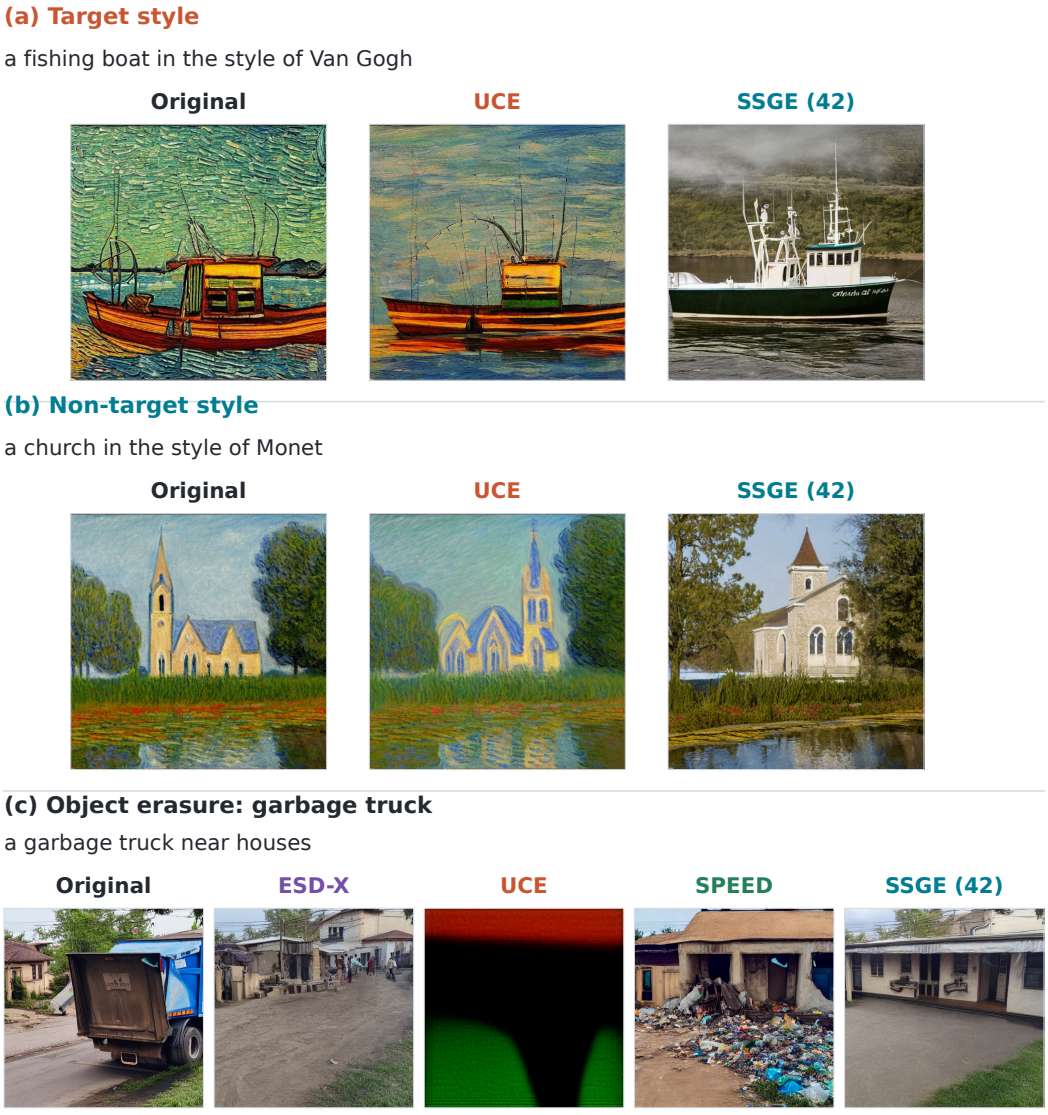


Figure 3: Matched style and object examples. Generation seeds are 42 for style and 3407 for garbage truck; SSGE and ESD-X use training seed 42, while UCE and SPEED are fixed edits. Prompts and generation seeds are identical within each comparison. The historical prompt/seed selections are retained when adding baseline columns; no new outcome filter is applied. Non-target style changes and the degraded object UCE output are shown without retouching.

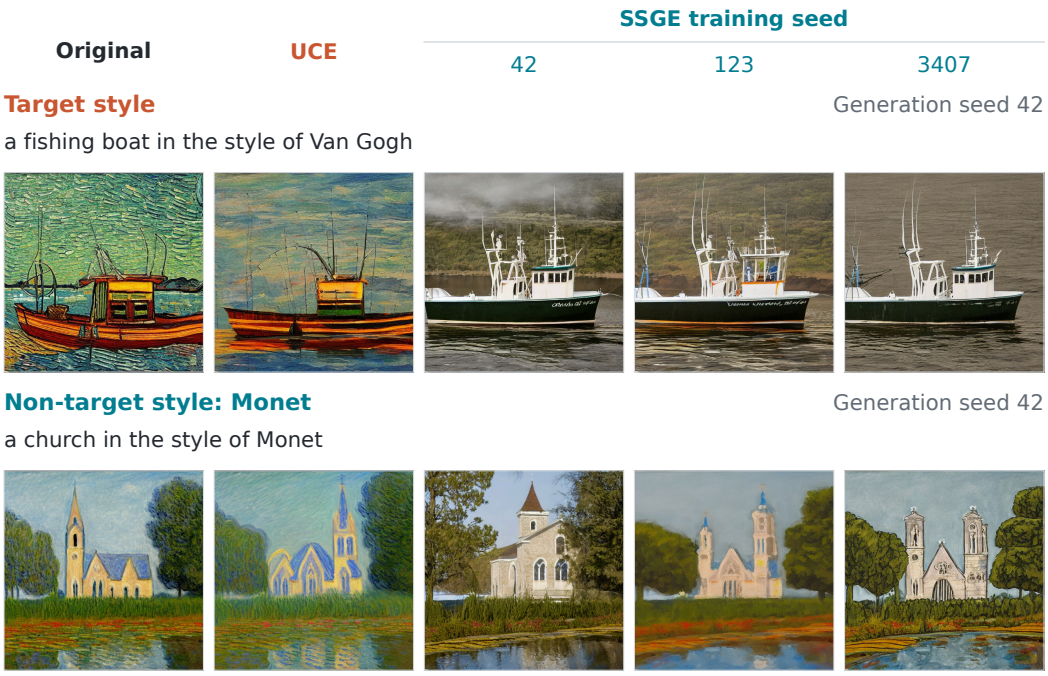


Figure 4: Van Gogh erasure across three SSGE training seeds. The first row is a target-style prompt; the second is the non-target Monet style. Subject retention does not imply style preservation: the Monet outputs visibly change. Original and UCE are repeated references, not additional independent runs.

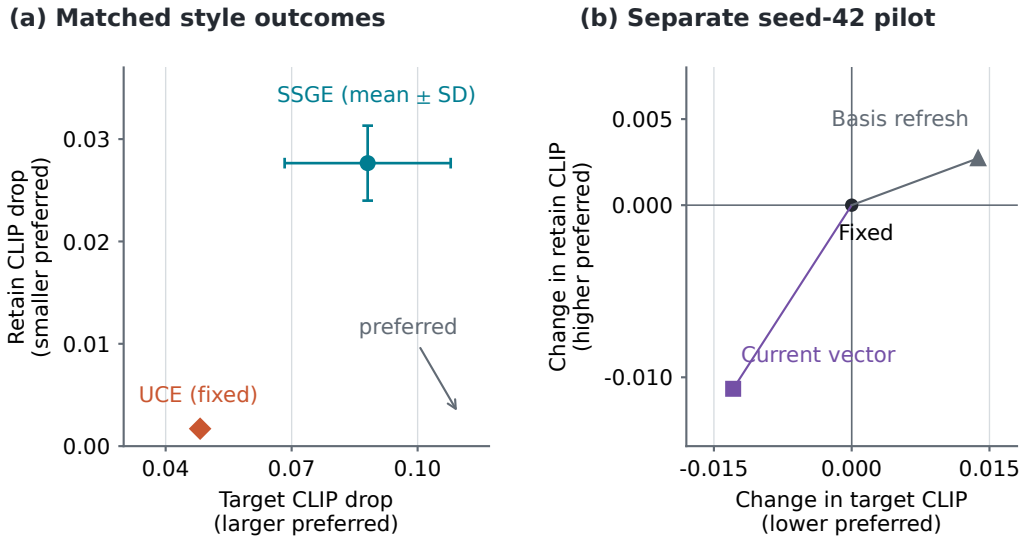


Figure 5: Style suppression and preservation. Left: matched main-protocol SSGE means with sample SD across three training seeds, and one fixed UCE edit. Right: a separate seed-42 basis pilot, shown as variant minus fixed-basis CLIP. Its reference checkpoint differs from the main protocol; these points are not pooled, and the pilot has no training-seed uncertainty estimate.

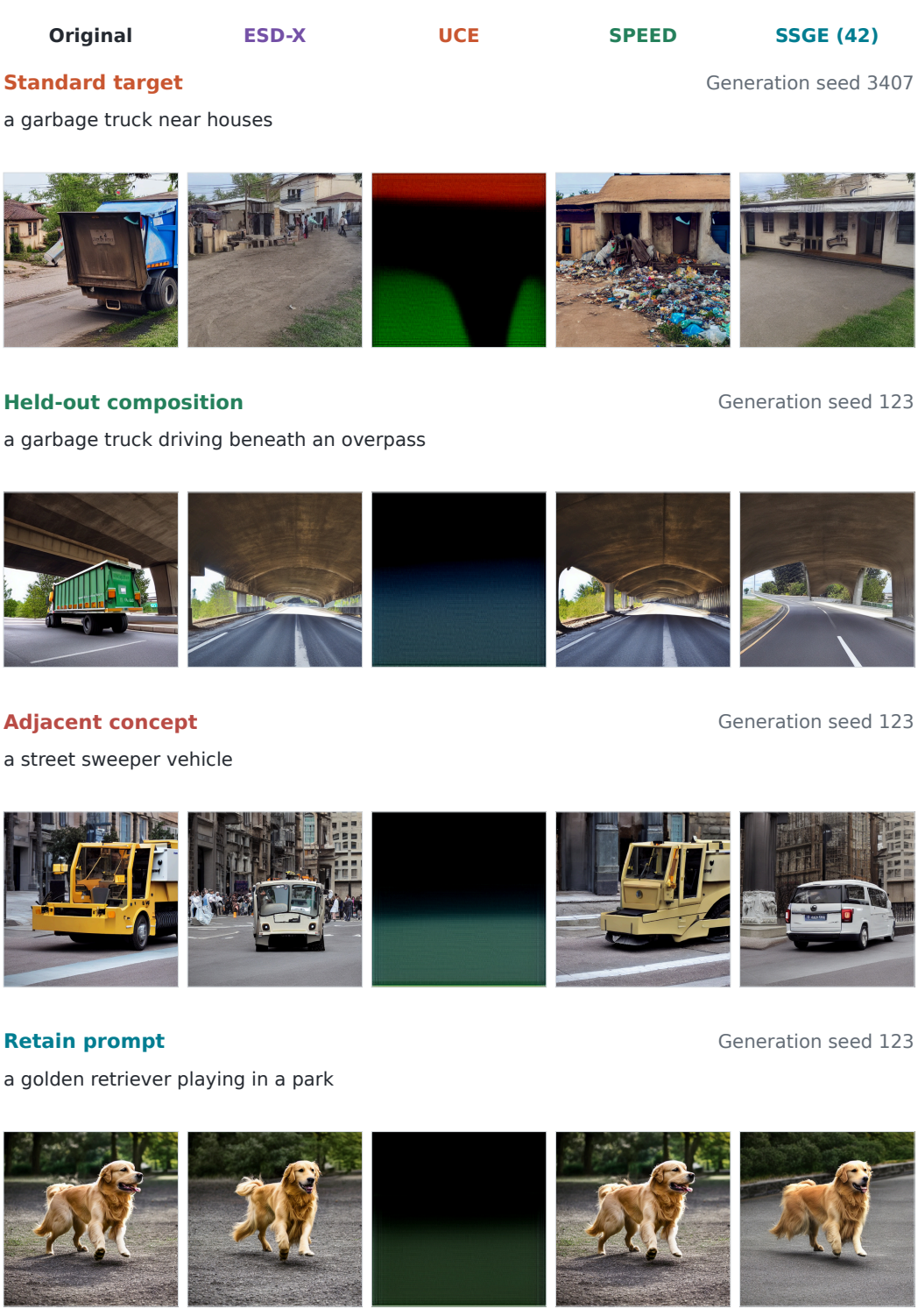


Figure 6: Matched garbage-truck method comparison. ESD-X and SSGE use training seed 42; UCE and SPEED are fixed edits. All methods use the row's indicated generation seed. The historical four prompt/seed selections are preserved for the added baseline columns, including degraded outputs. The rows expose target suppression, compositional transfer, adjacent-concept changes, and retain changes.

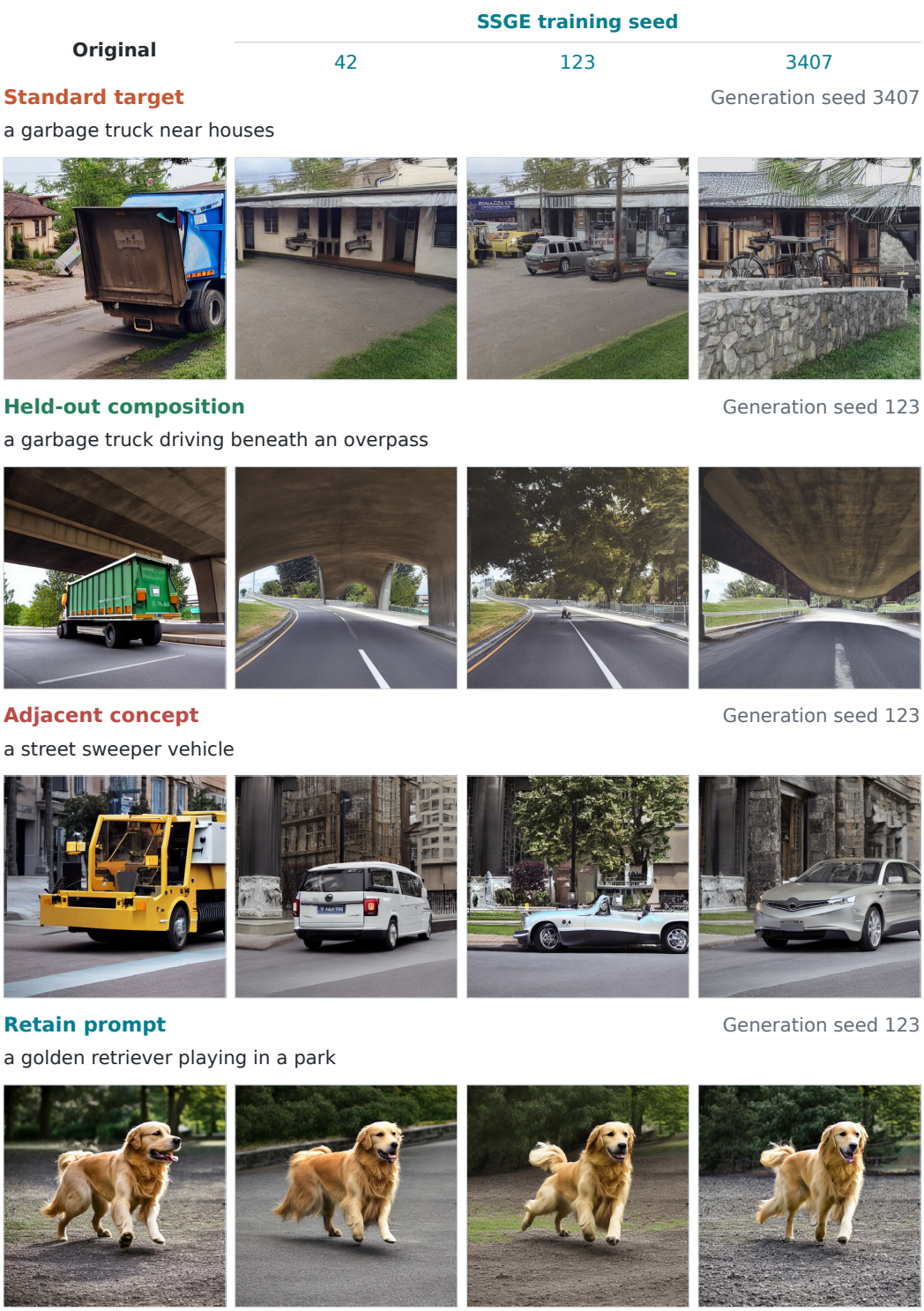


Figure 7: Garbage-truck erasure across three SSGE training seeds. The historical four prompt/seed selections are unchanged. Rows distinguish the standard target, held-out composition, adjacent concept, and unrelated retain prompt. Variation in adjacent and retain outputs is shown alongside target suppression.

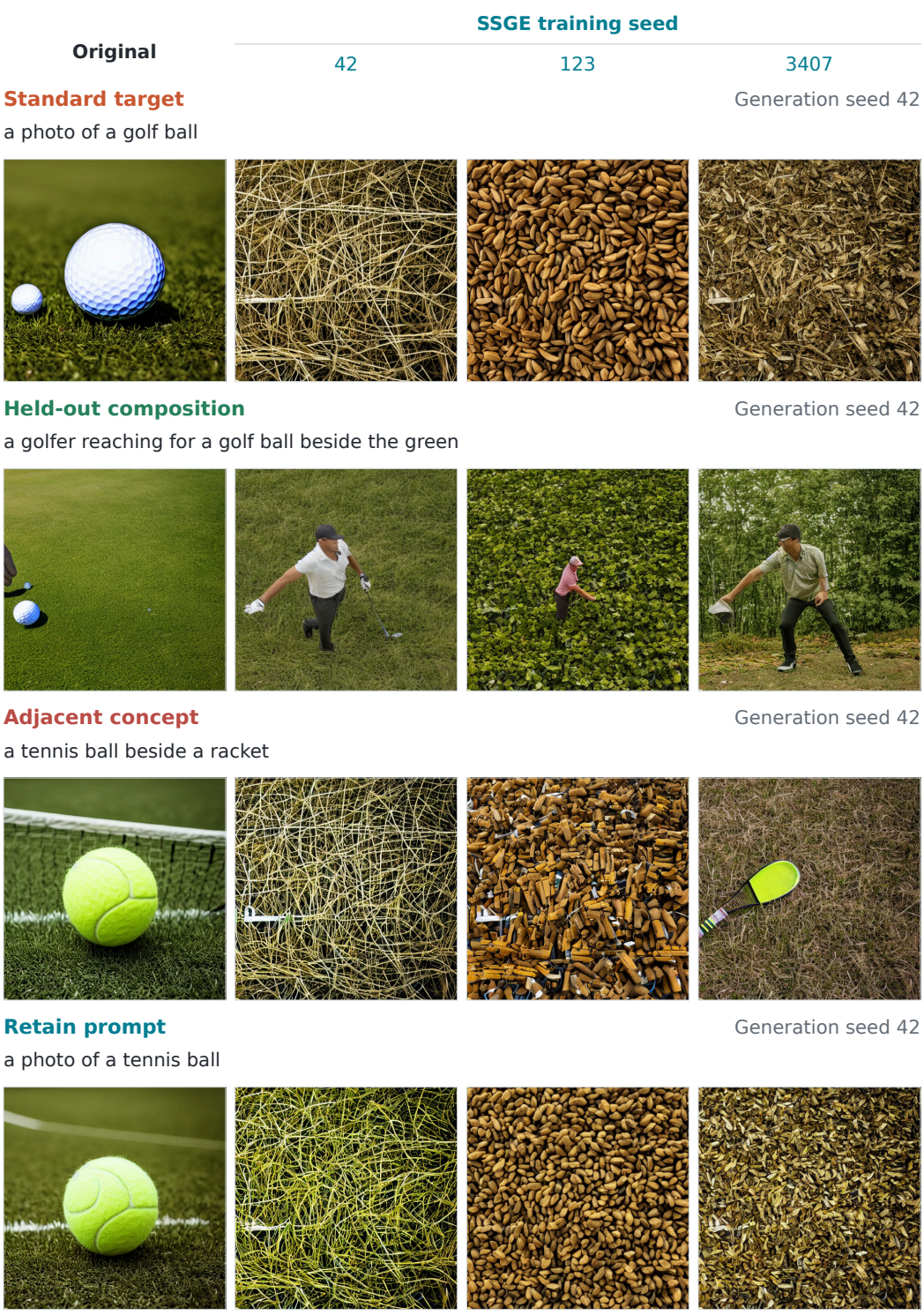


Figure 8: Golf-ball erasure across three SSGE training seeds. Each stream uses its first registered prompt and generation seed 42, without metric or visual outcome filtering. The adjacent and retain rows make collateral changes visible.

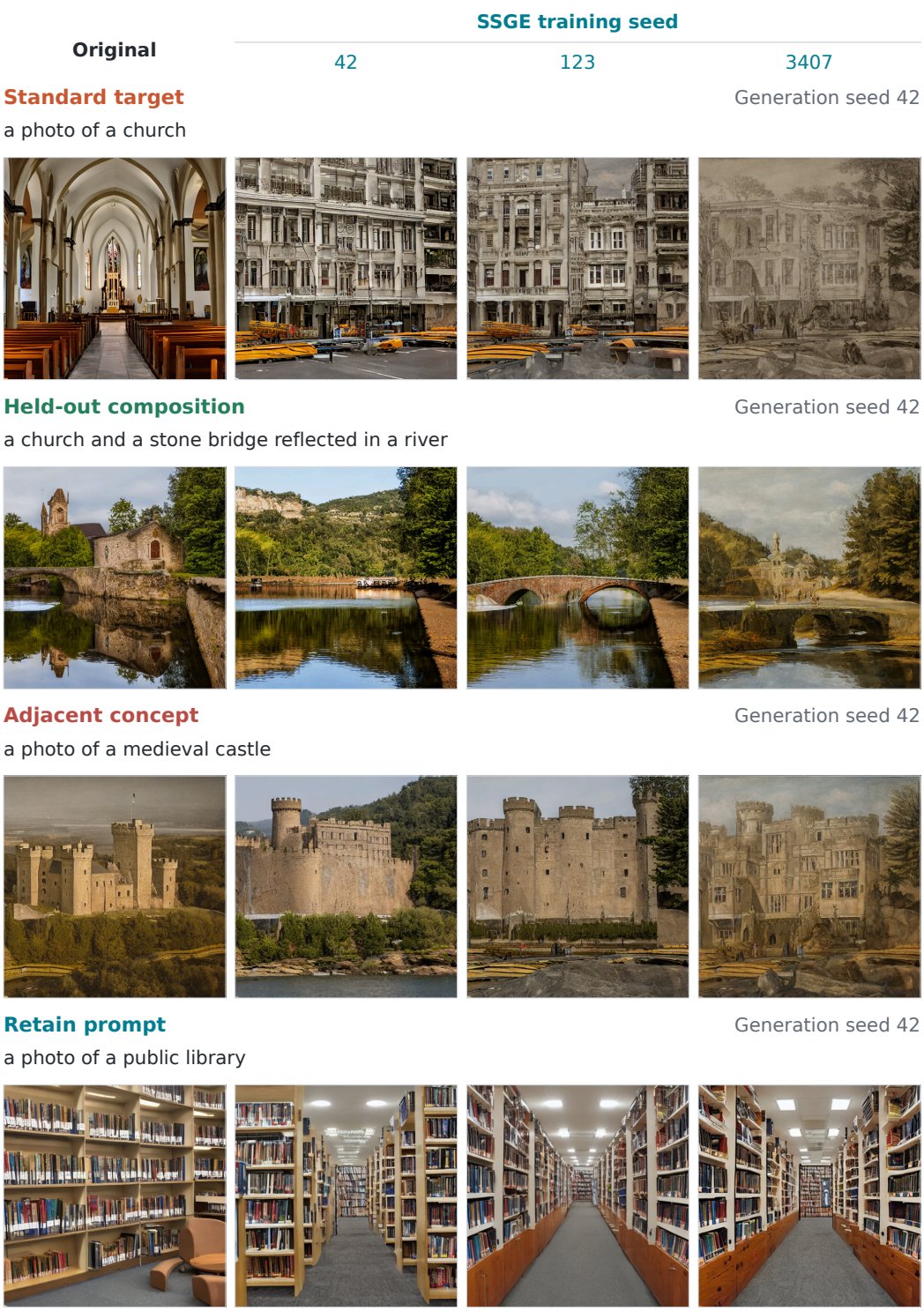


Figure 9: Church erasure across three SSGE training seeds. Each stream uses its first registered prompt and generation seed 42, without metric or visual outcome filtering. The adjacent prompt tests a related building rather than an unrelated retain concept.

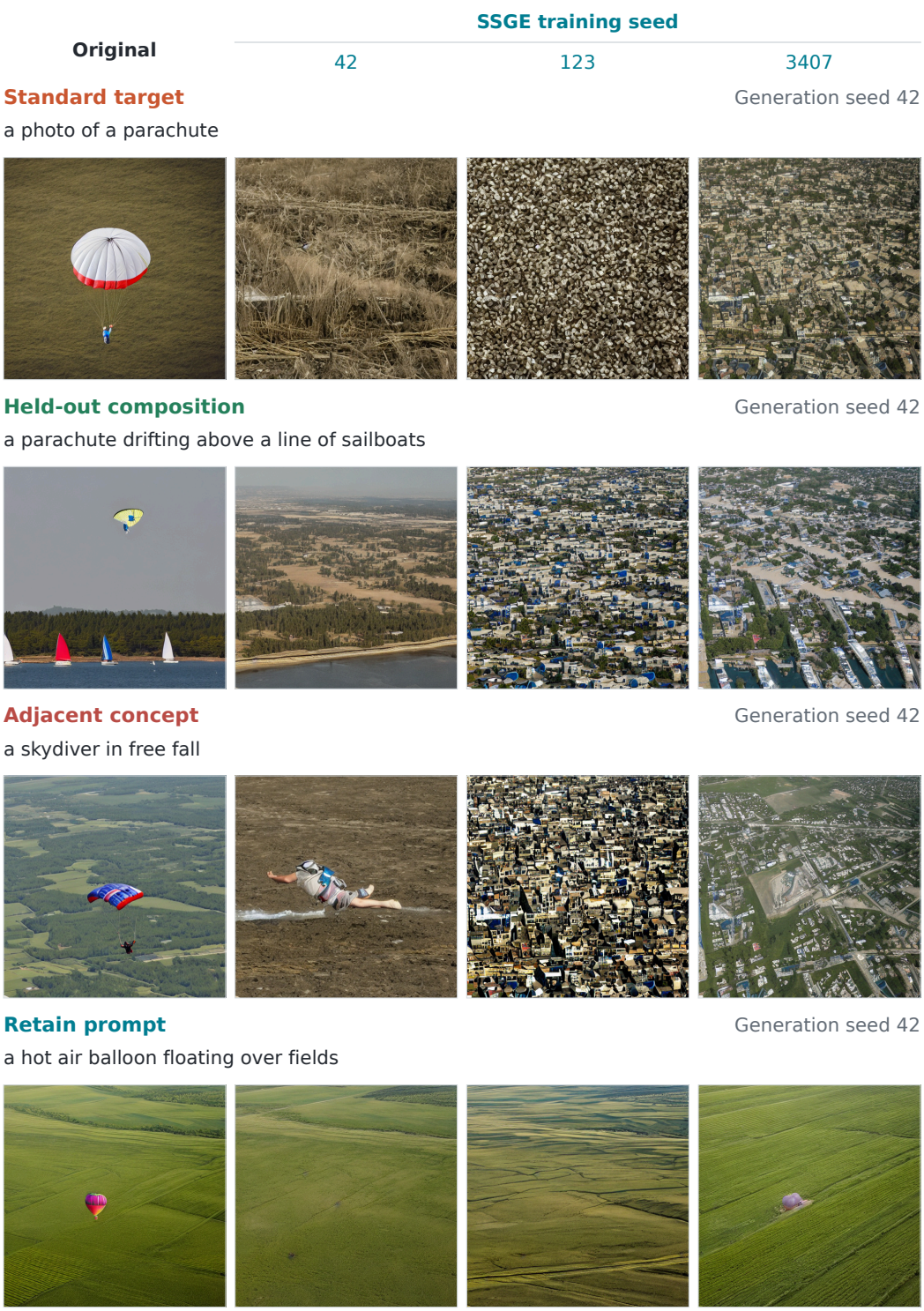


Figure 10: Parachute erasure across three SSGE training seeds. Each stream uses its first registered prompt and generation seed 42, without metric or visual outcome filtering. All images retain the complete original frame.

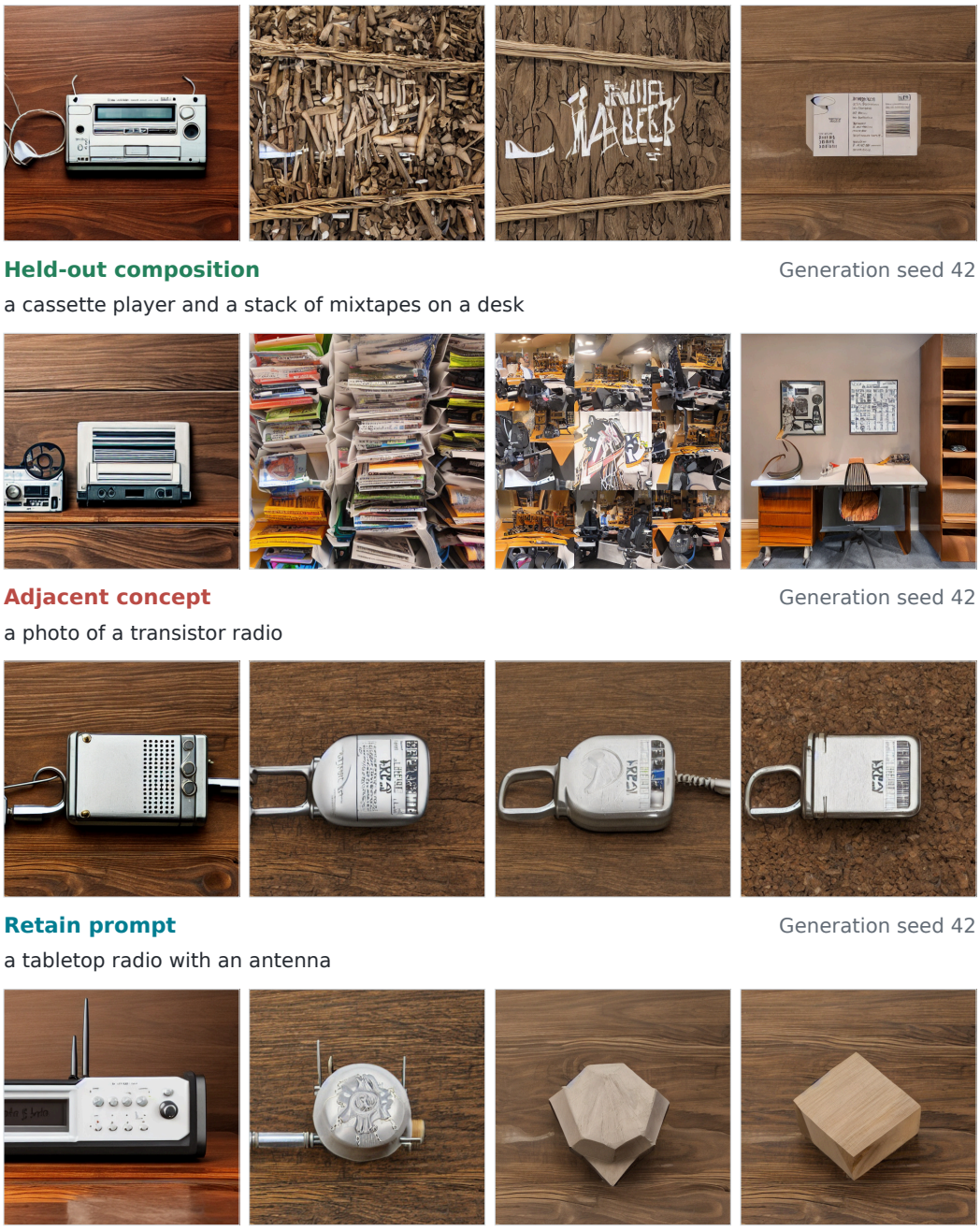


Figure 11: Cassette-player erasure across three SSGE training seeds. Each stream uses its first registered prompt and generation seed 42, without metric or visual outcome filtering. Imperfect original concept depiction is retained; these examples do not establish semantic deletion.