ERASED BUT NOT GONE: HOW MODEL MERGING UNDOES CONCEPT ERASURE IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept erasure suppresses a target concept in a diffusion model, but the edited weights may later be merged with another checkpoint. We study whether suppression survives composition with a public style or domain model. Averaging weakens the erasure update as it incorporates the partner, whereas additive task arithmetic keeps that update intact. We probe these changes separately and compare merge operators across suppression and partner-appearance transfer. Our study covers three object concepts in Stable Diffusion 1.5, with a 95-setting checkpoint/partner grid and a balanced 33-setting strength sweep. Attenuation restores target-concept evidence, but partner addition can also restore it without weakening the erasure update. At common minimum appearance-transfer targets, addition improves normalized suppression retention by 26.8 percentage points over interpolation; the tested parameter guards add 1.0 point over addition. These paired comparisons are retrospective and depend on the measured appearance estimator. The results identify strength-tuned addition as an important baseline and show why preserving an update in parameter space does not guarantee preservation of its behavioral effect.

1 INTRODUCTION

Concept-erasure methods modify a diffusion model to reduce generation of a selected object, style or other target (Gandikota et al., 2023; 2024; Fan et al., 2024). Yet suppression at the edited checkpoint need not survive later use. Adversarial prompts can expose residual capabilities (Tsai et al., 2024; Lu et al., 2025), and benign fine-tuning can restore suppressed concepts (Suriyakumar et al., 2026; Kim et al., 2024). We examine another common operation: merging the erased checkpoint with a public style or domain model.

Weight averaging combines compatible checkpoints without changing inference cost (Wortsman et al., 2022); Diffusion Soup applies this idea to diffusion models and style composition (Biggs et al., 2024). Our partner is derived from the same base as the erased model and is intended to contribute appearance characteristics. It is not optimized to attack erasure, although it may retain the target concept. The question is whether the composed model preserves suppression while acquiring that appearance.

Figure 1 illustrates why preserving an erasure update need not preserve suppression. Weight averaging entangles two changes: increasing the partner's weight introduces more of its update but reduces the erasure update's contribution. Task arithmetic (Ilharco et al., 2023) separates these changes: it adds the partner update to the erased checkpoint without reducing the erasure coefficient. This supplies an injection-only control and a practical baseline. It does not guarantee preserved behavior, because the partner update can change how the retained parameters affect generation.

A fair comparison must also account for what merging achieves. Low target scores may simply reflect weak appearance transfer. We therefore compare tested operators at common minimum appearance targets, then ask whether guards on selected parameters or directions improve on tuned addition. This connects interference-aware merging (Yadav et al., 2023) to the suppression–appearance trade-off. Earlier work combines separately trained safety and task adapters (Kim et al., 2024) or merges independently erased models (Lee et al., 2026). Our focus is a controlled comparison with external full-checkpoint partners, without retraining the partner or learning a guard.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

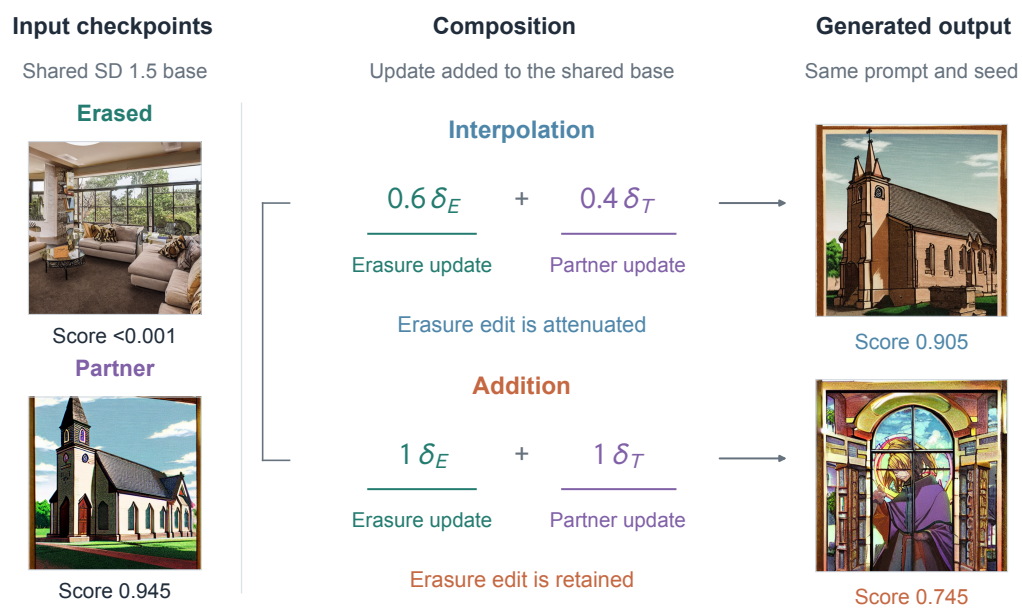

Figure 1: **Preserving an erasure update does not ensure preserved suppression.** The same erased and partner checkpoints supply both compositions. The center shows their update coefficients relative to the shared base: interpolation attenuates the erasure update δ_{E}, whereas addition retains it while adding the partner update δ_{T}. All images are unaltered ESD-u/church–anime outputs for "a church, anime", seed 13. Both compositions recover target evidence in this selected example. Scores are classifier outputs, not calibrated probabilities or recovery rates.

1.1 Contributions

- **A controlled test of erasure after composition.** We study three object concepts in Stable Diffusion 1.5, using a 95-setting checkpoint/partner grid and a balanced 33-setting strength sweep. Separate attenuation, injection and norm-matched random controls show that target evidence can return through both attenuation and partner injection.
- **An appearance-constrained operator comparison.** At common minimum appearance targets, addition gains 26.8 retention points over interpolation, while the tested guard family adds 1.0 point over addition. These are paired means under the recorded estimator; uncertainty and candidate-budget sensitivity are reported in Section 5.
- **An analysis of what preservation does and does not establish.** We distinguish preservation of an update from preservation of suppression, quantify the cost of stronger suppression in appearance transfer, and test sensitivity to cohort selection, metric definition and search budget. The results support tuned addition as a baseline for this setting, while leaving broader safety and generalization claims open.

2 Related Work

Concept erasure and resurgence. ESD fine-tunes a denoiser against a negatively guided target (Gandikota et al., 2023). UCE edits cross-attention projections in closed form (Gandikota et al., 2024). SalUn combines weight saliency with an unlearning objective, including random labeling for image generation (Fan et al., 2024). SPM uses adapters and MACE addresses many concepts (Lyu et al., 2024; Lu et al., 2024). These methods differ in objective, access and edited parameters, so results on our implementations should not be interpreted as a ranking of the published methods. Robustness evaluations include adversarial prompts, residual-capability probes and broader unlearning benchmarks (Tsai et al., 2024; Richardson et al., 2025; Lu et al., 2025; Zhang et al., 2024; Moon et al., 2025).

Adaptation after erasure. Suriyakumar et al. (2026) show that benign fine-tuning can restore previously suppressed concepts. Modular LoRA separates safety and task adaptation, combining their modules at inference (Kim et al., 2024). Thus, additive separation of a safety update is not new in itself. Our experiments instead start with independently obtained full checkpoints and compare training-free operators without retraining an adapter. Multi-concept immunization uses differentiable merging during training to make subsequent adaptation difficult (Zheng & Yeh, 2025); this addresses a different intervention and access model.

Merging diffusion models. Model Soups averages weights, task arithmetic combines updates relative to a common base, and TIES trims updates and resolves sign conflicts (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023). Diffusion Soup applies averaging to diffusion models trained on data shards and studies style mixing and shard-based unlearning (Biggs et al., 2024). Continual unlearning work merges independently erased diffusion models and reports sharp erasure transitions along base-to-erased interpolation paths (Lee et al., 2026). Our setup combines an erasure update with an external style/domain update, and evaluates their distinct effects and the resulting trade-off. SafeMERGE studies selective layer merging to restore alignment in language models (Djuhera et al., 2026). It is relevant to the broader safety-composition problem, but its task and selection procedure differ from those measured here.

3 Composition and measurement

3.1 Access model and operators

Figure 2 summarizes the pipeline. The composer has the erased UNet θ_{e}, the partner UNet θ_{p}, and their common base θ_{b}. Base access is necessary to recover $\delta_{\mathrm{E}} = \theta_{\mathrm{e}} - \theta_{\mathrm{b}}$ and $\delta_{\mathrm{T}} = \theta_{\mathrm{p}} - \theta_{\mathrm{b}}$. We assume compatible parameter shapes and retain the base text encoder and VAE throughout, including for the partner reference. No training data or gradients are used by the merge operators. The re-erasure comparison additionally uses the target concept, text encoder and eraser implementation.

Interpolation and addition. We evaluate

$$\theta_{\mathrm{interp}} = \alpha\theta_{\mathrm{e}} + (1-\alpha)\theta_{\mathrm{p}} = \theta_{\mathrm{b}} + \alpha\delta_{\mathrm{E}} + (1-\alpha)\delta_{\mathrm{T}}, \tag{1}$$

$$\theta_{\mathrm{add}} = \theta_{\mathrm{e}} + \lambda\delta_{\mathrm{T}} = \theta_{\mathrm{b}} + \delta_{\mathrm{E}} + \lambda\delta_{\mathrm{T}}. \tag{2}$$

Equation (1) changes two coefficients. We call reducing the erasure coefficient *attenuation*, and adding the partner delta *partner injection*. Injection may transfer target-generating behavior, disrupt suppression, or do both; a behavioral change alone cannot distinguish these possibilities. The attenuation-only control is $\theta_{\mathrm{b}} + a\delta_{\mathrm{E}}$. Injection-only is Eq. (2). A further control adds a fixed Gaussian direction to θ_{e}, scaled separately in each tensor to match $\lambda\delta_{\mathrm{T}}$'s norm.

Parameter guards. To test whether preserving selected directions improves on addition, we use

$$\theta_P = \theta_{\mathrm{e}} + \lambda(I-P)\delta_{\mathrm{T}}, \qquad P^2 = P. \tag{3}$$

This gives $P(\theta_P - \theta_{\mathrm{b}}) = P\delta_{\mathrm{E}}$ in exact arithmetic. Addition preserves the coefficient of δ_{E}; a guard additionally removes the projection of δ_{T}. Neither condition guarantees the same generated images or concept scores. We evaluate three choices:

- **Coordinate guard**: retain erased values where $|\delta_{\mathrm{E}}| \geq 10^{-4} \max|\delta_{\mathrm{E}}|$, with the maximum taken within each tensor; protect no entries of an unchanged tensor.
- **Attention guard**: retain all parameters whose names contain `attn2`, including query, key, value and output projections.
- **Row-space guard**: for each changed two-dimensional weight matrix, remove $\delta_{\mathrm{T}} V_r^\top V_r$, where V_r contains the leading r right-singular vectors of δ_{E}. Other tensors receive the full partner delta.

The strength sweep implements the row-space rule above. The earlier broad-grid implementation also projects unchanged matrices using their zero-matrix SVD basis; its results are labeled as the

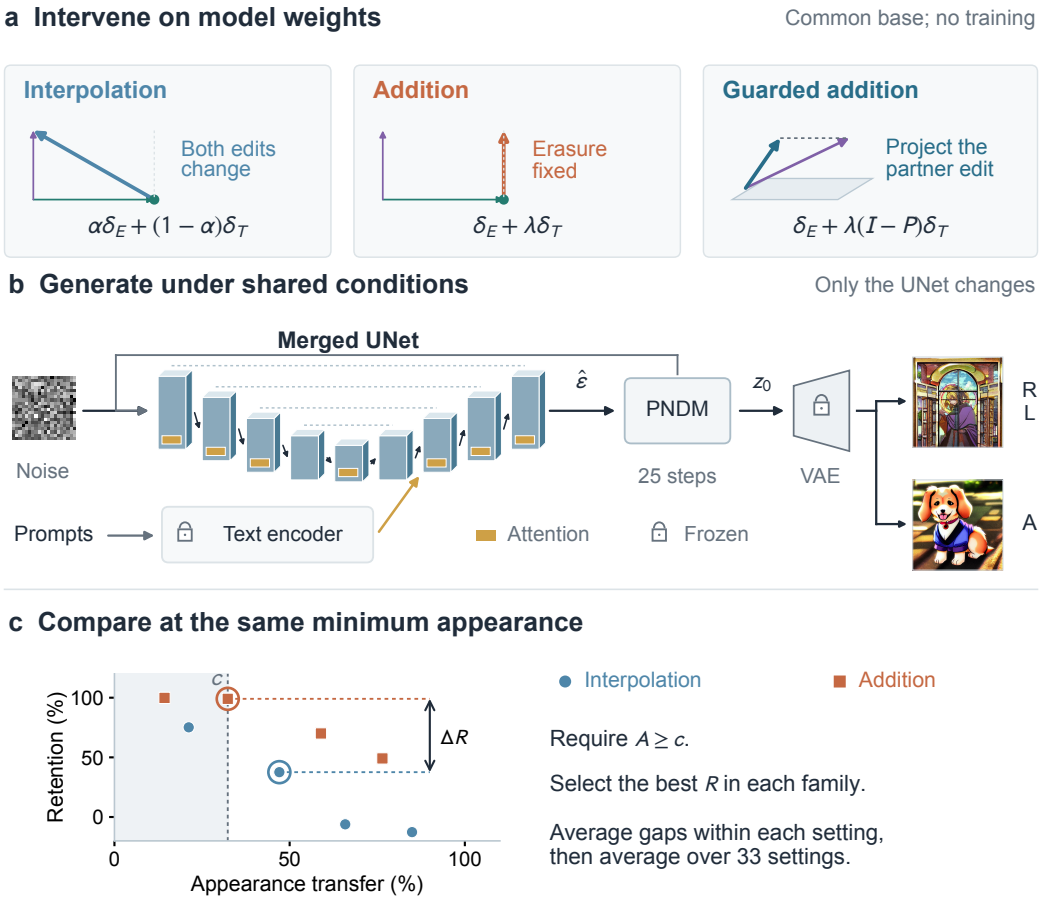

Figure 2: Compose, generate, compare. (a) Schematic update geometry; each expression is added to the common base. (b) Only the UNet changes. The sampler updates its latent input; the base text encoder and VAE stay frozen. Blocks summarize resolution stages, abstracting repeated layers and conditioning paths. Thumbnails are archived outputs; retention R additionally uses literal prompts. (c) Eight recorded ESD-u/church–anime candidates. Rings select the best feasible retention at minimum appearance c; achieved appearances differ. Gaps are averaged across reachable targets within settings, then across settings. Selection and evaluation share observations.

legacy variant. Appendix H documents this discrepancy. We do not treat the two variants as an identical implementation.

TIES is implemented with per-tensor magnitude trimming, sign election and averaging of agreeing entries, at density 0.5 and overall scale 1. Re-erasure applies the in-house UCE update to either interpolation at $\alpha = 0.6$ or addition at $\lambda = 1$. These are measured baselines, not exhaustive optimizations of either method.

3.2 Suppression, appearance and content

Let $s(\theta)$ be the mean ConvNeXt (Liu et al., 2022) ImageNet-1k target-class softmax score on literal concept prompts. We report normalized retention

$$R(\theta) = 100\left[1 - \frac{s(\theta) - s(\theta_e)}{s(\theta_p) - s(\theta_e)}\right]. \tag{4}$$

The erased endpoint has $R = 100$ and the partner has $R = 0$. Retention is not clipped and can lie outside $[0, 100]$. Paraphrase retention uses the same normalization with paraphrase scores. The *styled score* $L(\theta)$ is the raw mean target-class score when the concept is requested using the partner's trigger template. A classifier score is not a calibrated probability that the target is visibly present.

For each of eight neutral content prompts, let z_{is}^{θ} denote the normalized CLIP (Radford et al., 2021) image embedding of seed s, after applying the same partner template to every model. The appearance similarity is

$$C(\theta) = \frac{1}{8S^2} \sum_{i=1}^{8} \sum_{s=1}^{S} \sum_{t=1}^{S} (z_{is}^{\theta})^{\top} z_{it}^{\theta_{\mathrm{p}}}, \qquad A(\theta) = 100 \frac{C(\theta) - C(\theta_{\mathrm{e}})}{C(\theta_{\mathrm{p}}) - C(\theta_{\mathrm{e}})}. \tag{5}$$

Thus A measures movement toward the partner's generated appearance relative to the prompted erased model. It mixes style, content and composition. All seed pairs are included, including matched seeds; the partner endpoint also includes self-similarities. This reduces reliance on a single paired seed but does not remove common-noise effects. Appendix D tests exclusion of the diagonal on saved grid embeddings. Content alignment Q is 100 times CLIP image–text cosine against the unstyled content prompt, a complementary proxy rather than a comprehensive quality metric.

3.3 Comparing operator families at common appearance targets

A fixed coefficient is not comparable across different merge rules: it can transfer different amounts of the partner while retaining different amounts of suppression. For setting j, operator family F and minimum appearance target c, define the best observed feasible retention as

$$V_{j,F}(c) = \max_{m \in F:\, A_j(m) \geq c} R_j(m). \tag{6}$$

The candidates m are the measured strengths or guard configurations. We compare two families only at targets both can reach. Specifically, T_j contains their distinct observed appearance values with a feasible candidate in each family. Their setting-level difference and its overall mean are

$$D_j = \frac{1}{|T_j|} \sum_{c \in T_j} \big[V_{j,F}(c) - V_{j,G}(c) \big], \qquad \overline{D} = \frac{1}{J} \sum_{j=1}^{J} D_j. \tag{7}$$

A positive value favors the first family. Averaging targets before settings gives each checkpoint/partner pair equal weight, even when the number of reachable targets differs. We neither interpolate unmeasured candidates nor extrapolate beyond a family's observed range.

This comparison answers a constrained question: which tested family retains more suppression while meeting the same minimum appearance requirement? Its selected candidates can achieve different appearances above that requirement. The statistic is therefore not an exact appearance match or an area under a continuous frontier. Moreover, it selects and evaluates candidates on the same observations. We use it as a retrospective summary of the measured trade-off, with candidate-count sensitivity, rather than as evidence that a selection policy will generalize to new prompts or partners.

4 Experimental protocol

Models and cohorts. All experiments use Stable Diffusion 1.5 (Rombach et al., 2022) and three ImageNet objects: church, garbage truck and parachute. Four in-house configurations produce one erased checkpoint per object: ESD-u, ESD-x, UCE-EOS and saliency-masked ESD (SM-ESD). ESD-u updates parameters outside `attn2`; ESD-x updates only `attn2`. UCE-EOS edits key/value maps using end-of-text embeddings, an empty-string anchor and 38 preservation prompts. SM-ESD computes a saliency mask and applies the ESD loss to its selected non-cross-attention weights. It uses SalUn's masking idea but not its published generation objective. Training details are in Appendix A.

We retain the historical exclusion of ESD-x/church, whose separate verification recorded top-1 target frequency 0.28 and mean target score 0.238, exceeding the later gate of 0.1 on each quantity. The archive does not establish a uniform preregistered gate across all configurations. We therefore state the selection explicitly and give ungated grid results. The broad grid comprises 95 retained settings, where a setting is an erased-checkpoint/partner pair. It covers 13 partners with unequal counts from 1 to 11. The decisive strength sweep is balanced across anime, Realistic Vision and DreamShaper 8: 11 retained erased checkpoints times three partners, giving 33 settings. Appendix B shows the full coverage.

Table 1: **Composition operators on 95 grid settings.** Failure columns give percentages with $R < 90$ or $L > 0.1$; Q is content CLIP cosine multiplied by 100. Bold and underline mark best and second distinct displayed values, excluding references; ranks ignore appearance costs. Shading groups guards, all at $\lambda = 1$; the row-space guard uses legacy rank 4.

	Erasure retained (%)			Styled target		Appearance	Content
Method	Literal ↑	Para. ↑	Fail (%) ↓	Mean ↓	Fail (%) ↓	(%) ↑	Q ↑
Interpolation ($\alpha = 0.6$)	77.4	52.6	56.8	0.218	51.6	**43.0**	**26.82**
TIES (per tensor)	51.8	32.3	48.4	0.372	47.4	<u>28.6</u>	<u>26.53</u>
Addition ($\lambda = 1$)	96.1	85.2	11.6	0.045	14.7	27.8	26.07
Row-space guard (legacy)	97.3	86.6	8.4	0.030	<u>8.4</u>	27.1	26.00
Attention guard	<u>99.6</u>	<u>87.5</u>	<u>2.1</u>	<u>0.029</u>	<u>8.4</u>	22.5	25.89
Coordinate guard	**99.9**	**93.7**	**0.0**	**0.007**	**2.1**	13.4	25.72
Erased reference	100.0	100.0	0.0	0.006	1.1	0.0	25.51
Partner reference	0.0	0.0	100.0	0.728	96.8	100.0	27.40

Generation and sweeps. Sampling uses PNDM, 25 steps, guidance 7.5, 512×512 images and fp16 UNets. Within each run, all conditions share prompt/seed pairs. The grid uses seeds 0, 1 and 2; the strength and channel sweeps use 0, 1, 2 and 3. Literal suppression uses two templates, styled suppression one template, and appearance eight neutral prompts. The grid additionally uses four paraphrases per concept. Interpolation sweeps $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$ and addition sweeps $\lambda \in \{0.25, 0.5, 0.75, 1\}$. The guard family has six candidates: attention at $\lambda \in \{0.5, 1\}$, row-space at $r \in \{1, 4, 16\}$ with $\lambda = 1$, and coordinate at $\lambda = 1$. TIES is not fully strength-tuned. The channel controls use two partners, anime and Realistic Vision, with all 11 retained checkpoints.

Comparison and uncertainty. We use the within-setting comparison in Section 3.3. Addition and interpolation each have four candidates; the guard family has six. To assess sensitivity to that difference, we repeat the guard comparison for every four-candidate subset. All candidates and excluded settings remain available in the supplement.

We report 95% percentile intervals from 20,000 bootstrap resamples of erased-checkpoint clusters, retaining every partner and target for a sampled checkpoint. These intervals condition on the selected partners and do not represent uncertainty over unseen architectures or concepts. We also report setting-bootstrap and leave-one-partner-out sensitivity. The thresholds $R < 90$ and $L > 0.1$ describe settings with degraded suppression; they do not certify safety. Threshold sensitivity is included in the appendix.

Metric and utility limitations. A recorded consistency check on 1,584 selected images finds 98.1% threshold agreement between ConvNeXt and zero-shot CLIP. Its positives and negatives are assigned by prompt provenance, not human annotation, so it cannot estimate real detection error. Moreover, all eight neutral content prompts occur in UCE-EOS's preservation set. The re-erasure utility comparison therefore measures retention of these preserved contents, not transfer to a held-out content set.

5 Results and Analysis

5.1 Interpolation often degrades suppression

Table 1 summarizes the broad grid. At $\alpha = 0.6$, interpolation has mean retention 77.4% on literal prompts and 52.6% on paraphrases. Retention is below 90% in 54/95 settings, and styled score exceeds 0.1 in 49/95. Addition at $\lambda = 1$ improves these counts to 11/95 and 14/95, but transfers less appearance on average. A fixed-strength comparison alone therefore cannot determine the best trade-off.

The guards lower failure counts at their reported operating points, with the coordinate guard reducing them to 0/95 and 2/95. It also lowers mean appearance transfer from 27.8% for addition

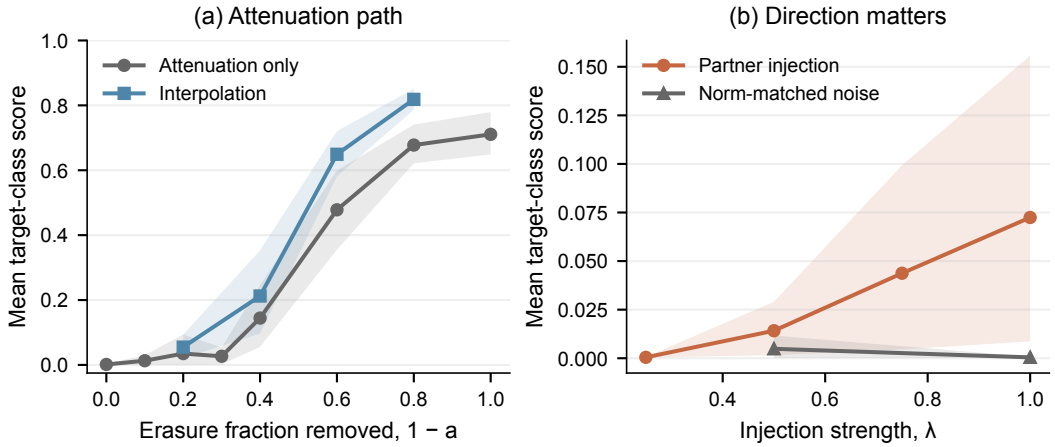

Figure 3: Separate weight-space controls. Left: attenuation alone versus interpolation, plotted against the fraction of the erasure update removed. Right: partner injection with the erasure coefficient fixed, versus a per-tensor norm-matched Gaussian direction. Lines are means over 11 erased checkpoints, averaging the two partners within checkpoint where applicable; bands are 95% checkpoint-bootstrap intervals. The right panel uses a narrower vertical scale to show residual recovery.

to 13.4%. TIES performs poorly in this particular per-tensor, fixed-density implementation. These outcomes motivate appearance-constrained comparisons; they do not establish that guarding is uniformly better or that TIES cannot be tuned successfully.

5.2 Both coefficient attenuation and partner direction matter

The separate paths in Figure 3 resolve what the grid cannot. Reducing the erasure coefficient from 1 to 0.6 raises the mean literal target score from 0.00145 to 0.14445. Interpolation at the same erasure coefficient reaches 0.21242. Attenuation alone therefore accounts for a large change, but the path comparison is not an additive causal decomposition of a nonlinear generator.

Full partner injection reaches 0.07244, whereas norm-matched noise reaches 0.000445. Their paired difference is 0.072 [0.008, 0.155]. Keeping the erasure coefficient fixed consequently does not preserve its behavioral effect. The random control tests one fixed direction per tensor, reused across conditions, and measures concept score only; a low score under noise is not evidence that image quality is maintained.

The strength sweep also injects only the partner's cross-attention delta or only its complement. For literal score, the interaction contrast is

$$I = s(\theta_{\mathrm{e}} + \delta_{\mathrm{T}}) - s(\theta_{\mathrm{e}} + \delta_{\mathrm{T},x}) - s(\theta_{\mathrm{e}} + \delta_{\mathrm{T},\neg x}) + s(\theta_{\mathrm{e}}). \tag{8}$$

Its mean is 0.033 [-0.003, 0.082]. This contrast measures departure from additivity of the score at these endpoints. It cannot identify a unique concept-carrying subspace or show that all useful guards must sacrifice appearance in the same proportion.

5.3 Strength-tuned addition is the stronger starting point

Across the balanced sweep, addition exceeds interpolation by 26.8 retention points [17.8, 35.8] at common minimum appearance targets (Figure 4). Leaving out each partner in turn gives mean improvements from 22.1 to 33.0 points; the setting-bootstrap interval is in Appendix C. This supports addition as a baseline for this composition setting, without implying that any fixed strength is safe.

The six-candidate guard family gains 1.0 retention points [0.2, 2.0] over four-candidate addition. Across all 15 four-guard subsets, the median gain is 0.32 points and the range is -0.09 to 0.65. Subsets also change attainable appearance targets; the range is not a confidence interval or a pure

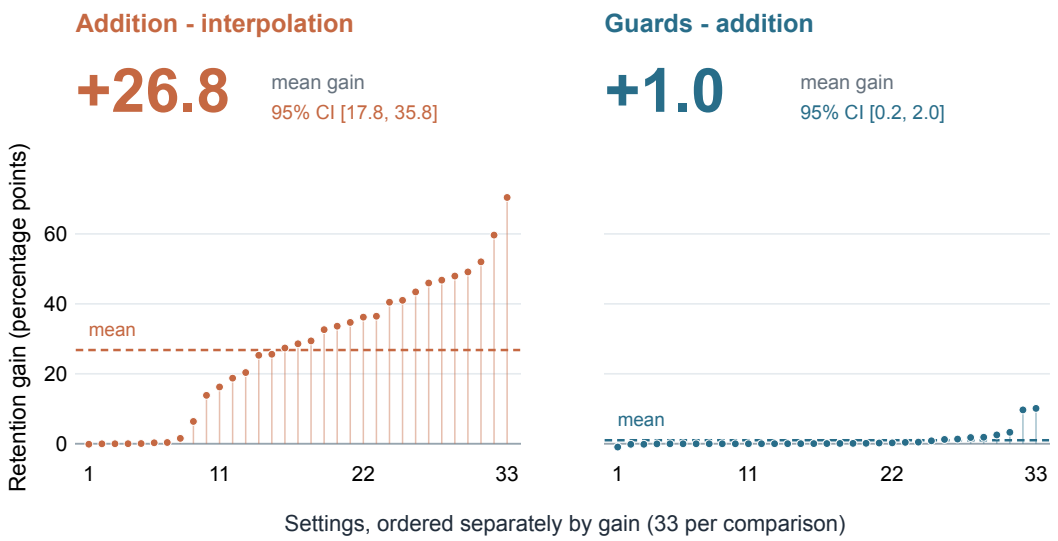

Figure 4: Paired retention gains, ordered within each comparison. Each dot is one setting's gain; the 33 settings are sorted separately, so horizontal ranks do not pair settings across panels. Both panels use the same vertical scale. Dashed lines and headline values show means; intervals are 95% checkpoint bootstraps. Settings share erased checkpoints. Guards use six candidates versus four for addition; the equal-budget median gain is 0.32 points

estimate of search-budget bias. The tested guards add a smaller, candidate-sensitive increment, without ruling out improvements from other guards or denser tuning.

The advantage varies across partners: mean additive gains are 36.1 points for anime, 14.4 for Realistic Vision and 29.9 for DreamShaper 8. Appendix C reports their intervals and guard increments. These groups share erased checkpoints and do not establish transfer to unseen partners.

5.4 Re-erasure and the cost of stronger suppression

Table 2 gives selected operating points, with the full sweep in Appendix E. Addition at $\lambda = 1$ satisfies both reporting thresholds in 21/33 settings; reducing λ to 0.25 raises this to 31/33 and lowers mean appearance transfer from 53.9% to 15.4%. The attention guard at $\lambda = 1$ and addition at $\lambda = 0.75$ each pass in 26/33 settings, at similar mean appearance transfer.

Re-erasing the additive model attains mean appearance transfer 32.4% and passes both thresholds in 31/33 settings. Relative to the additive envelope at its achieved transfer, its retention advantage is 3.1 [0.4, 6.4] over 30 comparable settings. This is a useful measured operating point, not evidence of universal dominance: re-erasure changes the access model, only two such points were tested, and its utility prompts overlap the preservation set.

An illustrative failure beyond the aggregate. Figure 1 shows a saved ESD-u/church example in which addition recovers target evidence despite retaining the erasure coefficient. The attention guard also recovers recognizable target structure, while the coordinate guard keeps a low score but alters the neutral appearance (Appendix I). The complete saved seed panels illustrate why an average operator advantage cannot certify individual outputs; they do not estimate recovery frequency.

Geometry is informative but not a behavioral guarantee. In the 89 grid settings with recorded geometry, the log ratio $\log_{10}(\|\delta_T\|/\|\delta_E\|)$ correlates with additive appearance transfer at $r = 0.83$. This is a descriptive association, partly reflecting distinct partner groups, not a validated predictor for unseen checkpoints. Guarded update energy is computable without generation, but appearance cost must be measured (Appendix G).

Table 2: **Selected operating points over 33 balanced settings.** "Both" counts $R \geq 90$ and $L \leq 0.1$, not certified safety. Ranks follow Table 1; time is unranked. Times are median recorded composition timers, not uniform end-to-end runtimes (Appendix E). Re-erasure requires the concept and preservation set.

Composition	Appearance (%) ↑	Retention (%) ↑	Styled score ↓	Both of 33 ↑	Compose time (s)
Interpolation $\alpha = 0.6$	<u>44.9</u>	72.4	0.271	8	5.7
TIES (per tensor), $d = 0.5$	43.7	19.2	0.649	5	92.0
Addition $\lambda = 0.25$	15.4	<u>99.8</u>	0.013	<u>31</u>	2.5
Addition $\lambda = 0.75$	43.2	94.3	0.059	26	2.6
Addition $\lambda = 1$	**53.9**	91.1	0.084	21	2.4
Attention guard $\lambda = 1$	41.6	96.5	0.042	26	2.7
Coordinate guard $\lambda = 1$	27.2	**99.9**	0.012	<u>31</u>	3.8
Interpolation + UCE-EOS	28.0	99.6	**0.003**	**32**	4.6
Addition + UCE-EOS	32.4	99.5	<u>0.007</u>	<u>31</u>	4.4
Erased reference	0.0	100.0	0.003	33	–
Partner reference	100.0	0.0	0.706	0	–

Sensitivity to the appearance estimator. Removing matched-seed similarities from the saved grid embeddings changes mean transfer from 43.0% to 57.1% for interpolation and from 27.8% to 33.4% for addition. The operator ordering is stable, but the scale depends on the estimator. The strength sweep lacks the embeddings needed to repeat its paired comparison off-diagonal, so this check does not validate that comparison under an alternative metric (Appendix D).

6 Limitations and discussion

The study covers one architecture, three benign object concepts, a selected set of public partners and one trained checkpoint per erasure configuration/concept. It does not establish robustness for harmful content, identities, artistic styles, newer diffusion architectures or adversarially chosen weights. The 95-setting grid is unbalanced, and the 33-setting sweep contains only three partners. Bootstrap intervals condition on this design; additional generation seeds would not replace new concepts, partners or training replicates.

The metrics introduce further limits. ConvNeXt and CLIP agreement is not ground truth. Appearance transfer depends on the prompts, shared seeds and the partner reference, and it does not isolate style or perceptual quality. Selecting operator candidates on the evaluation observations makes the envelope an optimistic retrospective summary. Re-erasure is evaluated on contents also present in its preservation set. A held-out evaluation with independent seeds and blinded labels is needed before turning these comparisons into a deployment policy.

The erasers are in-house implementations, the two archived row-space implementations differ, and checkpoint-revision metadata are incomplete. These limits restrict exact reproduction and comparison with published baselines. Evaluate the composed model itself, include strength-tuned addition, and report residual concept scores alongside appearance and content transfer. Preserving parameter coefficients alone does not establish preserved suppression.

7 Conclusion

Composing an erased SD 1.5 checkpoint with a public partner can restore target-concept scores. Interpolation attenuates the erasure update; addition avoids this change and improves the measured retention–appearance trade-off in our balanced sweep. Partner updates still cause recovery beyond norm-matched noise. Guard gains are small and candidate-sensitive, while re-erasure provides distinct operating points. These findings support post-composition evaluation and tuned additive baselines, without establishing safety guarantees or transfer to other architectures.

AI USE STATEMENT

Yes. Generative AI tools were used to aid and polish the writing, improve clarity and presentation, and assist with minor code and manuscript editing. All scientific content, results, analyses, and conclusions were reviewed and verified by the authors.

ETHICS STATEMENT

The experiments use benign object concepts to study the persistence of suppression under weight composition. Reversibility of concept erasure is relevant to evaluating released models, but the same observations may inform attempts to recover suppressed outputs. The study does not demonstrate safe deployment or removal of training-data influence. We do not test harmful-content targets, and the existing capabilities and risks of the base and partner models remain. Model use and redistribution are subject to their respective licenses.

REPRODUCIBILITY STATEMENT

The supplement includes the archived grid, channel and strength-sweep records, analysis scripts, figure-generation code, selected original generated images, and the model-generation implementations. Appendix A specifies the recorded protocol, and Appendix H documents implementation differences. Numerical tables and figures can be regenerated without model inference. Erased checkpoints and full checkpoint-revision manifests are absent from the available archive; model generations have not been independently rerun for this revision. The scripts support regeneration, but the missing metadata prevent a claim of exact end-to-end reproduction.

REFERENCES

Benjamin Biggs, Arjun Seshadri, Yang Zou, Achin Jain, Aditya Golatkar, Yusheng Xie, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Diffusion soup: Model merging for text-to-image diffusion models. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2406.08431.

Aladin Djuhera, Swanand Ravindra Kadhe, Farhan Ahmed, Syed Zawad, and Holger Boche. Safemerge: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. In *Findings of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2503.17239.

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 (ICLR)*, 2024.

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

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

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Sanghyun Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Safety alignment backfires: Preventing the re-emergence of suppressed concepts in fine-tuned text-to-image diffusion models. *arXiv preprint arXiv:2412.00357*, 2024.

Justin Lee, Zheda Mai, Jinsu Yoo, Chongyu Fan, Cheng Zhang, and Wei-Lun Chao. Continual unlearning for text-to-image diffusion models: A regularization perspective. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.07970.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.17013.

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 (CVPR)*, 2024.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Saemi Moon, Minjong Lee, Sangdon Park, and Dongwoo Kim. Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2410.05664.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Alex D. Richardson, Kaicheng Zhang, Lucas Beerens, and Dongdong Chen. Rethinking the vulnerability of concept erasure and a new method. *arXiv preprint arXiv:2502.17537*, 2025.

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 (CVPR)*, pp. 10684–10695, 2022.

Vinith M. Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C. Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=Vj0Z2wspQ5.

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.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, pp. 23965–23998, 2022.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Kompella, Xiaoming Liu, and Sijia Liu. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Amber Yijia Zheng and Raymond A. Yeh. Multi-concept model immunization through differentiable model merging. In *AAAI Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2412.15320.