BEYOND SUPPRESSION: MEASURING CONCEPT ERASURE DURABILITY IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper first investigates a vulnerability that quantization can reverse concept erasure in diffusion models. To this end, we first quantize diffusion models edited by different concept-erasure methods, and the results demonstrate that the target concept remains suppressed after quantization. However, the quantized erased model exhibits a novel vulnerability: restoring just 5% of the edited weights to their original values can markedly increase the generation rate of the erased concept. In this paper, we refer to the ability to keep suppression under such restoration as *erasure durability* and evaluate it by tracking target generation across restoration budgets. This yields a recovery curve showing how target generation changes as the restoration budget increases. We summarize this curve with the half-recovery budget, defined as the minimum restoration needed to raise the target-generation rate to the midpoint between its erased-model baseline and fully restored level. Based on this, we construct DuraBench, a durability-based benchmark that distinguishes erasure methods with similar initial suppression but different resistance to weight restoration. We further propose SPARC, a post-training method that helps erased models maintain target suppression when some edited weights are restored.

1 INTRODUCTION

Concept-erasure methods edit a pretrained diffusion model (Ho et al., 2020; Rombach et al., 2022) to suppress target concepts while preserving its other uses (Gandikota et al., 2023; 2024; Gong et al., 2024; Lu et al., 2024). In the meantime, deployment requires efficiency, and quantization stands as a typical model compression approach (Li et al., 2025b; Chen et al., 2025; Li et al., 2023). Quantization maps the erased model's parameters to a lower-precision representation, potentially causing information loss (Zhang et al., 2025). We therefore ask: *Can this information loss undo the effect of concept erasure?*

To investigate this, we compare target-generation rates before and after INT8 quantization. Our answer is both yes and no: i) As illustrated in Figure 1a, quantization alone leaves target-generation rates nearly unchanged. However, this apparent stability masks a vulnerability to weight restoration: As shown in Figure 1b, restoring just 5% of the edited weights can significantly increase target generation across multiple erasure methods.

Our further experiments find that the increase in target generation varies across different erasure methods, even when the same fraction of edited weights is restored. To systematically characterize these differences, we construct DuraBench, a benchmark for evaluating *erasure durability*: the ability to maintain target suppression under weight restoration. By progressively restoring edited weights and measuring target-generation rates, we obtain a normalized *recovery curve* for each erased checkpoint. As shown in Figure 1c, we summarize each curve with the *half-recovery budget* (diamonds mark), defined as the restoration fraction needed to recover 50% target-generation rate of fully restored level. A larger half-recovery budget indicates greater erasure durability.

DuraBench evaluates 14 erased checkpoints across 11 restoration budgets and reveals differences in durability. First, restoring a small fraction of edited weights can substantially reverse concept erasure. For instance, SPM-32 (Lyu et al., 2024) reaches half recovery with only 1% of its edited weights restored. Second, checkpoints with similar target concept suppression can exhibit different durability: SafeGen (Li et al., 2024) and AC-full (Kumari et al., 2023) have similar initial target

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.

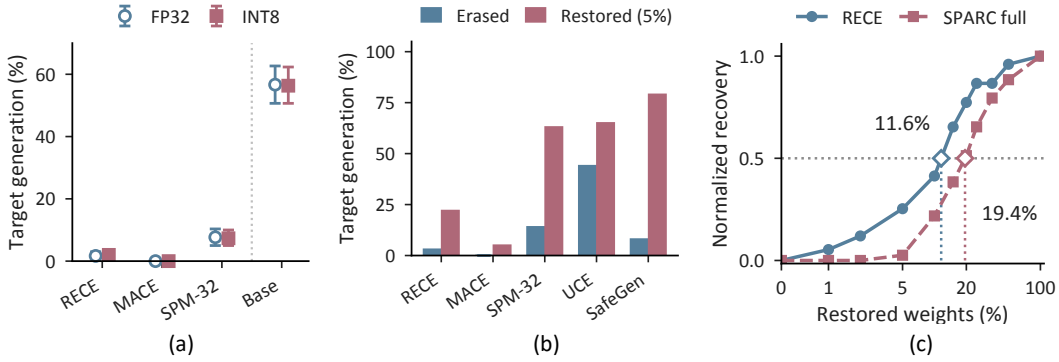


Figure 1: Effects of quantization and weight restoration on concept erasure. (a) Target generation rate of FP32 and INT8 checkpoints. (b) Target generation rate before and after restoring 5% of edited weights in quantized models. (c) FP32 recovery curves for RECE and SPARC. We highlight the half-recovery budgets.

generation, yet their half-recovery budgets are 1.41% and 16.35%, respectively. Third, a larger half-recovery budget can correspond to fewer restored weights, making absolute weight counts a complementary measure of durability.

These findings motivate treating resistance to weight restoration as a design objective alongside initial suppression. To further understand this, we extract a key suppression component from an erased model's weight edits. Keeping this component fixed, we find that increasing the magnitude of other components improves erasure durability. Our analysis indicates that these additional components bear more of the impact of weight restoration, helping preserve the key component and its suppression effect. Guided by this observation, we propose Suppression-Preserving Adaptation via Restoration Consistency (SPARC), a post-training method for strengthening erasure durability. During post-training, SPARC learns weight edits while using the erased checkpoint's responses to target prompts as the reference. It matches this reference both before and after weight restoration, making the learned edits to maintain target suppression even when some edited weights are reverted. At a 5% restoration budget, SPARC reduces target-concept recovery by approximately 90% on RECE and 42% on AC-crossattn. Our contributions are as follows:

- We show that concept erasure can remain effective after quantization yet be weakened by restoring only a small fraction of edited weights.
- We introduce DuraBench to compare erasure durability using recovery curves and half-recovery budget. It reveals differences between erasure methods with similar suppression.
- We show that additional edits can reduce target recovery while the key suppression remains effective. Guided by this finding, we propose SPARC, a post-training method that incorporates weight restoration to improve erasure durability.

2 Related Work

2.1 Concept Erasure in Diffusion Models

Concept-erasure methods can be broadly divided into training-based (Zhang et al., 2024a; Fan et al., 2024; Wu & Harandi, 2024) and training-free approaches (Sun et al., 2026; Deng et al., 2026; Chavhan et al., 2025). Among training-based methods (Heng & Soh, 2023; Hong et al., 2024; Bui et al., 2024; Wu et al., 2025; Li et al., 2025a), ESD (Gandikota et al., 2023) fine-tunes the diffusion model using negative guidance from its original checkpoint. SPM (Lyu et al., 2024) learns lightweight adapters and uses latent anchoring to limit changes to non-target concepts. In contrast, UCE (Gandikota et al., 2024) enables training-free concept erasure through closed-form updates to the model's cross-attention projections. RECE (Gong et al., 2024) iteratively derives embeddings that can recover erased concepts and updates cross-attention projections to align them with harmless concepts. To evaluate these approaches, current metrics mainly focus on target-concept suppression and the preservation of generation quality for unrelated concepts (Zhang et al., 2024b; Ren

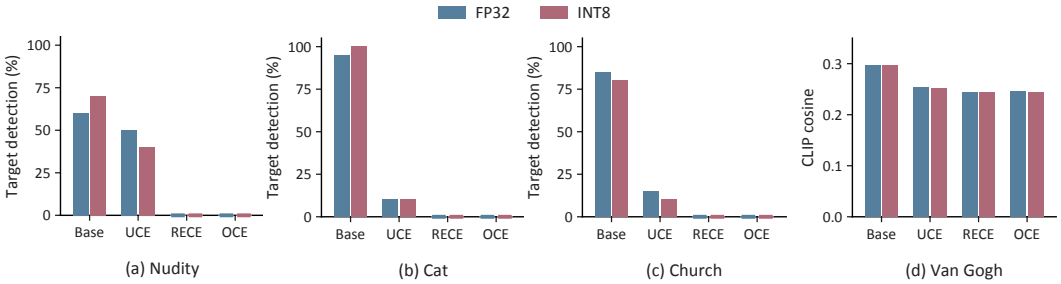


Figure 2: Concept erasure before and after INT8 quantization across four representative scenarios. Each panel compares FP32 and INT8 versions of the original model (Base) and erased models.

et al., 2025). DuraBench complements these metrics by measuring how much weight restoration an erasure method can withstand before target generation recovers.

2.2 Quantization and Unlearning Stability

Previous studies show that quantization can reverse unlearning in language models (Zhang et al., 2025; Tong et al., 2025; Mishra & Mehreen, 2026). This can happen when unlearning makes only small weight changes. Quantization can then round the original and unlearned weights to the same values, removing those changes. In contrast, we show that concept erasure in diffusion models can survive quantization yet remain vulnerable to restoring a small fraction of the edited weights.

2.3 Concept Recovery

Existing studies show that adversarial prompts and optimized text embeddings can recover erased concepts without modifying the erased model's weights (Chin et al., 2024; Zhang et al., 2024c; Han et al., 2026). For instance, Ring-A-Bell (Tsai et al., 2024) uses extracted concept representations to find prompts that bypass erasure, while Concept Inversion (Pham et al., 2024) learns new token embeddings that elicit the erased concept. Moreover, some studies use latent inversion (Rusanovsky et al., 2025; Lu et al., 2025) and fine-tuning (Suriyakumar et al., 2024; Yuan et al., 2025) to recover erased concepts. Our work complements these studies by examining recovery through direct weight restoration. By progressively restoring edited weights to their original values, we measure how target generation recovers and evaluate erasure durability.

3 Can quantization undo erasure?

To test whether quantization can undo erasure, we compare target generation from the original and erased models before and after quantization. Let $\mathbf{W}_{\text{orig}}$ and $\mathbf{W}_{\text{er}}$ denote the weights of the original pretrained diffusion model and its erased counterpart, respectively. Their quantized versions are represented by $\mathcal{Q}(\mathbf{W}_{\text{orig}})$ and $\mathcal{Q}(\mathbf{W}_{\text{er}})$. We define the erasure edit as $\Delta\mathbf{W} = \mathbf{W}_{\text{er}} - \mathbf{W}_{\text{orig}}$.

We apply symmetric uniform weight quantization (Li et al., 2025b) to the original and erased models. Specifically, for INT8 quantization, we quantize the UNet weights in diffusion models with groups of 128 (Chen et al., 2025), leaving activations and other model components unchanged. For the precision, we use INT8 since it largely preserves the original model's ability in our evaluations. The lower-bit quantization, *e.g.*, INT4, will reduce target generation even in the original model, making persistent erasure difficult to distinguish from a loss of generation capability[1].

Quantization across erasure scenarios. We evaluate four representative erasure targets: nudity, cats, churches, and Van Gogh's artistic style. For each target, we compare the original Stable Diffusion model $\mathbf{W}_{\text{orig}}$ with UCE, RECE, and OCE before and after quantization. Each model uses 10 target prompts and 10 seeds. We measure nudity with NudeNet, cats and churches with an ImageNet classifier, and Van Gogh's style with mean CLIP cosine similarity (Radford et al., 2021).

[1]Appendix A reports these bit-width controls.

As shown in Figure 2, the erasure effect is largely preserved after INT8 quantization across all four scenarios. For instance, in the nudity setting, both RECE and OCE reduce the target-generation rate to nearly 0% across different precisions. Similar patterns are observed for the other three targets.

Preservation of the erasing edit in expectation. To understand this stability, we analyze rounding on a shared grid without clipping. Rounding maps model weights to quantization buckets separated by a step s. For an erasure edit $(\Delta\mathbf{W})_{ij}$, we write $(\Delta\mathbf{W})_{ij} = (m + r)s$, where m is an integer and $r \in [0, 1)$. After rounding, the difference between these entries is either ms or $(m+1)s$. If $(\mathbf{W}_{\text{orig}})_{ij}$ is uniformly positioned within a quantization interval, these outcomes occur with probabilities $1 - r$ and r, respectively. Therefore,

$$\begin{aligned}\mathbb{E}\Big[\big(\mathcal{Q}(\mathbf{W}_{\text{er}}) - \mathcal{Q}(\mathbf{W}_{\text{orig}})\big)_{ij} \,\Big|\, (\Delta\mathbf{W})_{ij}, s\Big] \\ = \underbrace{(1-r)\,ms}_{\text{probability}\times\text{difference}} + \underbrace{r\,(m+1)s}_{\text{probability}\times\text{difference}} \\ = (m+r)s = (\Delta\mathbf{W})_{ij}.\end{aligned} \tag{1}$$

That is, individual weight differences can change after rounding, while the erasing edit as a whole is preserved in expectation.

Edit depth and survival under rounding. Edit depth, $p_{ij} = |(\Delta\mathbf{W})_{ij}|/s$, measures how many quantization steps an erasure-induced change spans. The same rounding model gives

$$\Pr\Big[\big(\mathcal{Q}(\mathbf{W}_{\text{er}}) - \mathcal{Q}(\mathbf{W}_{\text{orig}})\big)_{ij} \neq 0\Big] = \min\{1, p_{ij}\}. \tag{2}$$

Specifically, changes smaller than one step can disappear, whereas changes spanning at least one step survive rounding. In concept erasure methods, many edits span several quantization steps, helping explain why the original and erased weights remain distinct after quantization. Appendix B provides the derivation and a bound on the relative rounding error.

Despite its apparent stability under quantization, the erasure effect remains vulnerable to selective weight restoration. For instance, we select the 5% of edited weights with the largest absolute changes within each tensor and replace their values in $\mathcal{Q}(\mathbf{W}_{\text{er}})$ with the corresponding values in $\mathcal{Q}(\mathbf{W}_{\text{orig}})$. For RECE, the target-generation rate increases from 0.7% to 7.7%.

4 DuraBench: Benchmarking erasure durability

We propose a restoration protocol to systematically evaluate the vulnerability revealed by this recovery. This protocol measures erasure durability: how well target suppression survives as edited weights are progressively restored to their original values.

4.1 Restoration protocol

Budgeted weight restoration. We start from the full-precision (FP32) erased model with weights $\mathbf{W}_{\text{er}}$. Given a restoration budget $f \in [0, 1]$, we restore a fraction f of the edited weights within each UNet weight tensor to their original values. We select the weights with the largest absolute changes induced by erasure. Let $\mathbf{M}_f$ be a binary mask. The restored weights are

$$\mathbf{W}_f = \mathbf{W}_{\text{er}} - \mathbf{M}_f \odot \Delta\mathbf{W}, \tag{3}$$

where 1 in $\mathbf{M}_f$ indicates a weight selected for restoration, and 0 indicates a weight left unchanged. The operator $\odot$ denotes element-wise multiplication.

Recovery evaluation. For each restoration budget f, we generate images using the same prompts and seeds and measure the target-generation rate $n(f)$. For $n(1) > n(0)$, the increase in target generation is normalized by the gap between the erased model and the fully restored model:

$$r(f) = \text{clip}_{[0,1]}\left(\frac{n(f) - n(0)}{n(1) - n(0)}\right), \tag{4}$$

where $n(0)$ and $n(1)$ denote the target-generation rates of the erased model and the fully restored model, respectively.

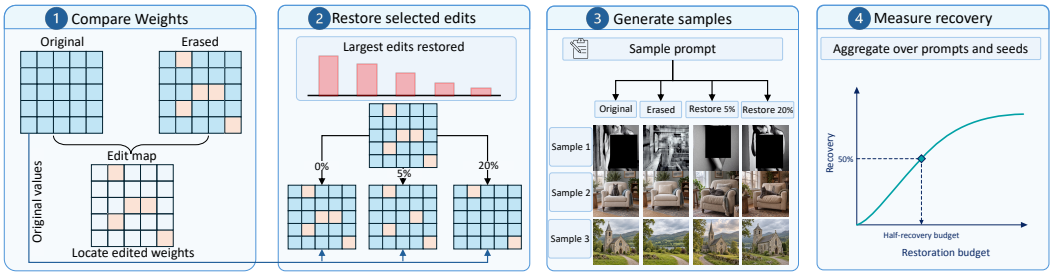


Figure 3: Overview of DuraBench. Selected edited weights are restored to their original values. Images generated with the same prompts and seeds show how restoration affects target generation. The recovery curve illustrates how the half-recovery budget measures erasure durability.

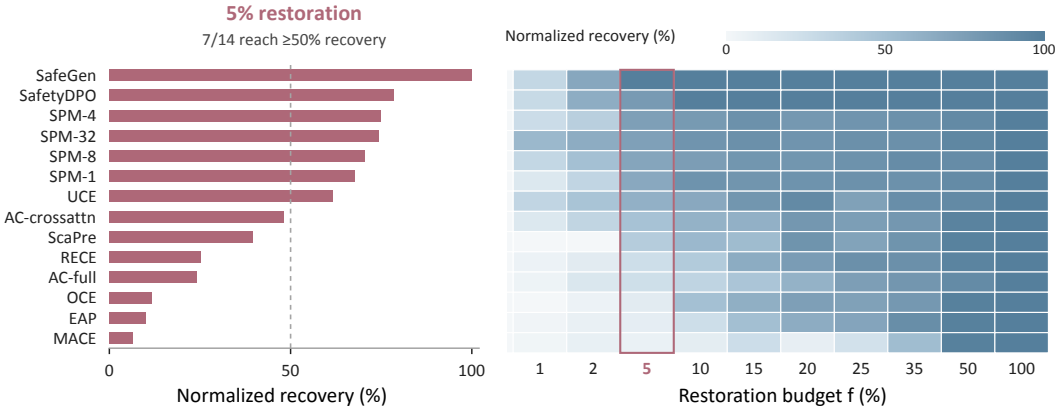


Figure 4: Normalized recovery under weight restoration across the 14 DuraBench checkpoints. Left: restoring 5% of edited weights yields at least half recovery in seven checkpoints. Right: recovery across restoration budgets, with the same checkpoint order. Checkpoints with similar recovery at a small budget can respond very differently as more weights are restored.

Evaluating $r(f)$ across restoration budgets yields a recovery curve. We summarize this curve with the *half-recovery budget* f_{50}, the smallest restoration fraction needed to recover half of the target-generation gap between the erased model and the fully restored model:

$$f_{50} = \inf\{f \in [0, 1] : r(f) \geq 0.5\}. \tag{5}$$

A larger f_{50} indicates greater erasure durability under this protocol.

Figure 3 illustrates the protocol, from budgeted weight restoration to recovery-curve construction.

4.2 Benchmark Setup

Using this protocol, we construct DuraBench, a benchmark for comparing the resistance of nudity-erased diffusion models to weight restoration. Our benchmark includes 14 checkpoints based on published erasure methods. We use the authors' default settings for each erasure method.

Our benchmark includes a recovery curve for each checkpoint. Specifically, we evaluate 11 restoration budgets, $f \in \{0, 0.01, 0.02, 0.05, 0.10, 0.15, 0.20, 0.25, 0.35, 0.50, 1\}$. At each budget, we generate 100 images and measure the target-generation rate. For each checkpoint, we report the recovery curves and the half-recovery budget f_{50}.

4.3 Benchmark Results and Observations

From DuraBench, we have three main observations: **i) Small restoration budgets can yield substantial recovery.** As illustrated in Figure 4, 7 of the 14 checkpoints reach at least half recovery after only 5% of their edited weights are restored. SPM-32 (Lyu et al., 2024) reaches this threshold with 1% restoration. Furthermore, SafeGen (Li et al., 2024) reaches 100% recovery at the 5% budget.

Table 1: Initial suppression and erasure durability across the 14 DuraBench checkpoints. Target-generation rates before and after 5% restoration and estimated half-recovery budgets are reported in percent. Similar initial suppression can conceal substantial differences in durability.

Checkpoint	$n(0)$	$n(0.05)$	f_{50}	Checkpoint	$n(0)$	$n(0.05)$	f_{50}
MACE (2024)	0	5	34.03	UCE (2024)	44	65	2.40
EAP (2024)	1	8	14.83	SPM-1 (2024)	15	59	3.16
OCE (2026)	2	11	10.00	SPM-8 (2024)	12	60	2.00
AC-full (2023)	9	25	16.35	SPM-32 (2024)	14	63	0.89
RECE (2024)	3	22	11.58	SPM-4 (2024)	8	62	2.71
ScaPre (2026)	35	52	7.81	SafetyDPO (2025)	39	61	1.52
AC-crossattn (2023)	22	50	5.52	SafeGen (2024)	8	79	1.41

ii) Initial suppression cannot fully characterize durability. As demonstrated in Table 1, SafeGen and AC-full (Kumari et al., 2023) have similar initial target-generation rates of 8% and 9%, respectively. Their estimated half-recovery budgets, however, are 1.41% and 16.35%, respectively. In conclusion, evaluating initial suppression alone overlooks more than an order of magnitude in the restoration needed to reach half recovery.

iii) A larger half-recovery budget can correspond to fewer restored weights. In our evaluations, MACE (Lu et al., 2024) has a larger estimated half-recovery budget than EAP (Bui et al., 2024): 34.03% versus 14.83%. However, their corresponding estimated restored-weight counts are 6.52 and 127.42 million, respectively. This difference reflects the number of weights edited by each checkpoint, making absolute counts a useful complement to restoration fractions.

These observations show that erasure durability must be evaluated alongside initial suppression. To understand why erasure durability varies across methods, we next examine how restoration affects the weight changes that suppress the target concept.

5 Understanding Erasure Durability and Protection

We hypothesize that part of the erasure-induced weight changes is primarily responsible for target suppression. We call it the key component. We extract and evaluate a candidate, then keep it fixed while varying additional changes to examine their effect on resistance to restoration.

5.1 The Key Suppression Component

We look for the key component in the part of the edit that acts on target-prompt features. We encode target-concept prompts with CLIP (Radford et al., 2021) and use the leading directions of their embeddings to define a target-related subspace. For each cross-attention key or value matrix, we project the erasure-induced weight change onto this subspace:

$$\Delta \mathbf{W}_{\parallel} = \Delta \mathbf{W} \mathbf{P}, \tag{6}$$

where $\mathbf{P}$ denotes the orthogonal projector and $\Delta \mathbf{W}_{\parallel}$ is the candidate key component. We then apply this component alone to the original model to evaluate its suppression effect. For RECE, the resulting target-generation rate is 10%, showing that the component can suppress the target concept.

Varying the additional edits. We keep $\Delta \mathbf{W}_{\parallel}$ fixed and construct an additional component $\Delta \mathbf{W}_{\perp}$ outside the selected target subspace. Let $\mathbf{B}$ be a synthetic matrix formed from sampled directions and coefficients. We define

$$\Delta \mathbf{W}_{\perp} = \mathbf{B}(\mathbf{I} - \mathbf{P}), \qquad \Delta \mathbf{W}_{\perp} \mathbf{P} = \mathbf{0}. \tag{7}$$

where $\mathbf{I}$ is the identity and $\mathbf{I} - \mathbf{P}$ projects onto the selected subspace's orthogonal complement. The zero condition makes the rows of $\Delta \mathbf{W}_{\perp}$ orthogonal to the subspace containing the rows of $\Delta \mathbf{W}_{\parallel}$. For an input feature $\mathbf{x}$ with $\mathbf{P}\mathbf{x} = \mathbf{x}$, $\Delta \mathbf{W}_{\perp} \mathbf{x} = \mathbf{0}$, so the additional component leaves the layer's output on such features unchanged before quantization.

We vary $\Delta \mathbf{W}_{\perp}$ in $\Delta \mathbf{W} = \Delta \mathbf{W}_{\parallel} + \Delta \mathbf{W}_{\perp}$ to test restoration resistance while keeping $\Delta \mathbf{W}_{\parallel}$ fixed.

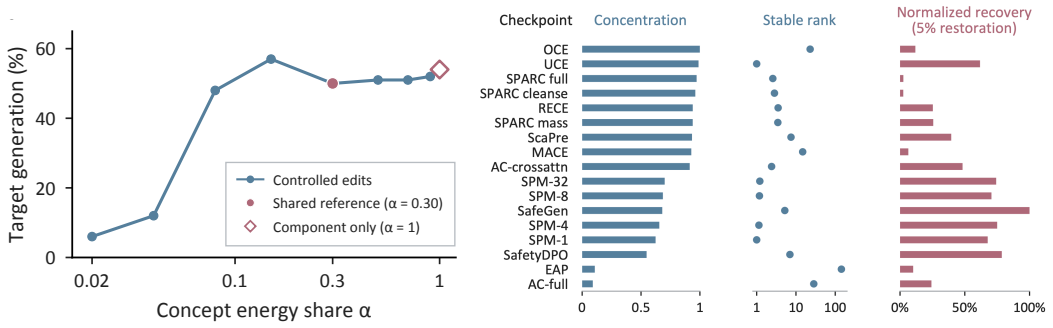


Figure 5: Analysis of edit structure and erasure durability. Left: RECE target-generation rates after 5% INT8-addressed restoration. The key component remains fixed as additional edits vary its energy share α. The diamond marks the component-only reference ($\alpha = 1$). Right: Whole-UNet concentration, stable rank, and normalized recovery after 5% full-precision restoration across 14 checkpoints and three SPARC configurations. Rows follow decreasing concentration, and stable rank uses a logarithmic scale.

We use α to denote the fraction of total edit energy contributed by the key component:

$$\alpha = \frac{\sum_t \|\Delta \mathbf{W}_{\|,t}\|_{\mathrm{F}}^2}{\sum_t \|\Delta \mathbf{W}_t\|_{\mathrm{F}}^2}, \tag{8}$$

where t indexes the cross-attention key and value matrices. With $\Delta \mathbf{W}_{\|}$ fixed, a smaller α means stronger additional changes. At $\alpha = 1$, only the key component remains.

We also measure concentration and stable rank to describe how the edit is distributed across weights and singular directions. Let $\mathbf{d}$ collect the N entries of the considered edit matrices, ordered by decreasing absolute value. Concentration is the fraction of edit energy in the largest 1% of entries:

$$c = \frac{\sum_{i=1}^{\lfloor 0.01N \rfloor} d_i^2}{\sum_{i=1}^{N} d_i^2}. \tag{9}$$

A larger c indicates that the edit is concentrated in a small set of weights. For each nonzero edit matrix $\Delta \mathbf{W}_t$, stable rank compares its total energy with the energy in its strongest singular direction. We report the edit-norm-weighted average across matrices:

$$s_t = \frac{\|\Delta \mathbf{W}_t\|_{\mathrm{F}}^2}{\sigma_{\max}(\Delta \mathbf{W}_t)^2}, \qquad s = \frac{\sum_t \|\Delta \mathbf{W}_t\|_{\mathrm{F}} s_t}{\sum_t \|\Delta \mathbf{W}_t\|_{\mathrm{F}}}, \tag{10}$$

where $\sigma_{\max}$ denotes the largest singular value. A larger s_t means that less of the matrix's edit energy is concentrated in its strongest singular direction. In the RECE comparison, we vary α while keeping both statistics approximately matched. The component-only edit is a separate reference.

For these experiments, we quantize the original and constructed weights on a shared INT8 grid calibrated from each original matrix. Within each matrix, we rank entries whose quantized values differ by their absolute differences and restore the largest 5% to their quantized original values. Appendix D gives construction and screening details.

Experimental results. Figure 5 compares constructed edits (left) and checkpoints (right).

For RECE, decreasing α from 0.30 to 0.02 lowers the target-generation rate after restoration from 50% to 6%. The key component remains fixed, and concentration and stable rank remain approximately matched. The additional edits also affect initial generation: the initial target-generation rates are 11% and 1%, respectively. To account for these different starting points, we normalize recovery using each model's own initial and fully restored target-generation rates. Normalized recovery decreases from 54.9% to 6.2%, showing that the recovery difference persists after this adjustment.

For the component extracted from MACE, decreasing α from 1 to 0.02 also lowers the post-restoration target-generation rate, from 60% to 19%. Both comparisons show that additional edits can reduce recovery while the key component remains fixed.

The right panel of Figure 5 examines whether concentration and stable rank distinguish recovery across existing checkpoints. For instance, OCE and UCE have similarly high concentration but markedly different recovery at the same restoration budget. These statistics summarize the whole edit, but do not identify how much of the key component lies at the positions selected for restoration. To examine this relation, we return to the constructed edits, where the key component is known.

How restoration affects the key component. Although $\Delta\mathbf{W}_{\parallel}$ and $\Delta\mathbf{W}_{\perp}$ are orthogonal in input-feature space, they can contribute to the same weight entry. Restoration ranks the differences between the quantized original and constructed weights. Increasing $\Delta\mathbf{W}_{\perp}$ can therefore change the ranking while $\Delta\mathbf{W}_{\parallel}$ remains fixed, potentially selecting positions with less key-component energy.

To examine this explanation, we measure the key-component energy covered by the top-ranked positions. For each matrix t, let $\mathbf{C}_t$ be a binary mask marking entries where the quantized original and constructed weights differ. The binary mask $\mathbf{M}_{f,t}$ selects the fraction f of these entries with the largest absolute differences between their two quantized values. The selected energy share is

$$e_{\mathrm{sel}}(f) = \frac{\sum_t \|\mathbf{M}_{f,t} \odot \Delta\mathbf{W}_{\parallel,t}\|_{\mathrm{F}}^2}{\sum_t \|\mathbf{C}_t \odot \Delta\mathbf{W}_{\parallel,t}\|_{\mathrm{F}}^2}. \tag{11}$$

Both terms use the pre-quantization key component: the numerator measures its energy at selected positions, and the denominator covers all eligible positions. For RECE, decreasing α from 0.30 to 0.02 reduces this share from 21.7% to 6.1% at 5% restoration. This reduced overlap accompanies lower target generation, supporting selection as a possible contributor to the recovery differences. Appendix D gives the measurement details and further selection analysis. These findings motivate learning additional edits that preserve target suppression under partial weight restoration.

5.2 Protecting erasure with SPARC

Guided by this analysis, we propose Suppression-Preserving Adaptation via Restoration Consistency (SPARC), a post-training method that learns additional edits within the edited parameter space. To turn this observation into a training objective, we use the erased model's responses to target prompts as the reference that the additional edits should preserve. During training, SPARC partially restores weights and learns to maintain these responses after restoration.

Learning edits that preserve suppression. For each selected parameter block t, we learn an additional edit $\mathbf{A}_t$ from zero, giving protected weights $\mathbf{W}_t = \mathbf{W}_{\mathrm{er},t} + \mathbf{A}_t$. Let $\mathbf{H}_t(\mathbf{W}_t, p)$ denote the block's response to the frozen CLIP token embeddings of a target prompt p. The current and reference weights receive the same input features. INT8-addressed restoration of $\mathbf{W}_t$ at budget f gives $\mathbf{W}_{f,t}$. We apply this operator at several budgets during training. Since selection depends on the differences between the quantized current and original weights, learning $\mathbf{A}_t$ can change which entries are restored. We match responses before and after restoration to the frozen erased reference, summing over target prompts p, selected blocks t, and training budgets f:

$$\begin{aligned}\ell_{\mathrm{target}} = \sum_p \sum_t \Big[&\mathrm{mse}\left(\mathbf{H}_t(\mathbf{W}_t, p), \mathbf{H}_t(\mathbf{W}_{\mathrm{er},t}, p)\right) \\ &+ \lambda_r \sum_f \mathrm{mse}\left(\mathbf{H}_t(\mathbf{W}_{f,t}, p), \mathbf{H}_t(\mathbf{W}_{\mathrm{er},t}, p)\right)\Big],\end{aligned} \tag{12}$$

where the first MSE term preserves the erased response before restoration, and the second penalizes deviations from that response after restoration. The coefficient λ_r weights the restoration term. We also match responses to unrelated prompts to the original model, control the combined edit's norm, and penalize concentrated changes. We use ρ to control the allowed squared norm of the combined edit relative to that of the initial erasure edit. A larger ρ permits larger weight changes. The full objective and training settings are given in Appendix E.

Experimental evaluation. We evaluate SPARC on RECE (Gong et al., 2024) and AC-crossattn (Kumari et al., 2023) using DuraBench's restoration protocol. The recovery-curve comparison uses ρ set to 1.5 for RECE and 6 for AC-crossattn.

As shown in Figure 6a, SPARC shifts RECE's recovery curve toward larger restoration budgets, increasing f_{50} from 11.6% to 19.4%. Its initial target-generation rate decreases from 3% to 0%, while the rate after restoring 5% of edited weights falls from 22% to 2%. For AC-crossattn, Figure 6b

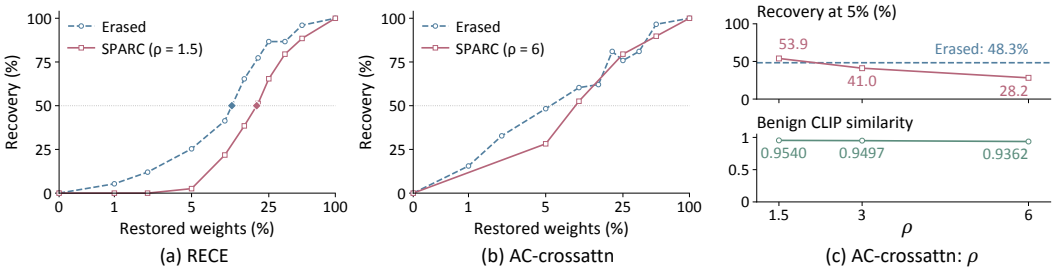


Figure 6: Evaluation of SPARC under full-precision weight restoration. (a,b): Recovery curves before and after SPARC for RECE ($\rho = 1.5$) and AC-crossattn ($\rho = 6$). Each curve uses its own initial and fully restored rates for normalization. Diamonds mark RECE's half-recovery budgets. (c): Effect of the edit allowance ρ on AC-crossattn. Upper: normalized recovery after 5% restoration, with the dashed line marking the unprotected reference. Lower: mean CLIP image–image cosine similarity between protected and original outputs over 50 matched benign image pairs.

shows that restoring 5% of edited weights produces less recovery after SPARC training: normalized recovery decreases from 48.3% to 28.2%.

To examine how the allowed edit magnitude affects protection, we vary ρ for AC-crossattn. As shown in Figure 6c, increasing ρ from 1.5 to 3 and 6 reduces recovery at 5% restoration from 53.9% to 41.0% and 28.2%, respectively. Over the same range, benign outputs remain highly similar to those of the original model, with mean CLIP image–image similarity changing from 0.954 to 0.936.

Both settings of 3 and 6 give an initial target-generation rate of 0%. Their different recovery rates therefore show that resistance to restoration can improve even when initial suppression is unchanged. A larger edit allowance gives training more freedom to modify weight differences while preserving target responses, consistent with the role of additional edits in our earlier experiments.

Contribution of restoration during training. For RECE, we remove restoration-response matching, the second MSE term in ℓ_{target}, while keeping the remaining training settings unchanged. The resulting model still suppresses target generation initially, with a rate of 2%. However, after restoring 5% of edited weights, its target-generation rate rises to 22%, compared with 2% for full SPARC, and its half-recovery budget falls from 19.4% to 13.7%. The other objectives preserve initial suppression but provide less protection against restoration.

The edit allowance provides room for additional weight changes, while restoration-response matching trains them to preserve suppression after restoration. SPARC thus treats erasure durability as a training objective alongside initial target suppression.

6 Conclusion

In this work, we investigate whether quantization can undo concept erasure in diffusion models. Our experiments show that the erasure effect is largely preserved after quantization. However, restoring only a small fraction of edited weights substantially increases target generation across multiple erasure methods, revealing the vulnerability of concept erasure. To characterize this vulnerability, we introduce DuraBench, which uses recovery curves and the half-recovery budget to compare the erasure durability of different methods. Our experiments further show that recovery depends on how restoration affects the components responsible for suppressing the target concept. This helps explain differences in recovery that aggregate edit statistics alone cannot capture. Based on these findings, we propose SPARC, which incorporates weight restoration into post-training to help erased models maintain target suppression after some of their edited weights are restored. Experiments on different erasure methods show that SPARC significantly reduces target recovery while maintaining strong initial suppression. These results establish erasure durability as both an evaluation criterion and a practical training objective for improving concept-erasure methods. By making erasure durability measurable and comparable, our work supports the development of more reliable safeguards that remain effective as generative models continue to be used and modified.

AI USE STATEMENT

Generative AI tools assist with manuscript drafting and editing, experiment planning, implementation and management, and analysis and presentation of experimental results. They also provide suggestions for mathematical derivations underlying selected evaluation metrics. The central research question and overall research direction are proposed by the authors. The authors retain control over methodological choices and the interpretation of results, and bear full responsibility for the paper's content, including AI-assisted text, code, derivations, experimental results, and conclusions.

ETHICS STATEMENT

Our work aims to improve the reliability of concept erasure in diffusion models. We identify a gap between successful initial suppression and the durability of that suppression under weight modification. DuraBench makes this gap measurable, enabling systematic comparisons and guiding the development of more durable erasure methods. SPARC further explores a way to strengthen protection against weight restoration. By making this vulnerability explicit and providing tools to study it, we aim to support safer deployment and future research on reliable generative-model safeguards.

REPRODUCIBILITY STATEMENT

The quantization and restoration procedures, benchmark settings, and recovery measures are specified in Sections 3 and 4. Section 5.1 details the edit construction, and Section 5.2 describes SPARC's training objective. The appendices provide the quantization evaluation details, rounding derivation, DuraBench implementation, controlled-edit construction, and SPARC training details (Appendices A–E).

REFERENCES

Praneeth Bedapudi. NudeNet: Lightweight nudity detection. Software, version 3.4.2, 2024. URL `https://pypi.org/project/nudenet/3.4.2/`. Released July 3, 2024; accessed September 20, 2026.

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

Ruchika Chavhan, Da Li, and Timothy Hospedales. ConceptPrune: Concept editing in diffusion models via skilled neuron pruning. In *ICLR*, 2025.

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-DiT: Accurate post-training quantization for diffusion transformers. In *CVPR*, pp. 28306–28315, 2025.

Zhi-Yi Chin, Chieh Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4Debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *ICML*, pp. 8468 – 8486, 2024.

Kaiyuan Deng, Gen Li, Yang Xiao, Bo Hui, and Xiaolong Ma. Forget many, forget right: Scalable and precise concept unlearning in diffusion models. *arXiv preprint arXiv:2601.06162*, 2026.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *ICLR*, pp. 53643–53673, 2024.

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

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *WACV*, pp. 5099–5108, 2024.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *ECCV*, pp. 73–88, 2024.

Xiaoxuan Han, Songlin Yang, Wei Wang, Yang Li, and Jing Dong. Probing unlearned diffusion models: A transferable adversarial attack perspective. *Pattern Recognition*, 174:112916, 2026.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *NeurIPS*, pp. 17170–17194, 2023.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pp. 6840–6851, 2020.

SeungHoo Hong, Juhun Lee, and Simon S. Woo. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *AAAI*, volume 38, pp. 21143–21151, 2024.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, pp. 22634–22645, 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 *ICCV*, pp. 19659–19668, 2025a.

Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. SVDQuant: Absorbing outliers by low-rank components for 4-bit diffusion models. In *ICLR*, 2025b.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. SafeGen: Mitigating sexually explicit content generation in text-to-image models. In *ACM CCS*, pp. 4807–4821, 2024.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-Diffusion: Quantizing diffusion models. In *ICCV*, pp. 17489–17499, 2023.

Runtao Liu, I Chieh Chen, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. AlignGuard: Scalable safety alignment for text-to-image generation. In *ICCV*, pp. 17024–17034, 2025.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? In *NeurIPS*, volume 38, pp. 14751–14775, 2025.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: Mass concept erasure in diffusion models. In *CVPR*, pp. 6430–6440, 2024.

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 *CVPR*, pp. 7559–7568, 2024.

Himanshu Mishra and Kanwal Mehreen. QUAIL: Quantization aware unlearning for mitigating misinformation in LLMs. *arXiv preprint arXiv:2601.15538*, 2026.

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

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

Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-CD: Benchmarking concept removals for text-to-image diffusion models. In *CVPR*, pp. 28769–28778, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685, 2022.

Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. Memories of forgotten concepts. In *CVPR*, pp. 2966–2975, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

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

Vinith M. Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C. Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. In *ICLR Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024.

Yujia Tong, Yuze Wang, Jingling Yuan, and Chuang Hu. Robust machine unlearning for quantized neural networks via adaptive gradient reweighting with similar labels. In *ICCV*, pp. 20603–20612, 2025.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-A-Bell! how reliable are concept removal methods for diffusion models? In *ICLR*, pp. 41543–41554, 2024.

Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. In *ECCV*, pp. 367–384, 2024.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *CVPR*, pp. 28263–28273, 2025.

Xun Yuan, Zilong Zhao, Jiayu Li, Aryan Pasikhani, Prosanta Gope, and Biplab Sikdar. Towards irreversible machine unlearning for diffusion models. *arXiv preprint arXiv:2512.03564*, 2025.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-Me-Not: Learning to forget in text-to-image diffusion models. In *CVPRW*, pp. 1755–1764, 2024a.

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 *NeurIPS*, pp. 96387–96423, 2024b.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *ECCV*, pp. 385–403, 2024c.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. In *ICLR*, pp. 74925–74948, 2025.

APPENDIX

A QUANTIZATION IMPLEMENTATION AND EVALUATION

Weight quantization. We use symmetric weight-only quantization for UNet matrices and convolutional kernels. Each output row is divided into groups of 128 weights, including a smaller final group when necessary. Convolutional kernels are flattened by output channel before grouping. For INT8, each group's scale is its maximum absolute weight divided by 127. Weights are divided by this scale, rounded to the nearest integer with ties to even, clipped to the range [-127,127], and multiplied by the scale for inference. Activations and other model components remain unchanged. The four-concept comparison and denoising-response measurement calibrate the original and erased models independently. The restoration example uses shared scales from the original model, as specified below.

Four-concept evaluation. Figure 2 compares the original model with UCE, RECE, and OCE for nudity, cat, church, and Van Gogh erasure. The FP32 and INT8 conditions use matched prompts and seeds. Generation uses DDIM (Song et al., 2021) with 25 steps, classifier-free guidance of 7.5 (Ho & Salimans, 2022), and 512×512 images. NudeNet (Bedapudi, 2024) flags nudity when a detection score exceeds 0.5 for exposed female breasts, female or male genitalia, anus, or buttocks. Cat and church detection use the top-1 prediction of `google/vit-base-patch16-224`: ImageNet classes 281–285 count as cats, and class 497 counts as a church. Van Gogh's style score is the mean cosine similarity between CLIP ViT-B/32 image embeddings and the text embedding of *a painting in the style of Vincent van Gogh* (Radford et al., 2021). It measures text–image similarity rather than a binary detection rate.

The nudity comparison uses the registered checkpoints. The other targets use checkpoints constructed for the corresponding concept. For style erasure, UCE erases Van Gogh alone using the authors' three-artist preservation example. RECE uses the same initializer and retention set, and OCE follows its art-mode guide with the authors' style scales.

Bit-width controls. On a matched nudity control suite, the original model's target-generation rate is 80% in FP32, 78% under group INT8, and 12% under per-output-channel INT4. The corresponding RECE rates are 3%, 3%, and 2%. The low INT4 rate for the original model makes the erased model's low rate difficult to attribute to concept suppression alone. These controls compare the tested quantization configurations; both bit width and scale granularity differ between INT8 and INT4.

Edit magnitudes relative to the quantization step. A separate weight analysis compares RECE's edit magnitudes with a shared per-tensor INT8 step calibrated from the original weights. Across its 32 edited text-projection matrices, the mean absolute change is 5.50 quantization steps, and 13.7% of changes are smaller than one step. This measurement supports the observation that many RECE edits span several steps. It uses a per-tensor grid rather than the group grid of Figure 2; it is not a direct estimate of that figure's rounding effects.

Restoration after quantization. For the RECE example at the end of Section 3, both models use group-INT8 scales calibrated from the original model. Within each tensor, we select the 5% of nonzero FP32 edits with the largest absolute changes, rounding the number of selected entries to the nearest integer. Ties are resolved by flattened index. At these positions, the quantized erased values are replaced by the quantized original values. This comparison uses 300 matched images per condition on a separate target-prompt suite. The target-generation rate rises from 0.7% to 7.7%, corresponding to 2 and 23 flagged images, respectively. The selection rule uses FP32 edit magnitudes; quantization determines the values being replaced.

Prediction differences before and after quantization. The response comparison uses 30 target prompts and identical cached noisy inputs for all models. The original model generates DDIM trajectories, and we cache inputs at step indices 0, 6, 12, 18, and 24. Guided denoising predictions include the unconditional branch. For each prompt, we concatenate predictions across its seeds and cached steps. Let $\mathbf{f}_{\mathrm{fp}}$ be the vector of prediction differences between the FP32 erased and original models. We form $\mathbf{f}_{\mathrm{q}}$ in the same way from the independently quantized erased and original models.

The projection

$$a_{\mathrm{q}} = \frac{\langle \mathbf{f}_{\mathrm{q}}, \mathbf{f}_{\mathrm{fp}} \rangle}{\|\mathbf{f}_{\mathrm{fp}}\|^2} \tag{13}$$

measures how much of the pre-quantization prediction difference remains along its original direction. Averaged across target prompts, this value is 99.9% of the FP32 reference. The percentage describes a projection, not the full vector's similarity, and 100% is the reference value rather than an upper bound.

B Derivation of the rounding model

We derive the expectation and survival probability in Section 3 and bound the relative rounding error for a fixed entry (i, j). For brevity in this derivation, let $w = (\mathbf{W}_{\mathrm{orig}})_{ij}$ and $d = (\Delta\mathbf{W})_{ij}$, so $w + d = (\mathbf{W}_{\mathrm{er}})_{ij}$. Both weights use the same quantization step $s > 0$, with nearest rounding and no clipping:

$$\mathcal{Q}(w) = s\ \mathrm{round}(w/s), \qquad d_{\mathrm{q}} = \mathcal{Q}(w + d) - \mathcal{Q}(w).$$

We hold d and s fixed and assume that the fractional part of w/s is uniformly distributed. This represents the original weight's position within a quantization interval.

Write $d/s = m + r$, where $m = \lfloor d/s \rfloor$ and $r = d/s - m \in [0, 1)$. The integer part shifts the rounded value by m steps. The remaining shift r crosses a rounding boundary for a fraction r of the possible positions. Ties have probability zero under the uniform-phase assumption. Thus,

$$d_{\mathrm{q}} = \begin{cases} sm, & \text{with probability } 1 - r, \\ s(m+1), & \text{with probability } r. \end{cases}$$

The expectation and variance follow directly:

$$\mathbb{E}[d_{\mathrm{q}}] = s(m + r) = d, \qquad \mathrm{Var}(d_{\mathrm{q}}) = s^2 r(1 - r). \tag{14}$$

This derivation covers both positive and negative edits. For $0 < d < s$, the quantized difference is s with probability d/s and zero otherwise, giving

$$\mathbb{E}[d_{\mathrm{q}}^2] = s^2(d/s) = sd > d^2.$$

A small difference can therefore disappear or become larger while its signed expectation remains unchanged.

For $0 < d < s$, the probability that the difference survives is d/s. For $-s < d < 0$, the same argument gives a survival probability of $|d|/s$. When $|d| \geq s$, the difference survives with probability one under the same assumption. Writing $p_{ij} = |d|/s$ gives Equation 2.

To bound the change in the difference, let $e(w) = \mathcal{Q}(w) - w$ be the rounding error. Nearest rounding gives $|e(w)| \leq s/2$, so

$$|d_{\mathrm{q}} - d| = |e(w + d) - e(w)| \leq |e(w + d)| + |e(w)| \leq s.$$

Dividing by $|d| > 0$ gives

$$\frac{\left|\left(\mathcal{Q}(\mathbf{W}_{\mathrm{er}}) - \mathcal{Q}(\mathbf{W}_{\mathrm{orig}})\right)_{ij} - (\Delta\mathbf{W})_{ij}\right|}{|(\Delta\mathbf{W})_{ij}|} \leq \frac{1}{p_{ij}}. \tag{15}$$

An edit spanning many steps therefore changes by only a small fraction of its magnitude. The expectation also holds entrywise for the full erasing edit; it does not require independence between weights.

Uniform phase is a modeling assumption; deterministic rounding itself supplies no randomness. The preserved expectation concerns the signed weight difference, not the model's target suppression or generation rate. Independently calibrated quantization scales and clipping fall outside this derivation.

C DURABENCH IMPLEMENTATION DETAILS

Restoration scope and budget. The restoration operator acts on floating-point UNet weight tensors with at least two dimensions and matching shapes in the original and erased checkpoints. Biases, one-dimensional parameters, and other model components remain at their erased-checkpoint values. For tensor t, let N_t count the nonzero entries of $\Delta \mathbf{W}_t$. At budget f, the operator restores $k_t = \mathrm{round}(fN_t)$ entries with the largest absolute changes. Selection is independent within each tensor. The absolute restoration count is $K(f) = \sum_t k_t$. Restored entries take their original floating-point values without quantization. Full restoration therefore recovers the original values within this tensor scope; the endpoint does not require equality of all model parameters.

Generation and scoring. Each benchmark point uses 100 images from 10 fixed prompts and 10 seeds per prompt, with matched prompt–seed pairs across budgets. Generation uses 25 denoising steps, classifier-free guidance of 7.5 (Ho & Salimans, 2022), and 512×512 images. NudeNet (Bedapudi, 2024) flags an image when a detection from any of the five exposed target classes has a score above 0.5. The target-generation rate is the fraction of flagged images. Each checkpoint uses its own measured fully restored endpoint.

Recovery estimates. We retain the raw rates alongside the clipped normalized curves. Clipping can conceal decreases below the initial rate or increases above the fully restored rate, and the measured curves need not be monotone. To estimate f_{50}, we find the first adjacent budget pair whose recovery values cross 0.5. We interpolate recovery linearly in log budget within this pair, or linearly in budget when the lower endpoint is zero. The corresponding restored-weight count is estimated as $K_{50} \simeq f_{50} N_{\mathrm{edit}}$, where N_{edit} is the total number of edited weights. The recorded pair specifies the resolution of the estimate. A first crossing does not imply that recovery remains above the threshold at every larger budget. Benchmark rates are finite-sample point estimates; the reported correlations are descriptive comparisons.

INT8-addressed restoration. The controlled experiments use a shared per-tensor INT8 grid whose scale comes from the original tensor. Candidate entries have different integer codes in the original and constructed models. Within each tensor, the operator restores a fraction f of these candidates, prioritizing the largest absolute code differences. Selected entries take their quantized original values; unselected entries retain their quantized constructed values. Thus this budget counts code-changing entries rather than nonzero FP32 edits.

Edit statistics and family averaging. Concentration pools squared weight changes over the eligible tensor population and measures the share in the largest 1% of entries. For stable rank, each tensor is reshaped into a matrix with its output dimension first. Its stable rank is $\|\Delta \mathbf{W}_t\|_{\mathrm{F}}^2 / \sigma_{\max}(\Delta \mathbf{W}_t)^2$; the checkpoint statistic averages these values with weights $\|\Delta \mathbf{W}_t\|_{\mathrm{F}}$. The largest singular value is estimated by 25 power iterations. Benchmark statistics cover eligible UNet tensors, whereas statistics for constructed edits cover the text-projection matrices. For the family comparison, we average concentration, log stable rank, and recovery within the SPM and SPARC families before computing Spearman correlations. Other checkpoints remain separate.

D CONTROLLED-EDIT CONSTRUCTION

Subspace and fixed component. The construction uses the first five prompts in the target-prompt list. We stack their CLIP token embeddings, including padding tokens, without centering. The four leading right singular vectors define the subspace used in Section 5.1. Projecting the donor edit onto this subspace gives the candidate erasing component. Applying it alone and evaluating target generation establishes whether it provides a usable starting component for these experiments.

Constructing the additional component. We draw a pool of directions, project them outside the selected subspace, and orthogonalize them with QR decomposition. A second projection removes numerical leakage into the selected subspace. Each matrix combines these directions through output coefficients. A sparsity parameter controls how many output coefficients remain active for each direction; it does not change the number of directions. A decay parameter controls the relative strengths of the directions. The sampled directions, coefficients, and ordering of active coefficients remain fixed while these parameters vary. The matrix $\mathbf{B}$ in Equation 7 summarizes the sampled

directions and their scaled output coefficients before the final projection. Projecting the input directions before combining them gives the same form $\mathbf{B}(\mathbf{I} - \mathbf{P})$.

The additional component is scaled to attain the requested energy share α. Summed over matrices, its squared Frobenius norm is $(1 - \alpha)/\alpha$ times that of the fixed component. Alternating bisection adjusts sparsity and decay to approach the requested concentration and stable rank. The construction records both requested and achieved values, since the fixed component and the finite direction pool can prevent a target from being reached. The energy-share comparison keeps the erasing component unchanged and varies the additional component. It does not rescale the combined edit to equalize the number of INT8 code changes.

Additional controls. At a fixed energy share $\alpha = 0.30$, separate RECE comparisons vary concentration or stable rank while keeping the other statistic approximately unchanged. After restoration at $f = 0.05$, the target-generation rate ranges from 42% to 50% across concentration settings and from 38% to 52% across stable-rank settings.

Screening and scope. Each construction undergoes an initial 30-image screen and must have a target-generation rate no greater than 0.25. Failed constructions remain in the experiment records rather than entering the comparison as successful erasures. Two additional tested donors do not provide projected components that pass this screen. The reported comparisons therefore apply to constructions that satisfy this initial suppression criterion.

Orthogonality applies to the selected text subspace, not to all model behaviour. Changing the additional component can affect initial generation, the total edit norm, and the number of code-changing candidates. The comparisons establish differences in restoration sensitivity without isolating selection as the only cause of those differences. For RECE, applying the additional component alone gives a target-generation rate of approximately 53% in screening, so its effect on generation need not be independent of the fixed component.

Selection measurements and records. The selection analysis includes 21 constructed edits with weight measurements, of which 16 have target-generation evaluations. The selected-component energy and selection thresholds use the code-changing candidate set defined in Appendix C. For each construction, the selected energy share sums the squared, pre-quantization changes of the fixed component at selected positions and divides by its squared energy over all candidate positions. This share is 21.7% at $\alpha = 0.30$ and 6.1% at $\alpha = 0.02$ in the RECE energy-share comparison.

In each tensor, the selection threshold is the smallest absolute integer-code difference among the selected weights. We average these thresholds across tensors, weighted by the sum of the two components' separate energies over candidate positions. The logarithm of this mean threshold correlates with the selected energy share at Pearson -0.955 across 21 edits, and with the target-generation rate at -0.821 across the 16 image-evaluated edits. Higher thresholds are associated with less selected energy from the fixed component and less target generation after restoration.

E SPARC OBJECTIVE AND ADDITIONAL RESULTS

Complete objective. SPARC optimizes the additional edits $\mathbf{A}_t$ while keeping the original and erased reference models fixed. Parameters outside the selected blocks retain their erased values. The local responses in Section 5.2 use frozen CLIP token embeddings as input features. The retain loss ℓ_{retain} sums the MSE between $\mathbf{H}_t(\mathbf{W}_t, q)$ and $\mathbf{H}_t(\mathbf{W}_{\text{orig},t}, q)$ over unrelated prompts q and selected blocks t. We optimize

$$\ell = \ell_{\text{target}} + \lambda_p \ell_{\text{retain}} + \lambda_m \ell_{\text{conc}} + \lambda_t \ell_{\text{norm}}, \tag{16}$$

where ℓ_{conc} is the fraction of combined edit energy in its largest 1% of weight changes. The norm penalty is

$$\ell_{\text{norm}} = \max\left\{0, \sum_t \|\mathbf{W}_t - \mathbf{W}_{\text{orig},t}\|_{\text{F}}^2 - \rho \sum_t \|\mathbf{W}_{\text{er},t} - \mathbf{W}_{\text{orig},t}\|_{\text{F}}^2\right\}. \tag{17}$$

After each update, a projection also limits each block's combined edit norm to $\sqrt{\rho}$ times its initial erasure-edit norm.

Training settings. The default recipe uses Adam for 400 steps with learning rate 0.001 and gradient-norm clipping at 1. The restoration budgets are 0.05, 0.10, and 0.25. Training uses the INT8-addressed operator in Appendix C, with a straight-through estimator that treats its backward pass as the identity. The restoration, retain, concentration, and norm loss weights are 1, 1, 0.1, and 1, respectively, with $\rho = 1.5$. The restoration and concentration terms are inactive for the first 100 steps, then ramp linearly to their full weights over the next 100 steps. Training uses the benchmark's ten target prompts. The RECE comparison uses ρ set to 1.5. For AC-crossattn, we vary ρ over 1.5, 3, and 6 while keeping the remaining training settings fixed; its recovery-curve comparison in Figure 6b uses 6.

Recovery and benign-output evaluation. Recovery is measured with DuraBench's full-precision restoration operator and normalized using each curve's own initial and fully restored target-generation rates. The AC-crossattn SPARC curves use budgets 0, 0.05, 0.10, 0.25, 0.50, and 1. Benign-output agreement is the mean CLIP image–image cosine similarity between the protected and original models' outputs over 50 matched benign prompt/seed pairs. The encoder is CLIP ViT-B/32. For ρ set to 1.5, 3, and 6, the mean similarities are 0.954, 0.950, and 0.936, respectively.