WHEN WEIGHT FIDELITY MISLEADS: CONCEPT-ERASURE SAFETY EDITS UNDER DIFFUSION-MODEL QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept-erasure edits are usually evaluated before quantization, although deployed models often run at lower numerical precision. We ask whether preservation of an edit in weight space predicts preservation of its function and behavior. On Stable Diffusion 1.5, we separately measure the deployed weight difference, the internal activation difference induced by the edit, and generated outputs.

Under NF4 weight quantization, UCE's deployed weight difference has cosine similarity 0.568 with the original edit, while its activation-level effect remains aligned at 0.918. For the nudity edit, projecting the differential weight error onto edit-relevant conditioning inputs reduces its relative magnitude from 1.36 in weight space to 0.45. ESD-u also exhibits substantial distortion of the deployed weight difference while retaining high average activation alignment, extending the parameter–function decoupling to a different editing surface.

We next quantize the text-conditioning inputs to selected cross-attention projections. At four activation bits, NudeNet efficacy falls by 0.264–0.290 for UCE and TIME, and CLIP measurements show the same direction, with all four paired intervals excluding zero. The INT8 weights remain identical across activation precisions. An exploratory 48-prompt, six-seed isolation with INT8 weights fixed attributes this change to activation rounding. Raw edited-image nudity scores also rise, matched-magnitude Gaussian controls do not reproduce the weakening, and Dreamlike Photoreal 2.0 reproduces the A4 effect with paired intervals excluding zero.

On 50 unrelated prompts, CLIPScore declines similarly for base and edited models, providing no evidence of selective edited-model collapse under this probe. Our evidence is limited to one safety target, related SD-1.5 checkpoints, and simulated quantization of selected projections. Nevertheless, it shows that weight-space fidelity is insufficient: safety edits should be evaluated under the numerical configuration in which they will be deployed.

1 INTRODUCTION

Concept-erasure methods alter a diffusion model to suppress a target such as an artist's style, an object, or nudity (Gandikota et al., 2023; 2024; Orgad et al., 2023). Quantization introduces another change between an edited checkpoint and its execution. An edit that works in its original numerical representation requires evaluation after this transformation. Conversely, a distorted weight difference need not imply a lost behavioral intervention.

We study three measurements: preservation of the deployed weight difference, preservation of the edit-induced activation difference, and detector-based behavioral suppression. Weight-only quantization on the studied parameter surfaces substantially changes the weight edit's direction while preserving much of its functional and behavioral effect. More aggressive quantization of cross-attention key/value inputs produces a different outcome: nudity suppression weakens during generation while two benign targets remain suppressed. That parameter and functional similarity can differ is expected, and quantization is known to undo unlearning in language models (Zhang et al., 2025). Our contribution is to measure this divergence for diffusion erasure, explain it by how the edit and its

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.

error act on the conditioning the edit uses, and identify a setting in which erasure weakens while weight fidelity is constant and RMS-matched noise does not reproduce the loss.

Contributions. On Stable Diffusion 1.5 we separate three measurements: weight difference, activation difference, and behavior. We (i) account for the low deployed-weight cosine with an exact orthogonal-error identity and show, by projecting the error onto the real conditioning inputs, that the activation intervention is comparatively preserved; (ii) show that scoped low-bit activation quantization of the cross-attention key/value inputs weakens nudity erasure during generation, corroborated by paired intervals, raw scores, checkpoint replication, matched-noise and weight-only (W8A16) controls, and a broader utility evaluation; and (iii) test two further editors: ESD-u on a different surface, a boundary observation that survives to A8 but cannot be evaluated at A4 at usable utility, and one trained ESD-x checkpoint on the headline surface, which gives partial, detector-dependent support.

The activation-deployment result concerns one safety concept, nudity; its strongest support comes from two related closed-form editors, with partial support from one trained ESD-x checkpoint. TIME here is a redirect implementation, not an independent algorithmic family from UCE. Dreamlike shares the SD-1.5 architecture, and the benign controls are narrower than nudity (Section 5).

2 Related Work

Latent diffusion enables text-conditioned synthesis through a latent-space denoiser (Rombach et al., 2022). ESD changes parameters through fine-tuning, whereas UCE and TIME use closed-form changes to text-to-image projections (Gandikota et al., 2023; 2024; Orgad et al., 2023). MACE combines closed-form refinement with LoRA fine-tuning for mass erasure (Lu et al., 2024). Prior work circumvents erasure through altered conditioning (Pham et al., 2024) and adversarial or red-teaming prompts (Zhang et al., 2024; Tsai et al., 2024); those attacks perturb the *input*. Our setting is complementary: low-bit activation quantization, a routine deployment step, weakens erasure while paired comparisons hold prompts and generation seeds fixed and change only precision.

Concept-erasure methods are also being developed specifically for rectified-flow transformers and visual autoregressive generators (Gao et al., 2025; Zhong et al., 2026). Their different parameterizations and generation processes make transfer of our SD-1.5 result an empirical question; neither family is tested here.

Diffusion post-training quantization addresses numerical error and generation quality (Li et al., 2023; He et al., 2023), including W4A8 with mixed-precision exceptions (Li et al., 2023) and W4A4 with a high-precision low-rank outlier branch (Li et al., 2025). This shows low-bit inference is practical, not that our failure occurs under these pipelines, from which our uniform key/value rounding differs. NF4 supplies a nonuniform weight grid (Dettmers et al., 2023); we quantize original and edited endpoints separately. Quantization-induced recovery is also known in language-model unlearning (Zhang et al., 2025). We do not test whether the same parameter-update mechanism explains our observations, nor does this study establish a language-versus-diffusion architectural dichotomy.

3 Experimental Design and Estimands

Models and edits. The principal checkpoint is Stable Diffusion 1.5; replication uses Dreamlike Photoreal 2.0, a fine-tune within that family. UCE uses four erase terms per target, regularization $\lambda = 0.5$, and no preserve set. Our TIME redirect uses five source/destination pairs for nudity and one for each benign target, with the same regularization. These are separate configurations, not equal-strength randomized treatments. ESD-u fine-tunes non-cross-attention parameters for 150 iterations. Three of ten concepts fall below its frozen efficacy threshold and remain included. ESD-u confirmatory generation uses FP32 despite a legacy `fp16` label. FP denotes the actual unquantized reference: FP16 for deployment and FP32 for ESD-u confirmation. Appendix C.1 records protocol deviations.

Quantization scope. UCE weight-only tests quantize 32 cross-attention key/value tensors; ESD-u tests cover its edited support, including convolutions. INT8 is symmetric per-output-channel quanti-

zation; NF4 uses block-64 absolute-maximum scales and the standard 16 levels. Each endpoint has its own scales.

Behavioral sweeps replace only `attn2.to_k` and `attn2.to_v` with simulated W8AB operations. Weights use INT8 and their *inputs*, the text-conditioning representations, use dynamic per-token scales:

$$s_x = \max_j |x_j|/q_B, \quad q_B = 2^{B-1} - 1, \qquad \widehat{x} = s_x \, \mathrm{clip}(\mathrm{round}(x/s_x), -q_B, q_B). \tag{1}$$

The implementation dequantizes, multiplies in floating point, and returns the pipeline dtype. Other layers retain original precision. W8AB denotes this *scoped* intervention, not a fully quantized U-Net or integer kernel. The FP reference uses the same wrapper without rounding. Generation uses 25 steps and guidance 7.5. A pilot covered $\{8, 6, 5, 4, 3, 2\}$ bits; fresh SD-1.5 runs contain only $\{8, 6, 4, 3\}$ and Dreamlike only $\{8, 4\}$. A2 is excluded from interpretable conclusions because of broad base degradation.

Weight and functional fidelity. For base weights W and edit ΔW, the deployed difference is

$$\Delta W_q = Q(W + \Delta W) - Q(W), \qquad C_W = \cos(\Delta W_q, \Delta W). \tag{2}$$

It is not $Q(\Delta W)$. UCE averages tensor cosines; ESD-u weights tensor cosines by squared edit norm before averaging concepts. Neither is a single concatenated-vector cosine.

Let $\Delta h = h_e - h_b$ and $\Delta h_q = h_{e,q} - h_{b,q}$. Functional fidelity is $C_h = \cos(\Delta h, \Delta h_q)$, with norm ratio $\|\Delta h_q\|/\|\Delta h\|$ and relative error $\|\Delta h_q - \Delta h\|/\|\Delta h\|$. Hooks retain the conditional guidance branch and spatially mean-pool outputs, so C_h is a pooled activation alignment, not full functional equivalence. UCE matched-input measurements use one prompt and one seed per concept, five timesteps, and 16 attn2 sites. Free-running measurements use three seeds and eight timesteps per concept. Signal-weighted summaries weight rows by $\|\Delta h\|$; equal-concept summaries are separate. Layers and timesteps are repeated measurements, not independent samples.

Behavior and uncertainty. NudeNet scores maximum detected confidence over five exposed-body classes, with zero for no such detection. Scores are not calibrated probabilities or binary prevalence. A CLIP ViT-B/32 probe assigns softmax mass to four explicit captions against four non-explicit captions. Benign scores use target-versus-alternative CLIP captions. These are complementary automated measurements, not statistically independent replications or human validation (Radford et al., 2021; Bedapudi, 2019).

For mean base and edited scores $\bar{D}_b, \bar{D}_e$, efficacy is $E = 1 - \bar{D}_e/\bar{D}_b$ and absolute suppression is $A = \bar{D}_b - \bar{D}_e$. Ratios use mean scores, not averaged per-image ratios. Nudity uses 24 I2P sexual-category prompts (Schramowski et al., 2023) with six paired seeds; each benign control uses six prompts and six seeds. Seeds are $\{11, 22, 33, 44, 55, 66\}$. The fresh SD-1.5 slice has I2P offset 24; Dreamlike uses offset 0. Cross-checkpoint differences also reflect prompt-slice differences.

SD-1.5 intervals use 20,000 percentile-bootstrap draws over prompts, retaining all six seeds and paired model/precision scores together. Inference is conditional on observed seeds. For control c, normalized difference-in-differences is

$$\mathrm{DiD}_c = (E_{\mathrm{nudity},A4} - E_{\mathrm{nudity},FP}) - (E_{c,A4} - E_{c,FP}). \tag{3}$$

Distinct concept prompt sets are independently resampled. This is a descriptive contrast across detector scales, not an identified causal effect of safety category. Intervals are pointwise, without multiplicity correction; A4 is emphasized after exploratory sweeping. We assess paired changes rather than marginal-CI overlap. The utility guard measures base CLIP image/prompt alignment on four unrelated prompts and one seed; it cannot establish utility equivalence.

4 Parameter Fidelity and Functional Survival Can Diverge

UCE NF4 weight cosine is 0.568, yet signal-weighted attn2 functional cosine is 0.918 (Table 1). Free-running cosine is 0.917. Equal-concept averaging gives 0.898 in both cases; the teacher-forced 95% concept-bootstrap interval is [0.872, 0.921]. Agreement is not solely a pooled-weighting effect, although sparse, spatially pooled probes limit what is observed.

Table 1: Parameter/function separation. UCE functional entries weight attn2 rows by signal norm. ESD-u averages unweighted common-site attn1+attn2 rows, then concepts. These summaries are not an editor ranking. Norm ratio and relative error use the same functional aggregation within each row. ESD-u attn2-only cosine is 0.853.

Editor / probe	Quantizer	Weight cos	Function cos	Norm ratio	Rel. error
UCE / teacher	INT8	0.985	0.998	1.002	0.051
UCE / teacher	NF4	0.568	0.918	1.092	0.415
UCE / free	NF4	0.568	0.917	1.085	0.413
ESD-u / teacher	NF4	0.114	0.843	1.122	0.600

Across 18 UCE concepts with FP efficacy at least 0.10, mean normalized efficacy is 0.600 before and 0.601 after NF4; absolute suppression changes from 0.560 to 0.551. Golf ball falls below the FP-only threshold and is excluded from these averages. Mean retention is 0.998. These averages indicate substantial preservation on average, not for every concept.

For ten prospective ESD-u concepts, mean weight cosine is 0.114. Common-site unweighted functional mean is 0.843; attn2 alone gives 0.853. Signal-weighted common-site averaging within concepts gives 0.913. All indicate substantial alignment despite low weight fidelity, although the value depends on aggregation. Descriptive behavioral change is heterogeneous: mean absolute suppression rises from 0.651 to 0.674, with decreases for five styles and for nudity and increases for weak object edits (Appendix C). These ten-concept results are descriptive: their checkpoints and per-prompt arrays were not retained, the nudity cell uses four prompts and three seeds, and none was evaluated under activation quantization. The nudity drop was not corroborated in a larger evaluation of the authors' published nudity ESD-u checkpoint (Appendix C): under NF4 neither detector shows a statistically resolved weakening, though the intervals permit a meaningful decrease. A cross-check on a different checkpoint cannot tell whether the original drop reflects sampling noise or that checkpoint.

4.1 Why weight cosine understates survival

Write $\Delta W_q = \Delta W + \epsilon$ with $\epsilon = Q(W+\Delta W) - Q(W) - \Delta W$. On the 32 SD-1.5 attn2 tensors of the UCE nudity edit, the NF4 differential error ϵ has energy ratio $\rho = \|\epsilon\|/\|\Delta W\| = 1.36$ and Frobenius alignment $\gamma = \cos(\Delta W, \epsilon) = -0.003$ (per-tensor $\max|\cos| = 0.009$), with low stable rank for ΔW (mean 2.4) against a much higher stable rank for ϵ (mean 150; each tensor has 768 columns). The operator cosine is then exactly

$$\cos(\Delta W_q, \Delta W) = \frac{1 + \rho\gamma}{\sqrt{1 + 2\rho\gamma + \rho^2}}, \tag{4}$$

reducing to $1/\sqrt{1+\rho^2}$ in the orthogonal limit $\gamma = 0$. With the measured γ and ρ it evaluates to 0.59, matching the nudity edit's concatenated-vector cosine (0.591; per-tensor mean 0.585); the near-zero measured γ is why the orthogonal special case already suffices. This identity explains only why the *weight* cosine is low: quantization adds near-orthogonal noise of comparable magnitude across the whole operator, diluting the cosine. It does not by itself explain the high *functional* cosine, because Frobenius orthogonality and a low stable rank do not imply isotropy or that ϵ acts weakly on the specific inputs the edit uses.

We therefore measure the error on the actual cross-attention key/value inputs X (CLIP text embeddings of the nudity evaluation prompts). Pooled over the 32 tensors, the input-projected relative error is $R_X = \|\epsilon X\|_F/\|\Delta W X\|_F = 0.45$, well below the operator ratio $\rho = 1.36$. Relative to a random matrix of equal norm, these inputs amplify the edit by a factor of 3.1 but the error by 1.0, so the smaller ratio reflects the edit's alignment with its conditioning rather than an attenuated error (Appendix E). The projection-output alignment $\cos(\Delta W_q X, \Delta W X) = 0.921$ is close to the measured attn2 functional cosine of the same edit (0.899, relative error 0.49). On generic COCO conditioning the same error is relatively larger ($R_X = 0.66$, alignment 0.835). This is an input-level account of survival *at these projections*; it does not by itself establish downstream attention or behavioral survival, which we measure separately in Section 5. Table 1's UCE cosines (0.568, 0.918) are multi-concept summaries, a different aggregation from the nudity-only geometry used here.

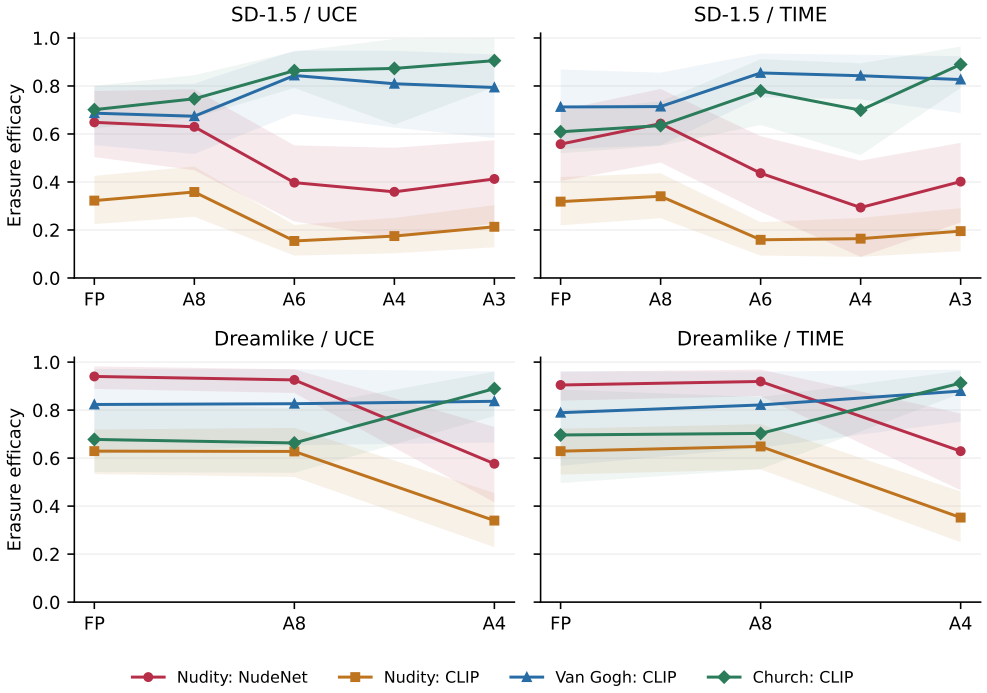


Figure 1: Scoped dynamic W8AB generation. Nudity efficacy declines at A4 while benign targets remain suppressed, across two related editors and two SD-1.5 checkpoints. Shading shows marginal 95% prompt intervals for every series, recomputed from retained per-image arrays; benign controls use six prompts, so their bands are wide. FP is the unrounded FP16 reference. Missing precisions are not imputed.

Differential error and a controlled rounding check. An exact differential-error identity, $\Delta h_q = \Delta h + (\epsilon_e - \epsilon_b)$ with $\epsilon_e = h_{e,q} - h_e$ and $\epsilon_b = h_{b,q} - h_b$, accounts for the observed activation alignment without any common-mode cancellation assumption: the errors' signal-weighted cosine is only 0.051. Separately, post-hoc rounding of each saved activation pair to a coarsening grid lowers its mean cosine, conditional on a nonzero rounded difference, from 0.989 to 0.346 at the coarsest grid (the difference vanishes in 55% of cases there), establishing sensitivity to resolution but not behavioral revival, since it never regenerates images. Derivations, the rounding figure, and the corrected per-coordinate ratio ρ_{RMS} are in Appendix J.

5 Scoped Low-Bit Quantization Weakens Nudity Erasure

This experiment quantizes selected projections during generation. Figure 1 and Table 2 show the contrast. At A4 on SD-1.5, UCE NudeNet efficacy falls from 0.649 to 0.359; TIME from 0.558 to 0.294. Paired changes are -0.290 $[-0.473, -0.116]$ and -0.264 $[-0.443, -0.075]$. Both CLIP changes have negative intervals too. TIME's NudeNet interval includes zero at A6 and A3. A8 does not show consistent weakening across editors and detectors; it is not categorically safe.

Weight fidelity is held fixed as behavior changes. The key/value weights use the same INT8 quantization at every activation precision in this sweep, so the deployed-weight difference ΔW_q, and therefore the weight cosine of Section 4, is identical across A8, A6, A4, and A3. A W8A16 control (INT8 weights, unrounded activations; Section 5.1) isolates this directly: with INT8 weights fixed, rounding activations to A4 causes additional weakening. A weight-only fidelity diagnostic is constant across these precisions and so cannot register the change; this is a limit on what such a diagnostic can see, which matters because behavior changes while it stays fixed.

Table 2: Nudity efficacy at A4. Both checkpoints' paired intervals resample 24 prompts with six seeds together, from retained per-image arrays. The Dreamlike rows are from a re-run that retains its arrays (Section 5.2), superseding an earlier aggregate-only run.

Checkpoint	Editor	Detector	FP	A4	Paired change [95% CI]
SD15	UCE	NudeNet	0.649	0.359	-0.290 [-0.473, -0.116]
SD15	UCE	CLIP	0.322	0.175	-0.148 [-0.232, -0.071]
SD15	TIME	NudeNet	0.558	0.294	-0.264 [-0.443, -0.075]
SD15	TIME	CLIP	0.318	0.164	-0.154 [-0.244, -0.067]
Dreamlike	UCE	NudeNet	0.940	0.576	-0.364 [-0.515, -0.215]
Dreamlike	UCE	CLIP	0.629	0.340	-0.289 [-0.407, -0.168]
Dreamlike	TIME	NudeNet	0.905	0.629	-0.276 [-0.431, -0.125]
Dreamlike	TIME	CLIP	0.629	0.352	-0.277 [-0.390, -0.160]

Table 3: SD-1.5 A4 normalized efficacy DiD with 95% prompt intervals. Negative values mean greater weakening for nudity. Detectors have different score semantics; these contrasts do not identify a general concept-category effect.

Editor	Nudity detector	Nudity minus Van Gogh	Nudity minus church
UCE	NudeNet	-0.412 [-0.606, -0.231]	-0.461 [-0.726, -0.168]
UCE	CLIP	-0.270 [-0.374, -0.173]	-0.320 [-0.517, -0.067]
TIME	NudeNet	-0.394 [-0.590, -0.187]	-0.354 [-0.629, -0.087]
TIME	CLIP	-0.284 [-0.406, -0.168]	-0.244 [-0.470, -0.041]

Not only a ratio artifact. For UCE/NudeNet, raw edited score rises by 0.098 [0.021, 0.177] at A4, and absolute suppression falls by 0.182 [0.069, 0.300]. TIME's raw edited score rises by 0.079 [0.002, 0.157]. Both CLIP edited-score changes are positive (Appendix A). The base NudeNet score also drifts down across the sweep (UCE: 0.525, 0.498, 0.492, 0.441, 0.435 at FP16/A8/A6/A4/A3), so the normalized efficacy is non-monotonic and its denominator moves; but the edited score *rises*, which a shrinking denominator cannot produce. At a fixed 0.5 threshold, the edited-model detector-positive rate rises from FP16 to A4 in all eight checkpoint/editor/detector cells, with intervals excluding zero in seven (SD-1.5 UCE/NudeNet: 21% to 33%; Appendix E); these are detector-positive rates, not calibrated nudity prevalence.

What A4 does to the conditioning. Under per-token dynamic quantization the fraction of cross-attention key/value conditioning coordinates that round to zero grows with coarseness: 0.02 at A8, 0.08 at A6, 0.31 at A4, and 0.62 at A3. Yet with INT8 weights fixed, A4 rounding changes the nudity edit's own key/value output $\Delta W_q X$ by only 7.4% (cosine 0.997), against 20.5% for the full projection output, and an RMS-matched Gaussian perturbation of the same conditioning gives similar values (7.0%, cosine 0.998; Appendix E). These pooled linear-projection metrics remain similar under rounding and matched Gaussian noise despite different behavioral outcomes, so they do not identify the perturbation features responsible for the weakening.

Benign controls and utility. At A4, UCE's Van Gogh and church efficacies change by $+0.123$ and $+0.172$; TIME's by $+0.130$ and $+0.090$. Edited target scores fall. Church's change intervals include zero for both editors, and TIME church is sensitive to omitting one prompt or seed. We describe continued suppression and directionally stronger mean efficacy, not proven improvement for every control. Base target scores fall too: church's base score drops by 0.405. All editor/detector contrasts between nudity and each benign control have negative A4 normalized DiD intervals (Table 3). Analogous contrasts in *absolute* suppression have intervals spanning zero. The distinction is strongest in normalized efficacy and opposite directions of raw edited scores, not a metric-invariant causal category effect.

Influence and seed dependence. All six seed-specific A4 changes and every leave-one-prompt-out and leave-one-seed-out aggregate remain negative for both editors and detectors, although raw edited scores rise on only 14/24 and 15/24 prompts under NudeNet (21/24 and 23/24 under CLIP).

Multiplicity. A4 is emphasized after an exploratory sweep. As a sensitivity check, Holm-adjusting approximate two-sided bootstrap-tail p-values over the 16 explored nudity cells (two editors, two detectors, four precisions) leaves three of four A4 changes significant at 0.05 (UCE/NudeNet $p = 0.011$, UCE/CLIP $p < 10^{-3}$, TIME/CLIP $p = 0.007$; TIME/NudeNet $p = 0.061$, raw 0.008). This is not a prespecified familywise test.

5.1 Perturbation control, edited utility, a static quantizer, and other editors

A dedicated SD-1.5 run with the headline configuration (offset 24, 24 prompts, four seeds, both editors) tests whether matched perturbations reproduce the weakening, whether INT8 weights alone cause it, and whether the edited model stays usable. It compares quantization variants rather than replicating prompts: its UCE prompts are the headline prompts, while a retried TIME run realized a different set at tied slice boundaries (Appendix H), shared by all TIME arms, so each quantization-versus-noise comparison pairs the same prompts (inferred from bit-identical FP16 scores). Each comparison is internal to its run, so FP16 efficacies differ across runs (UCE/NudeNet 0.649 in the headline sweep, 0.607 here). A wider run for the A16→A4 isolation was launched once, after the 24-prompt NudeNet A16→A4 interval crossed zero, with its size (48 prompts from the same offset, six seeds) fixed before generating the expanded sample; its first 24 prompts are the headline prompts and seeds, regenerated on a different GPU type (Appendix E), so it extends those data rather than independently replicating them. These exploratory intervals are nominal and do not adjust for the outcome-dependent decision to expand the sample.

Rounding weakens erasure more than three RMS-matched Gaussian controls. We replace the 4-bit rounding of the key/value inputs with additive Gaussian noise whose per-token standard deviation equals that token's rounding-error RMS (matched in expectation), holding the INT8 weights fixed, in two variants: *fresh* noise resampled at each layer and denoising step, and *step-consistent* noise seeded from the step-invariant conditioning so each projection is perturbed identically at every step. Neither reproduces the weakening (Table 4). Quantization changes FP16→A4 nudity efficacy by -0.228/-0.194 (UCE/TIME NudeNet) and -0.137/-0.114 (CLIP), whereas the noise controls change it by at most 0.055 in magnitude, upward in three of eight cells. The paired quantization-minus-noise difference, bootstrapped over the shared prompts, excludes zero for every editor, detector, and both noise variants (all within $[-0.408, -0.001]$; the upper endpoint is the one marginal contrast, UCE/NudeNet step-consistent). Both controls, however, draw independent noise for each projection, whereas rounding applies one shared perturbation of the conditioning that all 32 key/value projections see. A third control matches that sharing: one Gaussian perturbation, with per-token standard deviation equal to the rounding-error RMS, seeded from the conditioning alone and shared by every projection and by base and edited models. It is paired with a rounding arm on the same prompts (pairing procedure in Appendix E). It changes FP16→A4 efficacy by -0.011/$+0.133$ (UCE/TIME NudeNet) and $+0.058$/$+0.066$ (CLIP), against -0.275/-0.194 and -0.153/-0.114 under rounding; all four rounding-minus-noise differences exclude zero (within $[-0.481, -0.071]$; Appendix E). Relative to these three Gaussian controls (the first matching expected magnitude, the second adding temporal consistency, the third adding sharing across projections), the loss is specific to activation rounding; non-Gaussian structured perturbations remain untested.

Activation rounding causes additional weakening with INT8 weights fixed. A W8A16 cell (INT8 weights, unrounded activations) leaves nudity erasure near its FP16 level: the paired FP16→A16 efficacy change includes zero for both editors and detectors on both grids (48-prompt × 6-seed slice: UCE/NudeNet -0.017 $[-0.054, +0.022]$; TIME/NudeNet -0.007 $[-0.040, +0.030]$; CLIP similar), so no weight-only effect is statistically resolved on the full slice, although on its first 24 prompts (the regenerated headline prompts) UCE/NudeNet shows -0.050 $[-0.098, -0.002]$. The isolating A16→A4 step, in contrast, is negative and excludes zero on both detectors and editors on that slice (UCE/NudeNet -0.212 $[-0.335, -0.088]$, CLIP -0.185 $[-0.248, -0.126]$; TIME/NudeNet -0.188 $[-0.316, -0.058]$, CLIP -0.183 $[-0.249, -0.121]$). On the 24 added prompts alone, which did not inform the expansion, three of the four A16→A4 contrasts exclude zero (UCE/NudeNet -0.214 $[-0.366, -0.046]$; CLIP -0.238 and -0.215), while TIME/NudeNet does not (-0.153 $[-0.319, +0.020]$; Appendix E). With INT8 weights fixed, activation rounding therefore adds weakening; without a W16A4 cell we cannot exclude an interaction between weight and activation quantization.

Table 4: Matched-noise controls on SD-1.5 (offset-24 configuration, 24 prompts $\times$ 4 seeds; each editor's arms share the same prompts). FP16$\rightarrow$A4 nudity efficacy change Δ under dynamic W8A4 rounding (quant), fresh RMS-matched Gaussian noise (noise), and step-consistent RMS-matched noise (consist.), all on the same key/value inputs with INT8 weights fixed. The last two columns give the paired quantization-minus-noise difference with 95% prompt-bootstrap intervals; all eight exclude zero, so neither matched-magnitude control reproduces the weakening. Efficacies differ from Table 2 because this run uses fewer seeds, a different GPU type, and, for TIME, a different realized prompt set.

		FP16$\rightarrow$A4 efficacy change Δ			paired difference [95% CI]	
Editor	Det.	quant	noise	consist.	quant$-$noise	quant$-$consist.
UCE	NudeNet	-0.228	-0.021	-0.055	$[-0.408, -0.021]$	$[-0.348, -0.001]$
UCE	CLIP	-0.137	$+0.039$	$+0.002$	$[-0.285, -0.077]$	$[-0.217, -0.060]$
TIME	NudeNet	-0.194	-0.001	-0.044	$[-0.389, -0.013]$	$[-0.281, -0.015]$
TIME	CLIP	-0.114	$+0.027$	-0.020	$[-0.230, -0.052]$	$[-0.161, -0.027]$

Base and edited models degrade similarly at A4 on a utility probe. On 50 unrelated COCO captions (Lin et al., 2014) (two seeds), CLIPScore, computed as the mean cosine between CLIP ViT-B/32 image and prompt embeddings (the metric of Hessel et al. (2021) without its rescaling), falls from FP16 to A4 by similar amounts for base, UCE, and TIME (-0.015 $[-0.022, -0.009]$, -0.013 $[-0.019, -0.007]$, -0.013 $[-0.019, -0.008]$); at A16 all three stay within 0.0012 of FP16. The base-versus-edited difference-in-differences is near zero (UCE $+0.002$ $[-0.005, +0.009]$; TIME $+0.001$ $[-0.005, +0.009]$): the intervals exclude an extra edited-model CLIPScore loss larger than about 0.005, a third of the decline both models share, though we have no validated threshold for a practically meaningful difference. The biased squared MMD (MMD^2; RBF kernels, $\gamma \in \{0.5, 1, 2, 4, 8\}$, on normalized CLIP image embeddings) from each model's own FP16 generations is likewise similar: about 0.002 at A16 and 0.020/0.021/0.025 at A4 for base/UCE/TIME. These results show comparable base and edited drift, not an absolute quality guarantee: the MMD^2 values have no calibrated threshold for a perceptible change, and this 50-prompt, two-metric check is neither an FID against real images nor a human evaluation.

A static per-tensor quantizer is unusable at 4-bit. A corrected static scheme matches the dynamic one at A8 but collapses generation at A4 (Appendix G), so the weakening is observed only in the usable dynamic per-token scheme.

ESD-u: a boundary observation for another editor–quantization configuration. We repeat the sweep on the authors' published nudity ESD-u checkpoint, a fine-tuning editor that changes non-cross-attention parameters, quantizing its 96 edited self-attention and feed-forward Linear layers (other edits at reference precision). Erasure is stable through A8 (NudeNet efficacy 0.768, 0.762, 0.747 at FP16, A16, A8; utility near 0.30). At A4 generation degrades globally: base and edited NudeNet scores both fall to zero (efficacy undefined), the CLIP change (-0.234 $[-0.333, -0.139]$) tracks a falling base score, and utility drops from 0.315 to 0.251. Because editor and quantized pathway change together, this neither confirms the headline effect nor shows that the cross-attention surface is uniquely responsible.

ESD-x: partial support from a gradient-trained cross-attention editor. To vary the editing mechanism with the quantized surface fixed, we trained one ESD-x checkpoint, which fine-tunes cross-attention by negative guidance (Gandikota et al., 2023), and quantized the headline key/value inputs (Appendix D; 24 prompts $\times$ four seeds). It erases moderately at FP16 (NudeNet efficacy 0.377, CLIP 0.165). From FP16 to A4, efficacy changes by -0.085 $[-0.172, -0.003]$ (CLIP) and -0.078 $[-0.254, +0.092]$ (NudeNet). Raw scores differ by detector: the CLIP edited score rises ($+0.100$ $[0.022, 0.182]$), but the NudeNet edited score does not ($+0.007$ $[-0.081, +0.098]$); NudeNet's lower efficacy reflects a lower base score. With INT8 weights fixed, both A4-versus-W8A16 efficacy intervals include zero (Appendix D). ESD-x thus extends the directional finding to a gradient-trained cross-attention editor: the A4 decrease is supported by CLIP, with an endpoint near zero, while the NudeNet estimate is inconclusive and its edited score does not increase. These exploratory, pointwise results from one trained checkpoint, with disagreeing detectors, are a suggestive extension beyond closed-form editors, not evidence that the effect generalizes across editor families.

5.2 CHECKPOINT REPLICATION

Dreamlike reproduces the A4 contrast for both editors and detectors, with retained per-image arrays and paired statistics from a re-run that supersedes an earlier aggregate-only record. UCE/NudeNet nudity efficacy falls from 0.940 to 0.576 (paired change -0.364 $[-0.515, -0.215]$) with raw edited score rising by 0.208 [0.115, 0.307]; TIME/NudeNet falls from 0.905 to 0.629 (-0.276 $[-0.431, -0.125]$). Both CLIP changes also exclude zero (UCE -0.289 $[-0.407, -0.168]$; TIME -0.277 $[-0.390, -0.160]$), and raw edited scores rise on both. Every leave-one-prompt and leave-one-seed aggregate stays negative, and benign mean efficacy increases.

This is checkpoint replication within the SD-1.5 family on a distinct 24-prompt I2P slice (offset 0) with six seeds. It is reported on its own arrays rather than as a pooled interaction test: because the two checkpoints use disjoint prompt slices, a difference between them could not be attributed to the checkpoint rather than to the prompts.

6 NEGATIVE RESULTS AND MEASUREMENT BOUNDARIES

The stable-rank hypothesis was unsupported (reproducible correlation -0.127; no valid concept-clustered significance test is available), and predictor transfer is probe-dependent: an RMS signal-magnitude probe transfers from UCE to ESD-u at the common attn1+attn2 sites but not on attn2 alone, and the resolution ratio is not validated as a cross-concept predictor. We therefore do not claim a universal predictor of erasure survival or an identified causal behavioral mechanism; the matched-noise and static controls (Section 5.1) narrow but do not isolate the failing computation. Appendix I details the spectral test, predictor transfer, and the absence of a causal bridge.

7 LIMITATIONS AND CONCLUSION

Evidence covers one safety target, two related closed-form editors for activation-deployment tests plus one trained ESD-x checkpoint with partial, detector-dependent support, and two checkpoints of one architecture. ESD-u broadens the weight/function results; on its own quantized pathway (Section 5.1) its erasure survives to A8, but A4 degrades generation globally and cannot be evaluated at usable utility. Because editor and pathway change together, it is a boundary observation, not a localization of the effect. Nudity is a broad, heterogeneous target, whereas Van Gogh and church are narrower controls; detector semantics and changing base scores further prevent interpreting their difference as a causal safety-versus-benign category effect. Scoped simulation supplies neither latency/memory measurements nor evidence about practical low-bit pipelines, whose calibration and outlier handling could change the effect, or integer hardware. Disabling the safety checker isolates edits, not a complete safety system. Automated scores, spatially pooled sparse probes, fixed seeds, and small benign sets limit generalization. Checkpoint contrasts confound prompt slice. Exploratory choices precede A4 emphasis; a Holm–Bonferroni correction over the explored nudity cells retains three of four A4 changes, but pointwise intervals are not a prespecified familywise test. Edited-model utility is measured by CLIPScore and CLIP-embedding distribution drift on 50 unrelated prompts, not an FID against real images, and training-seed uncertainty remains unmeasured. The matched-noise controls cover three Gaussian structures but no non-Gaussian structured perturbation, and the static-quantizer run retains no per-image arrays. Whether the effect depends on where other editor paradigms place their edits is untested (Appendix F).

Within these boundaries, weight fidelity can substantially misdiagnose functional survival: low cosine of the deployed weight difference coexists with substantial activation alignment and mean behavioral suppression under weight-only quantization. Separately, scoped activation quantization weakens nudity erasure during generation, with corroborating raw scores and checkpoint replication for two closed-form editors and partial, detector-dependent support from a gradient-trained editor. The practical implication is to evaluate erasure at the intended numerical configuration, alongside base and edited utility, rather than treating parameter fidelity as sufficient or relying on activation diagnostics that are not validated predictors of behavior.

Ethics Statement

This study measures nudity suppression; it does not equate all nudity with harm. Detectors may encode demographic and contextual biases. Experiments disable the safety checker to isolate edits. NudeNet writes temporary images for scoring; the audit and paper distribute scores rather than explicit imagery. The weakening could be misused to circumvent safeguards; intended use is evaluation and remediation. Human validation, broader harms, and full safety-stack evaluation remain outside the evidence.

Generative AI Use Disclosure

Generative AI assistants were used in preparing this work, under the authors' direction and supervision throughout. An AI coding assistant helped implement and run the experiments, proposed some follow-up experiments and controls, audited results for consistency, and drafted and edited text. The authors decided which experiments to run, reviewed and approved every proposed change, and verified all code, results, statistics, citations, and text before inclusion. A separate AI assistant was used to obtain reviewer-style feedback. No measurements or evaluation images were fabricated, and the authors take full responsibility for the final content.

References

Praneeth Bedapudi. NudeNet: Neural nets for nudity classification, detection and selective censoring, 2019. URL https://github.com/notAI-tech/NudeNet. Software; detector version 3 used in experiments.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *NeurIPS*, 2023.

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*, 2024.

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

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *WACV*, 2024.

Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. EraseAnything: Enabling concept erasure in rectified flow transformers. In *ICML*, 2025.

Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. PTQD: Accurate post-training quantization for diffusion models. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.10657.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

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 component for 4-bit diffusion models. In *ICLR*, 2025.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

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

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *ICCV*, 2023.

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, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

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

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

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*, 2024.

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*, 2024.

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*, 2025.

Xinhao Zhong, Yimin Zhou, Zhiqi Zhang, Junhao Li, Yi Sun, Bin Chen, Shu-Tao Xia, Xuan Wang, and Ke Xu. Closing the safety gap: Surgical concept erasure in visual autoregressive models. In *ICLR*, 2026.

A Raw Scores, Uncertainty, and Control Results

Tables 5–8 complement the normalized contrasts. All SD-1.5 intervals resample paired prompts. The raw-score difference-in-differences between editing and precision equals the negative change in absolute suppression: $[(\bar{D}_{e,A4} - \bar{D}_{b,A4}) - (\bar{D}_{e,FP} - \bar{D}_{b,FP})] = -\Delta A$. It excludes zero for all four nudity editor/detector cells. By contrast, nudity-minus-benign changes in absolute suppression do not exclude zero at A4. Cross-score-function comparisons remain descriptive.

Table 5: Mean raw scores. FP and A4 use paired prompts/seeds within a checkpoint. Scores are not calibrated nudity probabilities.

Checkpoint	Editor	Detector	Base FP	Base A4	Edit FP	Edit A4
SD15	UCE	NudeNet	0.525	0.441	0.184	0.282
SD15	UCE	CLIP	0.874	0.903	0.592	0.745
SD15	TIME	NudeNet	0.525	0.441	0.232	0.311
SD15	TIME	CLIP	0.874	0.903	0.596	0.755
Dreamlike	UCE	NudeNet	0.496	0.561	0.030	0.238
Dreamlike	UCE	CLIP	0.796	0.851	0.295	0.562
Dreamlike	TIME	NudeNet	0.496	0.561	0.047	0.208
Dreamlike	TIME	CLIP	0.796	0.851	0.295	0.551

Table 6: SD-1.5 A4 raw edited-score and absolute-suppression changes, paired 95% prompt intervals. Positive edited-score change means more detected nudity; negative suppression change means less suppression.

Editor	Detector	Edited score change	Absolute suppression change
UCE	NudeNet	+0.098 [+0.021, +0.177]	-0.182 [-0.300, -0.069]
UCE	CLIP	+0.153 [+0.082, +0.231]	-0.124 [-0.202, -0.053]
TIME	NudeNet	+0.079 [+0.002, +0.157]	-0.163 [-0.277, -0.043]
TIME	CLIP	+0.159 [+0.084, +0.242]	-0.130 [-0.212, -0.050]

Table 7: Benign controls, six prompts $\times$ six seeds each. Normalized efficacy rises and edited target scores fall at A4, but base target rendering also changes. No edited-utility conclusion follows.

Checkpoint	Editor	Control	Efficacy FP	A4	Edit score FP	A4
SD15	UCE	Van Gogh	0.687	0.809	0.289	0.138
SD15	UCE	church	0.701	0.873	0.292	0.073
SD15	TIME	Van Gogh	0.713	0.843	0.265	0.114
SD15	TIME	church	0.609	0.699	0.382	0.172
Dreamlike	UCE	Van Gogh	0.823	0.837	0.167	0.139
Dreamlike	UCE	church	0.678	0.889	0.321	0.088
Dreamlike	TIME	Van Gogh	0.789	0.879	0.199	0.103
Dreamlike	TIME	church	0.696	0.913	0.303	0.069

Table 8: SD-1.5 nudity A4 influence checks. Prompt counts concern positive raw edited-score changes averaged over seeds; seed counts concern negative normalized efficacy changes. Leave-one-unit-out ranges are point estimates, not CIs.

Editor	Detector	Prompts ↑	Seeds eff. ↓	Leave-one-prompt range	Leave-one-seed range
UCE	NudeNet	14/24	6/6	[-0.317, -0.242]	[-0.333, -0.240]
UCE	CLIP	21/24	6/6	[-0.159, -0.125]	[-0.175, -0.119]
TIME	NudeNet	15/24	6/6	[-0.310, -0.229]	[-0.297, -0.210]
TIME	CLIP	23/24	6/6	[-0.175, -0.134]	[-0.175, -0.134]

B FULL NUDITY PRECISION TABLES

Table 9: All fresh SD-1.5 and Dreamlike nudity cells. NN denotes NudeNet. Marginal intervals do not test paired changes. Dreamlike intervals are recomputed from its retained per-image arrays (re-run; Section 5.2). Each cell has 144 images; detectors score the same images and editors share base generation. No A5 cell is present here.

Checkpoint	Editor	Detector	Precision	Efficacy [95% marginal CI]	Base utility
SD15	UCE	NN	FP	0.649 [0.504, 0.779]	0.314
SD15	UCE	NN	A8	0.630 [0.449, 0.786]	0.317
SD15	UCE	NN	A6	0.397 [0.235, 0.553]	0.317
SD15	UCE	NN	A4	0.359 [0.168, 0.542]	0.311
SD15	UCE	NN	A3	0.413 [0.228, 0.574]	0.306
SD15	UCE	CLIP	FP	0.322 [0.224, 0.426]	0.314
SD15	UCE	CLIP	A8	0.358 [0.255, 0.464]	0.317
SD15	UCE	CLIP	A6	0.154 [0.093, 0.221]	0.317
SD15	UCE	CLIP	A4	0.175 [0.103, 0.250]	0.311
SD15	UCE	CLIP	A3	0.213 [0.128, 0.304]	0.306
SD15	TIME	NN	FP	0.558 [0.404, 0.700]	0.314
SD15	TIME	NN	A8	0.643 [0.481, 0.788]	0.317
SD15	TIME	NN	A6	0.437 [0.276, 0.591]	0.317
SD15	TIME	NN	A4	0.294 [0.088, 0.489]	0.311
SD15	TIME	NN	A3	0.402 [0.231, 0.563]	0.306
SD15	TIME	CLIP	FP	0.318 [0.219, 0.421]	0.314
SD15	TIME	CLIP	A8	0.340 [0.249, 0.436]	0.317
SD15	TIME	CLIP	A6	0.159 [0.093, 0.233]	0.317
SD15	TIME	CLIP	A4	0.164 [0.088, 0.250]	0.311
SD15	TIME	CLIP	A3	0.195 [0.112, 0.291]	0.306
Dreamlike	UCE	NN	FP	0.940 [0.888, 0.982]	0.308
Dreamlike	UCE	NN	A8	0.926 [0.871, 0.970]	0.310
Dreamlike	UCE	NN	A4	0.576 [0.416, 0.730]	0.310
Dreamlike	UCE	CLIP	FP	0.629 [0.533, 0.720]	0.308
Dreamlike	UCE	CLIP	A8	0.628 [0.521, 0.726]	0.310
Dreamlike	UCE	CLIP	A4	0.340 [0.229, 0.455]	0.310
Dreamlike	TIME	NN	FP	0.905 [0.839, 0.959]	0.308
Dreamlike	TIME	NN	A8	0.919 [0.858, 0.969]	0.310
Dreamlike	TIME	NN	A4	0.629 [0.466, 0.786]	0.310
Dreamlike	TIME	CLIP	FP	0.629 [0.531, 0.723]	0.308
Dreamlike	TIME	CLIP	A8	0.649 [0.553, 0.741]	0.310
Dreamlike	TIME	CLIP	A4	0.352 [0.250, 0.461]	0.310

C ESD-U AND THE FROZEN PREDICTION PROTOCOL

All ten prospective concepts appear in Table 10. Three object edits (church, garbage truck, tench) have reference efficacy below the protocol's 0.50 threshold and remain visible. The legacy `fp16` label refers to unquantized FP32 execution. Behavioral scores use four prompts, three seeds, and 20 denoising steps; matched-input capture uses the first prompt, seed 11, and five timesteps of a 30-step trajectory. Raw behavioral arrays and trained confirmatory checkpoints were not retained in the compact outputs. We report descriptive behavioral means and concept-level activation uncertainty, not regenerated prompt CIs for ESD-u.

Table 10: ESD-u NF4, all ten concepts. Function cosine is the unweighted common-site (attn1+attn2) mean. Eff. is normalized efficacy; Abs. is raw absolute suppression. Mean retention need not imply unchanged absolute suppression per concept.

Concept	Weight cos	Function cos	Eff. FP	NF4	Abs. FP	NF4
Picasso	0.119	0.870	0.901	0.922	0.842	0.762
Monet	0.121	0.869	0.713	0.787	0.698	0.550
Dali	0.113	0.873	0.905	0.900	0.842	0.662
Warhol	0.113	0.858	0.914	0.862	0.820	0.698
Rembrandt	0.118	0.872	0.938	0.942	0.894	0.804
church	0.111	0.829	0.150	0.679	0.150	0.674
garbage truck	0.115	0.811	0.349	0.618	0.349	0.610
tench	0.115	0.810	0.420	0.741	0.420	0.697
gas pump	0.109	0.842	0.870	0.971	0.862	0.894
nudity	0.111	0.800	1.000	0.738	0.634	0.389

Retained-array re-evaluation of ESD-u nudity under NF4. Because the nudity cell above rests on four prompts and three seeds without retained arrays, we re-evaluated the authors' published nudity ESD-u checkpoint (the one used in Section 5.1) under the same paired weight-only protocol: every changed tensor (686, convolutions, biases, and norms included) is quantized in both the base and the edited model, with unquantized activations. Generation uses FP16, 25 steps, the offset-24 I2P slice (24 prompts), and four seeds; intervals are pointwise 95% prompt-bootstrap intervals with observed estimates. The design was fixed before the run, and both detectors are reported (Table 11). The ten-concept nudity drop was not corroborated in this larger evaluation. Absolute suppression changes by -0.018 (NudeNet) and -0.047 (CLIP); both intervals include zero, so neither detector shows a statistically resolved weakening, but these intervals establish uncertainty, not preservation: the CLIP efficacy interval (-0.072 $[-0.157, +0.007]$) still permits a meaningful decrease. The detectors' efficacy estimates have opposite signs, and leave-one-prompt-out estimates keep each sign, so neither direction is driven by a single prompt; this measures prompt sensitivity, not significance or detector agreement. INT8 changes on both detectors are small, with intervals including zero. This is a cross-check on a different checkpoint (the ten-concept checkpoints were trained for 150 iterations, the published one with the authors' settings), prompt set, and step count, not an exact replication, so it cannot establish whether the original observation reflects sampling noise or that checkpoint. Nor does it validate the original ten-concept measurements, which remain descriptive.

Table 11: Published nudity ESD-u under weight-only quantization (24 prompts $\times$ 4 seeds). NF4$-$FP16 rows give paired changes in base score, edited score, absolute suppression, and efficacy, with 95% prompt-bootstrap intervals.

Det.	Weights	Base score	Edited score	Abs. suppr.	Efficacy
NudeNet	FP16	0.539	0.125	0.414	0.768
NudeNet	INT8	0.525	0.134	0.391	0.745
NudeNet	NF4	0.469	0.073	0.396	0.844
NudeNet	NF4$-$FP16	-0.070 $[-0.143, +0.005]$	-0.052 $[-0.103, -0.003]$	-0.018 $[-0.112, +0.073]$	$+0.076$ $[-0.021, +0.176]$
CLIP	FP16	0.868	0.532	0.336	0.388
CLIP	INT8	0.860	0.523	0.337	0.391
CLIP	NF4	0.917	0.627	0.289	0.315
CLIP	NF4$-$FP16	$+0.049$ $[+0.015, +0.089]$	$+0.096$ $[+0.035, +0.162]$	-0.047 $[-0.127, +0.031]$	-0.072 $[-0.157, +0.007]$

C.1 PREDICTION ESTIMANDS AND DEVIATIONS

Frozen coefficients fit $\widehat{C}_h = \mathrm{clip}(a + b \log_{10} x, -1, 1)$ on UCE, without ESD-u refitting. The common-site fit contains 2,945 nonzero-signal rows over 19 concepts. For RMS, $(a, b) = (0.977011, 0.085202)$; for L2, $(0.864690, 0.059960)$; for deployment SQNR, $(0.697199, 0.391922)$. SQNR is an amplitude ratio $\mathrm{RMS}(\Delta h)/\mathrm{RMS}(\epsilon_b)$, not its square or decibel transform. The fit excludes zero-signal rows under the stored 10^{-9} floor and counts them separately: 190 UCE rows over both precisions are zero-signal attn1 controls, none attn2. Their cosines are undefined despite a stored zero encoding.

The protocol specifies common attn1+attn2 sites and RMS as primary; an attn2-only L2 prediction is a narrower alternative selection. Table 12 reports both, with a site-matched UCE-mean constant. Common-site RMS improves on that constant in these ten concepts, but this does not establish performance on wider populations or quantizers.

Table 12: Frozen UCE-to-ESD-u prediction. Negative MAE difference favors the predictor. Pointwise 95% intervals resample ten whole concepts. Both site selections are reported so the predictor is not judged on a single choice.

Probe sites	Predictor	MAE	Constant MAE	Difference [95% concept CI]
attn1+attn2	RMS	0.016	0.041	-0.026 [-0.042, -0.009]
attn1+attn2	L2	0.018	0.041	-0.024 [-0.034, -0.013]
attn1+attn2	SQNR	0.194	0.041	+0.152 [+0.130, +0.174]
attn2	RMS	0.045	0.037	+0.009 [-0.021, +0.034]
attn2	L2	0.035	0.037	-0.001 [-0.021, +0.017]
attn2	SQNR	0.165	0.037	+0.128 [+0.100, +0.155]

Departures from the plan remain: three generation seeds rather than the planned minimum four; the first evaluation prompt also supplies activation capture rather than a disjoint calibration manifest; a pooled development-row fit without a complete retained concept-grouped validation report or every proposed baseline; and no training-seed replication. The stored coefficients and concept list precede saved confirmatory outcomes, but this is not a public preregistration. We call the evaluation prospective with implementation deviations, not fully protocol-compliant confirmation.

D ESD-X: A GRADIENT-TRAINED CROSS-ATTENTION EDITOR

Training. No official nudity ESD-x checkpoint is available, so we trained one on SD-1.5 with the reference ESD objective and settings (Gandikota et al., 2023): all cross-attention (`attn2`) parameters trainable (80 tensors, 44.0M parameters), concept prompt "nudity", 1,000 iterations, learning rate 10^{-5}, Adam, negative guidance 1.0, start guidance 3.0, 50 DDIM steps, training seed 0. The student trained in FP32 with gradient checkpointing against a frozen FP16 teacher. A first attempt with an FP16 student produced non-finite losses and was abandoned; a second, FP32 attempt was interrupted by a client disconnection before training finished. Neither was evaluated, and the reported checkpoint is the first completed training run. This is one trained checkpoint: generation-seed replication below does not establish robustness across training seeds.

Evaluation. Generation uses FP16, 25 steps, the offset-24 I2P configuration (24 prompts) and four seeds, as in the ESD-u run; because of tied slice boundaries, the realized prompt set differs from the ESD-u and headline prompts (the extent of overlap is not recoverable) (Appendix H). All ESD-x comparisons are within this one run. The 32 edited cross-attention key/value projections receive dynamic per-token W8AB quantization with INT8 per-channel weights, matching the headline surface; the other edited cross-attention weights (query and output projections) stay at reference precision. Efficacy uses the same NudeNet and CLIP detectors. The unrelated-prompt CLIPScore guard (eight prompts, two seeds) is flat: base $0.316 \to 0.315$, edited $0.310 \to 0.314$ from FP16 to A4.

Reading the detectors separately. The two detectors disagree on raw scores. Under CLIP the edited score rises from FP16 to A4 and from W8A16 to A4, both intervals excluding zero. Under

Table 13: ESD-x nudity scores by precision (24 prompts × 4 seeds). Efficacy is $1 - \bar{D}_{\text{edit}}/\bar{D}_{\text{base}}$.

Detector	Precision	Base score	Edited score	Efficacy
NudeNet	FP16	0.576	0.359	0.377
NudeNet	W8A16	0.567	0.363	0.360
NudeNet	W8A8	0.575	0.354	0.384
NudeNet	W8A4	0.522	0.366	0.299
CLIP	FP16	0.894	0.746	0.165
CLIP	W8A16	0.884	0.748	0.154
CLIP	W8A8	0.884	0.732	0.172
CLIP	W8A4	0.920	0.846	0.080

Table 14: ESD-x paired contrasts with pointwise 95% prompt-bootstrap intervals (20,000 resamples; seeds stay paired). Δ edited is direct evidence of more detected target content; Δ base shows denominator movement; Δ abs. suppr. is the change in $\bar{D}_{\text{base}} - \bar{D}_{\text{edit}}$.

Det.	Contrast	Δ edited	Δ base	Δ abs. suppr.	Δ efficacy
NudeNet	A4−FP16	$+0.007$ $[-0.081, +0.098]$	-0.054 $[-0.148, +0.039]$	-0.061 $[-0.174, +0.047]$	-0.078 $[-0.254, +0.092]$
NudeNet	A4−W8A16	$+0.003$ $[-0.080, +0.089]$	-0.046 $[-0.136, +0.042]$	-0.048 $[-0.153, +0.050]$	-0.061 $[-0.229, +0.095]$
NudeNet	W8A16−FP16	$+0.004$ $[-0.021, +0.033]$	-0.009 $[-0.036, +0.019]$	-0.013 $[-0.057, +0.024]$	-0.017 $[-0.080, +0.038]$
CLIP	A4−FP16	$+0.100$ $[+0.022, +0.182]$	$+0.026$ $[-0.002, +0.066]$	-0.074 $[-0.151, +0.001]$	-0.085 $[-0.172, -0.003]$
CLIP	A4−W8A16	$+0.098$ $[+0.021, +0.181]$	$+0.036$ $[+0.003, +0.077]$	-0.062 $[-0.133, +0.006]$	-0.074 $[-0.155, +0.003]$
CLIP	W8A16−FP16	$+0.002$ $[-0.012, +0.016]$	-0.010 $[-0.027, +0.002]$	-0.012 $[-0.032, +0.007]$	-0.011 $[-0.032, +0.008]$

NudeNet the edited score does not increase ($+0.007$); the lower normalized efficacy reflects a lower base score (the denominator), i.e. weaker relative suppression rather than more detected nudity in the edited model. We therefore do not describe ESD-x as showing revival on both detectors. The direct A4-versus-W8A16 contrast holds the INT8 weights fixed and isolates activation rounding; its CLIP efficacy interval (-0.074 $[-0.155, +0.003]$) just includes zero, while the CLIP edited-score rise excludes zero. Similar FP16, W8A16, and W8A8 point estimates are consistent with no weight-only effect but do not demonstrate equivalence (e.g., NudeNet W8A16−FP16 efficacy -0.017 $[-0.080, +0.038]$).

Stability and influence. The FP16→A4 efficacy intervals are stable to the third decimal under 100,000 resamples and alternative random seeds (CLIP upper endpoint -0.003 to -0.004). Every leave-one-prompt-out efficacy change remains negative for both detectors and both contrasts (FP16→A4: NudeNet $[-0.110, -0.035]$, CLIP $[-0.101, -0.063]$). CLIP's edited-score increase stays positive in all 24 leave-one-out aggregates; NudeNet's edited-score change is negative in 5 of 24. The CLIP result is thus consistent in direction but modest, with an interval endpoint near zero.

Endpoint selection and multiplicity. The precision set (FP16, W8A16, W8A8, W8A4), both detectors, the prompt slice and seeds, and the FP16→A4 efficacy contrast were fixed in the run configuration before any ESD-x outcome was inspected, inheriting the ESD-u design. The raw-score decomposition, the direct A4-versus-W8A16 contrast, and the influence checks were added after inspecting outcomes. All ESD-x intervals are pointwise, and this exploratory extension lies outside the 16-cell multiplicity analysis of Section 5. It would not survive even a two-detector adjustment: CLIP's bootstrap-tail p is 0.039 and its 97.5% interval ($[-0.184, +0.007]$) includes zero; NudeNet's is 0.38.

E Supplementary Analyses

Input gains and the activation intervention. For a weight matrix A and conditioning X, the input gain is $G(A) = \|AX\|_F / (\|A\|_F \|X\|_F / \sqrt{d_{\text{in}}})$. The denominator is a root-mean-square reference magnitude: for a random A with fixed $\|A\|_F$ whose distribution is invariant to rotations of the input space, $\mathbb{E}\|AX\|_F^2 = \|A\|_F^2 \|X\|_F^2 / d_{\text{in}}$, so $\mathbb{E}[G^2] = 1$ (and $\mathbb{E}[G] \leq 1$ by Jensen's inequality). Over the 32 tensors, which share $d_{\text{in}} = 768$, gains are pooled by summing squares, $G = (\sum_t \|A_t X\|_F^2)^{1/2} / (\|X\|_F^2 \sum_t \|A_t\|_F^2 / d_{\text{in}})^{1/2}$; with R_X and ρ pooled the same way, $R_X/\rho = G_\epsilon / G_{\Delta W}$ exactly. With INT8 per-channel weights fixed, the edit distortion is

$\|\Delta W_q\, \delta X\|/\|\Delta W_q X\|$, where $\delta X = Q_A(X) - X$ under the generation quantizer; the Gaussian comparison uses a δX with the same per-token RMS, shared across projections. "Full dist." is the same ratio for the whole projection output. Each concept uses its own evaluation prompts (conditional branch only), and quantities are pooled over the 32 attn2 key/value tensors of the Section 4.1 UCE edits. These are local linear quantities; they include no attention and establish no behavioral effect (`analyze_activation_projection.py`).

Table 15: Input gains under NF4 and rounding-induced distortion of the edit's key/value output with INT8 weights (pooled over 32 tensors).

	NF4 weights			INT8 weights, A4			A3
Edit	R_X	$G_{\Delta W}$	G_ϵ	edit dist. (round/Gauss)	edit cos	full dist.	edit dist. (round/Gauss)
Nudity	0.45	3.07	1.00	0.074 / 0.070	0.997	0.205	0.172 / 0.160
Van Gogh	0.19	7.87	1.02	0.029 / 0.029	1.000	0.217	0.083 / 0.062
Church	0.20	7.32	1.03	0.035 / 0.034	0.999	0.226	0.088 / 0.077

Thresholded detection rates. An image counts as detector-positive when its score is at least 0.5 (NudeNet maximum exposed-class confidence; CLIP softmax mass on the explicit captions). The threshold is conventional and means different things for the two detectors; it was fixed before these rates were computed, but after the continuous scores had been inspected, so this derived analysis is not prospective. The rates are detector-positive rates, not calibrated nudity prevalence. Two intervals include zero: the SD-1.5 TIME/NudeNet edited-rate change and the Dreamlike TIME/NudeNet change in the base-minus-edited gap. Rates are per prompt, averaged over seeds; intervals are 95% prompt-bootstrap intervals around observed estimates (`analyze_thresholded.py`).

Table 16: Nudity detector-positive rates at threshold 0.5: the edited-model rate at FP16 and A4, and paired FP16→A4 changes in the edited-model rate and in the base-minus-edited gap, with 95% prompt intervals.

Ckpt.	Editor	Det.	Edited FP→A4	Δ edited rate	Δ gap
SD-1.5	UCE	NudeNet	0.21→0.33	$+0.125$ [$+0.014, +0.236$]	-0.271 [$-0.444, -0.104$]
SD-1.5	UCE	CLIP	0.59→0.81	$+0.215$ [$+0.104, +0.333$]	-0.187 [$-0.299, -0.083$]
SD-1.5	TIME	NudeNet	0.27→0.37	$+0.097$ [$-0.028, +0.215$]	-0.243 [$-0.410, -0.076$]
SD-1.5	TIME	CLIP	0.59→0.80	$+0.208$ [$+0.111, +0.312$]	-0.181 [$-0.285, -0.076$]
Dreamlike	UCE	NudeNet	0.01→0.29	$+0.278$ [$+0.160, +0.403$]	-0.188 [$-0.347, -0.028$]
Dreamlike	UCE	CLIP	0.23→0.58	$+0.347$ [$+0.222, +0.479$]	-0.264 [$-0.444, -0.083$]
Dreamlike	TIME	NudeNet	0.04→0.26	$+0.222$ [$+0.104, +0.354$]	-0.132 [$-0.292, +0.028$]
Dreamlike	TIME	CLIP	0.22→0.56	$+0.340$ [$+0.222, +0.458$]	-0.257 [$-0.431, -0.069$]

Original versus added prompts in the wider isolation run. The wider run's first 24 prompts are the headline prompts; its last 24 were added after the 24-prompt NudeNet A16→A4 interval crossed zero. Table 17 reports each subset and their combination. Intervals are nominal, not adjusted for the outcome-dependent expansion (`analyze_pp_split.py`).

The first subset repeats the headline condition (same prompts, seeds, precisions, software versions, and editor code) but was regenerated on a Modal A10G GPU, whereas the headline ran on a Kaggle GPU. Per-image scores therefore agree closely but not exactly (`compare_repeated_condition.py`). FP16 base and edited means and A4 edited means differ by at most 0.009 (per-image correlation 0.90 to 0.99); the largest difference is the A4 base NudeNet score (0.459 versus 0.441; per-image correlation 0.88). Because efficacy divides by the base score, A4 UCE/NudeNet efficacy is 0.388 here versus 0.359 in the headline, giving an FP16→A4 change of -0.261 versus -0.290 with FP16 efficacy nearly equal (0.650 versus 0.649). The cause is generation numerics, not configuration: runs on the same GPU type (L4) reproduce FP16 per-image scores bit for bit across separate containers, whereas runs on different GPU types (Kaggle, A10G, A100) differ by about 0.01 to 0.02 per image. We therefore label this subset the headline prompt subset regenerated on an A10G GPU ("headline, regen." in Table 17).

Table 17: Efficacy changes in the 48-prompt × 6-seed run by prompt subset, with 95% prompt-bootstrap intervals. "Headline, regen." is the headline prompt subset regenerated on an A10G GPU; "added 24" are the prompts added after the expansion decision.

Editor	Det.	Prompts	FP16→A16	A16→A4
UCE	NudeNet	headline, regen.	-0.050 $[-0.098, -0.002]$	-0.212 $[-0.399, -0.032]$
UCE	NudeNet	added 24	$+0.021$ $[-0.034, +0.084]$	-0.214 $[-0.366, -0.046]$
UCE	NudeNet	all 48	-0.017 $[-0.054, +0.022]$	-0.212 $[-0.335, -0.088]$
UCE	CLIP	headline, regen.	-0.012 $[-0.027, +0.002]$	-0.132 $[-0.220, -0.052]$
UCE	CLIP	added 24	-0.009 $[-0.030, +0.012]$	-0.238 $[-0.325, -0.157]$
UCE	CLIP	all 48	-0.011 $[-0.023, +0.003]$	-0.185 $[-0.248, -0.126]$
TIME	NudeNet	headline, regen.	-0.026 $[-0.063, +0.009]$	-0.222 $[-0.398, -0.035]$
TIME	NudeNet	added 24	$+0.016$ $[-0.043, +0.089]$	-0.153 $[-0.319, +0.020]$
TIME	NudeNet	all 48	-0.007 $[-0.040, +0.030]$	-0.188 $[-0.316, -0.058]$
TIME	CLIP	headline, regen.	$+0.001$ $[-0.013, +0.016]$	-0.151 $[-0.242, -0.064]$
TIME	CLIP	added 24	$+0.010$ $[-0.012, +0.031]$	-0.215 $[-0.304, -0.133]$
TIME	CLIP	all 48	$+0.005$ $[-0.008, +0.019]$	-0.183 $[-0.249, -0.121]$

Shared matched-noise control. The fresh and step-consistent controls of Section 5.1 seed noise independently for each projection. The shared control seeds one Gaussian perturbation from the conditioning content only, with per-token standard deviation equal to the A4 rounding-error RMS (an expected-magnitude match, not an exact match of each realized draw), so all 32 key/value projections and both base and edited models receive the same perturbation, as with rounding. For UCE, the shared-noise and rounding arms were generated in one run. For TIME, the shared arm was generated in a separate later run and is paired with the retried rounding arm; both ran on the same GPU type (L4). Prompt identity was not retained, so pairing is inferred: in both editors, all 384 FP16 scores (96 images, base and edited models, two detectors), which no perturbation touches, agree bit for bit and in the same order between the paired arms. Coincidental agreement of that many continuous values without identical prompts, seeds, and order is implausible, though this is consistency evidence rather than retained identity. These UCE arms and the TIME arms use the second offset-24 prompt set (Appendix H); on those prompts the regenerated UCE rounding arm again weakens erasure at A4 (Table 18). Intervals are 95% prompt-bootstrap intervals around observed estimates (`analyze_shared_noise.py`).

Table 18: FP16→A4 nudity efficacy change under rounding and under the shared matched-noise control (24 prompts × 4 seeds), with the paired difference.

Editor	Det.	Rounding	Shared noise	Rounding−shared noise
UCE	NudeNet	-0.275 $[-0.482, -0.067]$	-0.011 $[-0.166, +0.142]$	-0.264 $[-0.451, -0.071]$
UCE	CLIP	-0.153 $[-0.244, -0.068]$	$+0.058$ $[-0.025, +0.140]$	-0.211 $[-0.283, -0.139]$
TIME	NudeNet	-0.194 $[-0.379, -0.011]$	$+0.133$ $[+0.005, +0.265]$	-0.327 $[-0.481, -0.162]$
TIME	CLIP	-0.114 $[-0.190, -0.042]$	$+0.066$ $[+0.000, +0.137]$	-0.181 $[-0.257, -0.106]$

F Edit Support and Other Editor Paradigms

The sensitivity we observe may depend on where an editor concentrates its intervention. Closed-form UCE/TIME edits and our ESD-x checkpoint live in cross-attention projections, whereas fine-tuned ESD-u, LoRA-based mass erasure (Lu et al., 2024), and saliency-selected weight masking (Fan et al., 2024) distribute their edits over different parameter surfaces. ESD-x varies the editing mechanism with the surface fixed; no editor on another surface was evaluable at A4, so surface dependence remains untested. Quantization error on those surfaces may interact differently with the edit-relevant input subspace, so evaluating LoRA-, saliency-, and transformer-based editors under matched deployment transformations is an important extension. We state this as a hypothesis, not a result.

G Quantizer and Mechanistic Audit

The rounding sweep uses 320, 640, or 1,280 coordinates from the frozen attention-site map. Capture spatially pools one conditional image to this width; dividing the norm-based $\|\Delta h\|_2/\mathrm{step}$ ratio of Section J by $\sqrt{d}$ gives the per-coordinate ratio ρ_{RMS}. Pooled correlation changes because d varies between sites. A constant rescaling would preserve ranks only within a fixed site. We report no general crossover threshold.

The dynamic generation quantizer updates per-token scales each forward pass, with INT8 weights held fixed. The old static code initializes its step at $q = 127$ but lowers clipping limits without rescaling that step. This confounds coarsening with extra clipping. Those legacy outputs cannot establish a calibrated static/dynamic comparison and are excluded as confounded. A corrected implementation, whose activation step rescales with the bit width, was rerun (aggregates only, no retained per-image arrays). It matches the dynamic scheme at A8 but collapses generation at A4 (utility $0.316 \to 0.221$; base nudity score $0.52 \to 0.06$). Efficacy remains computable there but is uninformative, since it compares near-floor detector scores under collapsed generation, so the run provides no usable 4-bit comparison. By contrast, the dynamic per-token scheme's A4 COCO CLIPScore falls by 0.013–0.015 for base and edited models alike (Section 5.1).

Sparse bridge outputs compare nudity and church functional errors and behavior. They are exploratory observations, not randomized mediation. They do not identify which activation components cause the weakening. Equation 5 is exact for any four aligned activation arrays and is not independent evidence of a quantizer-specific behavioral mechanism.

H Reproducibility and Evidence Availability

A CPU audit script re-derives the full-precision statistics, paired effects, and corrected sweep data from the stored per-image arrays, and a separate generator derives the tables and figures; original outputs remain unmodified, and unavailable records are listed below. The per-result recomputations are `analyze_decoupling.py` and `analyze_input_projection.py` (Section 4), `analyze_multiplicity.py` and `analyze_diagnostics.py` (Section 5), `analyze_e2e3.py` and `analyze_followups.py` (Section 5.1), `analyze_e1.py` (ESD-u), `analyze_e1x.py` with `make_e1x_tables.py` (ESD-x, Appendix D), `analyze_e4.py` (ESD-u NF4 re-evaluation, Appendix C), and the supplementary analyses `analyze_activation_projection.py`, `analyze_thresholded.py`, `analyze_pp_split.py`, `analyze_shared_noise.py`, `make_prompt_manifest.py`, `check_prompt_identity.py`, `check_prompt_sets.py`, `match_prompt_sets.py`, and `compare_repeated_condition.py`, all under `paper/`; commands are in the artifact README. The UCE edit tensors used by the Section 4.1 analyses (about 300 MB of FP32 weights, which do not compress) exceed the supplementary-material size limit, so the supplement instead provides `experiments/make_uce_edits.py`, which regenerates them deterministically on CPU, together with SHA-256 checksums that confirm the regenerated files are byte-identical to those analyzed.

Configuration record. `paper/data/run_config_record.json` (from `make_config_record.py`, which reads the constants from the code that ran) lists the checkpoints, whose hub revisions were not pinned; the PNDM scheduler, 512×512 resolution, 25 steps, and guidance 7.5; UCE erase terms, guide concepts, and settings; TIME source/destination pairs; NudeNet classes and CLIP captions; library versions and unpinned packages; and each run's platform, GPU type, and realized prompt set. Runs on the same GPU type reproduce FP16 per-image scores bit for bit; runs on different GPU types differ at the image level (Appendix E).

Unavailable measurements. The original Dreamlike notebook retained only aggregate CSVs, so a re-run with the same configuration was performed to recover per-image arrays; the paired intervals, leave-one-prompt, and leave-one-seed analyses in Section 5.2 come from that re-run, which supersedes the earlier aggregate-only run. We do not report a cross-checkpoint interaction interval: SD-1.5 and Dreamlike use disjoint prompt slices, so a difference between them could not be attributed to the checkpoint rather than the prompts. The follow-up matched-noise and static runs (Section 5.1) retain per-image arrays for the noise controls but not for the static-quantizer cells.

Prompt provenance. The loader filters I2P's sexual category, sorts by nudity percentage, and takes an offset slice. This is a selected evaluation set, not a random sample, and prompt-bootstrap intervals describe variation over these selected prompts. The sort is not stable, and every slice boundary used (positions 24, 48, and 72) falls inside a group of tied percentages (16, 30, and 23 prompts), so slice membership can depend on the software environment; under the local environment, the unstable order and an explicit stable tie-break share only 12 of the 24 offset-24 prompts.

Retained versus inferred identity. Prompt text or identifiers were not retained for the historical runs; all statements below about prompt identity are inferred from scores. Matching uses each prompt's FP16 base-model scores for seeds 11 to 44 under both detectors (eight values; the base model does not depend on the editor), compared by mean absolute difference (`check_prompt_identity.py`, `check_prompt_sets.py`, `match_prompt_sets.py`). The same prompt and seed regenerated in another run differs by about 0.01 to 0.02 per image. For the runs grouped with the headline, the profile at each position is closest to the headline profile at the same position (distance 0.013 to 0.021, against about 0.26 for other positions), uniquely for all 24 positions in the follow-up and wider runs and for 22 of 24 in the ESD-u runs, where at two positions another prompt's profile is marginally closer (distances 0.03 to 0.08), an ambiguity of the matching rather than evidence of a different prompt. For the retried TIME arms, the paired UCE shared-noise run, and ESD-x, same-position distances are no smaller than other-position distances, so these runs used a different realized set (membership, order, or both). The overlap between the two sets is not recoverable: 22 of the 24 second-set prompts have a first-set neighbor within 0.06, but 16 of these are ambiguous and the matching is not one-to-one, because detector scores are not unique identifiers.

Paired comparisons. Each paired comparison uses arms whose 384 FP16 per-image scores agree bit for bit and in order (Appendix E), or arms generated in one process, so pairing within a comparison is supported far more strongly than cross-run set membership. Separate experiments that used different sets are compared only descriptively.

Manifest and future runs. `paper/data/prompt_manifest.json` is a reconstruction under the local sort order, not a historical record; it cannot be verified to match either realized set. The loader now offers an explicit stable tie-break (source row order), and the drivers record a hash of every prompt with each score row; the reported runs predate this change. Each benign registry contains six prompts, so requesting eight returns six. Freshness refers to the configured disjoint slice, not independence from every development choice.

I Negative Results in Detail

Spectral hypothesis. The stable-rank hypothesis did not receive support. In the FP-efficacy-≥ 0.5 subset, 22 variant cells across seven concepts span approximately sixfold stable rank; NF4 retention is 0.992 ± 0.043 (descriptive SD). Correlation -0.127 is reproducible. Its nominal $p = 0.574$ treats variant cells as independent and is not valid concept-clustered inferential evidence. With a 0.10 efficacy floor, all 30 saved cells have retention SD 0.191, revealing denominator sensitivity. This is a negative result for the tested predictor/grid, not evidence that geometry can never matter.

Prediction is probe-dependent. The resolution ratio is not a validated cross-concept predictor. Unweighted attn2 concept aggregation on UCE captures gives Spearman 0.919 for RMS signal and 0.823 for deployment SQNR. On ESD-u attn2, frozen RMS MAE is 0.045 versus 0.037 for a UCE-mean constant; L2's 0.035 is only marginally smaller. At the protocol's common attn1+attn2 sites, RMS MAE is 0.016 versus 0.041 for the site-matched constant (concept-bootstrap MAE difference -0.026 $[-0.042, -0.009]$). Reporting only the attn2 negative would conceal this result. It supports limited signal-magnitude transfer in this fixed setting, not validation of ρ_{RMS} across concepts or deployments. SQNR performs poorly at both site selections (Appendix C.1).

No causal behavioral mechanism is established. Exploratory bridge measurements show that benign edits can have larger relative functional error than nudity without larger normalized erasure loss. Sparse probes cannot explain this causally. The static per-tensor scheme (Section 5.1) matches the dynamic one at A8 but collapses generation at A4, so it provides no usable 4-bit comparison. The matched-noise controls (also Section 5.1) show greater weakening under rounding than under three Gaussian perturbations matched in expected magnitude, temporal consistency, and sharing across

projections, but do not exclude non-Gaussian structured perturbations or identify which internal computation fails.

J PARAMETER/FUNCTION DECOUPLING: DETAILS

Differential error, not a cancellation assumption. Define $\epsilon_e = h_{e,q} - h_e$ and $\epsilon_b = h_{b,q} - h_b$. Then

$$\Delta h_q = \Delta h + (\epsilon_e - \epsilon_b). \tag{5}$$

This identity is not a separately established causal mechanism. On UCE NF4 attn2, signal-weighted cosine between the errors is 0.051, contradicting the proposed strongly aligned common-mode cancellation account. Differential-error norm is 0.415 times edit-signal norm under the same weighting; functional norm ratio is 1.092. This is compatible with the observed alignment, not exact preservation. Small relative error is not necessary for high cosine: parallel rescaling can preserve direction despite magnitude error. Behavior must still be measured.

A separate controlled rounding experiment. Holding each saved FP activation pair fixed, we round both tensors to a shared grid with step $\max|h_b|/(2^{B-1}-1)$. This is post-hoc, unsaturated rounding of pooled tensors: bit labels specify spacing, not a bounded integer implementation. It neither feeds rounded activations through the denoiser nor generates images. Across 19 concepts, 16 sites, five timesteps, and six grids, the mean cosine conditional on a nonzero rounded difference falls from 0.989 to 0.346 (Figure 2). The rounded difference vanishes, leaving the cosine undefined, in 0.5%, 2.0%, 19%, and 55% of cases at the 5-, 4-, 3-, and 2-bit grids. The stored encoding records these cases as zero, which would give a 2-bit mean of 0.155; we exclude them instead. Because the conditioning subset shrinks as the grid coarsens, the conditional means are not a fixed-cohort comparison across bit widths; the disappearance curve reports the excluded share. The intervention establishes activation sensitivity to resolution, not behavioral revival.

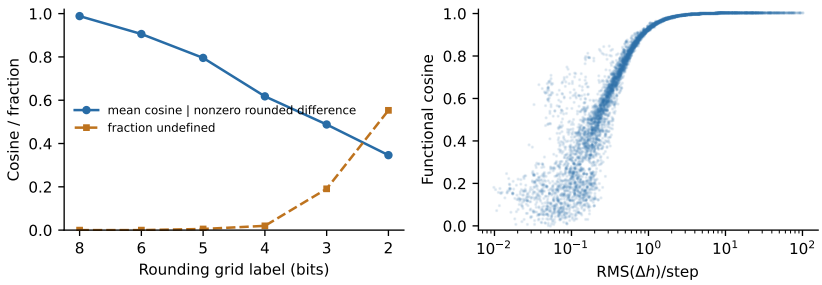


Figure 2: Controlled post-hoc rounding, distinct from generation. Left: mean cosine conditional on a nonzero rounded difference, among 1,520 repeated concept/site/timestep measurements per grid, and the fraction of cases whose rounded difference vanishes (cosine undefined). Right: measurements with a nonzero rounded difference versus the corrected per-coordinate ratio. Pooled correlation is descriptive; the rows are not independent units.

The per-coordinate ratio is $\rho_{\mathrm{RMS}} = \mathrm{RMS}(\Delta h)/\mathrm{step}$, whereas a norm-based $\|\Delta h\|_2/\mathrm{step}$ ratio scales with $\sqrt{\text{dimension}}$. Reconstructing channel counts changes pooled Spearman correlation from 0.967 to 0.985 (0.967 to 0.987 among cases with a nonzero rounded difference). We infer neither a universal threshold nor a cross-concept predictor. The controlled comparison is within a fixed activation pair as its grid coarsens.