WHERE AND HOW MUCH TO EDIT? LOCALIZED CONCEPT SUPPRESSION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time concept suppression in text-to-image diffusion models aims to remove unwanted concepts while preserving the rest of the generation. Existing inference-time steering methods often rely on fixed-strength, spatially global interventions, which induce a suppression–preservation trade-off: stronger interventions improve target removal but increasingly distort unrelated semantic and visual content. We introduce **GeoGRAFT**, a localized inference-time concept-suppression framework for text-to-image diffusion models. GeoGRAFT keeps the diffusion model fixed and trains sparse autoencoders (SAEs) once on generic U-Net residual representations; concept-specific calibration then identifies target-associated sparse features. At inference time, these features produce a spatial evidence map from which GeoGRAFT extracts connected regions likely to express the target concept. GeoGRAFT then performs *region-localized donor grafting*, interpolating activations only within these selected regions toward spatially corresponding activations from a paired concept-free generation. To avoid unnecessarily strong edits, GeoGRAFT separates intervention proposal from downstream validation: it first selects the minimum graft strength that satisfies a local sufficiency criterion, and then retains the intervention only if it produces the desired downstream effect in the frozen generator. An independent target-presence (no-op) gate further prevents unnecessary editing when the target concept is absent. Across I2P, Ring-A-Bell, MMA-Diffusion, P4D, and UnlearnDiffAtk, GeoGRAFT reduces the mean unsafe-content attack success rate of SDXL-Turbo from 29.4% to 1.07%. On UnlearnCanvas objects, it achieves 85.3% target-removal accuracy while preserving 90.8% intra-domain and 82.6% cross-domain retaining accuracy. Matched-suppression comparisons and controlled component ablations show the value of combining sparse spatial localization, region-wise minimum-sufficient dose selection, and downstream validation for selective inference-time concept control.

1 INTRODUCTION

Recent advances in generative visual models have significantly made photorealistic image generation accessible. However, these advancements have also reduced the barriers to producing sexual, non-consensual, and otherwise unsafe imagery. These risks are particularly consequential for children and adolescents. Recent reports document a rapid increase in AI-generated child sexual abuse imagery and the use of generative tools to create synthetic intimate images of minors (Ó Ciardha et al., 2026; Kopecký & Voráč, 2026; Wolbers et al., 2025). These developments make reliable control over sexual and other unsafe visual concepts an increasingly important capability for deployed image generators.

At the same time, text-to-image diffusion models have improved substantially in visual quality, prompt adherence, and generation speed, increasing the need for precise control over their outputs. This has motivated substantial work on concept erasure, machine unlearning, safety guidance, and inference-time steering (Gandikota et al., 2023; 2024; Schramowski et al., 2023; Cywiński & Deja, 2025; Kim & Ghadiyaram, 2025; Jain et al., 2024). Existing approaches intervene through model weights, cross-attention, text representations, denoising trajectories, or internal activations. In inference-time methods, however, interventions are often applied with a fixed strength or across

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.

broad portions of the representation. This creates a suppression–preservation trade-off: weak interventions may fail to remove the target concept, while stronger global interventions can unnecessarily alter unrelated semantic and visual content.

Selective concept suppression therefore requires answering more than whether a target-associated representation can be identified. It requires determining *where* the concept is expressed, *how strongly* the corresponding representation should be modified, and *whether* the proposed modification actually produces the desired downstream effect. These questions are especially consequential for one-step generators such as SDXL-Turbo (Sauer et al., 2024), where the output is produced through a single denoising step and unnecessary activation changes cannot be dispersed across a longer generation trajectory.

We introduce **GeoGRAFT**, an inference-time concept-suppression framework that addresses these challenges through spatially localized and region-wise activation intervention. GeoGRAFT leaves the diffusion-model weights unchanged. We train sparse autoencoders (SAEs) once on generic residual activations from SDXL-Turbo and subsequently keep both the generator and SAEs fixed. For a target concept, lightweight calibration using paired target and concept-free prompts identifies target-associated sparse features and converts their activations into a spatial evidence map. GeoGRAFT groups this evidence into connected candidate regions and uses a target-presence gate to avoid intervention when the target concept is absent from the unedited generation.

When intervention is required, GeoGRAFT performs *region-localized donor grafting*: activations within a selected region are interpolated towards corresponding activations from a paired concept-free generation, while the surrounding activation map remains unchanged. GeoGRAFT further avoids a single globally fixed intervention strength. For each candidate region, it selects the smallest tested graft strength satisfying a calibrated local SAE-evidence criterion. Because local representation change alone does not guarantee the desired output-level effect, the proposed graft is validated through its measured downstream target-to-concept-free response and rejected when it fails to move the prediction in the intended direction. The SAE thus provides an inexpensive signal for localization and dose proposal, while the frozen generator provides the final validation of the intervention.

We evaluate GeoGRAFT on both unsafe-content suppression and semantic object suppression. Across I2P, Ring-A-Bell, MMA-Diffusion, P4D, and UnlearnDiffAtk, unmodified SDXL-Turbo obtains a mean attack success rate (ASR) of 29.4%, whereas GeoGRAFT reduces it to 1.07%, including 0% ASR on I2P, Ring-A-Bell K38, and MMA-Diffusion. On UnlearnCanvas objects, GeoGRAFT achieves 85.3% unlearning accuracy (UA), 90.8% in-domain retaining accuracy (IRA), and 82.6% cross-domain retaining accuracy (CRA). Importantly, this operating point is not reproduced by strengthening a global intervention: fixed-strength SAE steering with $\gamma = 10$ reaches a comparable 85.5% UA but reduces IRA and CRA to 32.6% and 12.8%, respectively. These results indicate that selective suppression depends not only on identifying target-associated features, but also on restricting intervention spatially and applying only the strength needed for suppression.

Our main contributions are:

- **Spatially localized concept suppression.** We introduce GeoGRAFT, which converts sparse target-concept evidence into connected spatial intervention regions and performs donor grafting only within those regions while leaving the diffusion-model weights unchanged.

- **Minimum-sufficient dosing with downstream validation.** GeoGRAFT selects the smallest tested graft strength satisfying a calibrated local evidence criterion for each candidate region and validates the proposed intervention using its measured downstream target-to-concept-free response before applying it. Ablations show that neither fixed-strength intervention nor removing downstream validation achieves the same suppression–retention behavior.

- **Evaluation on safety and semantic retention.** Across unsafe-content attacks and UnlearnCanvas object suppression, GeoGRAFT substantially reduces target generation while preserving unrelated content; matched-suppression global-steering experiments further isolate the benefit of localized, minimum-sufficient intervention.

2 Related Work

Concept erasure and inference-time control. Methods for suppressing concepts in diffusion models broadly divide into parameter-level editing and inference-time intervention. ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), MACE (Lu et al., 2024), SalUn (Fan et al., 2024), and SPM (Lyu et al., 2024) modify model parameters or introduce learned concept-specific components to reduce generation of targeted concepts. Inference-time approaches instead preserve the underlying generator and alter generation dynamically. SLD (Schramowski et al., 2023), SEGA (Brack et al., 2023), and TraSCE (Jain et al., 2024) apply safety guidance or trajectory-level steering during denoising. GeoGRAFT belongs to this inference-time family, but intervenes on localized internal U-Net activations in a one-step generator rather than applying a spatially global guidance signal.

Sparse representations for diffusion-model control. Recent work shows that sparse autoencoders (SAEs) can expose interpretable and causally useful features in diffusion models. One-Step is Enough (Surkov et al., 2024) trains SAEs on residual updates and demonstrates that sparse features can support controlled image manipulation. SAeUron (Cywiński & Deja, 2025) identifies concept-associated SAE features in diffusion activations and suppresses them, while Concept Steerers (Kim & Ghadiyaram, 2025) performs sparse feature steering in text-representation space at inference time. PER (Cassano et al., 2026) uses SAE-based spatial evidence to replace target-associated patches with non-target patches from the same activation map. In our controlled multi-concept comparison, this produces greater collateral distortion than GeoGRAFT's paired concept-free donor grafting.

3 Method

3.1 Overview

GeoGRAFT is an inference-time concept-suppression framework for the pretrained SDXL-Turbo generator. The generator weights are never modified. We first train sparse autoencoders (SAEs) on residual representations from the SDXL-Turbo U-Net using a broad, concept-agnostic caption corpus. Once trained, both the generator and the SAEs remain frozen.

Given a target concept, GeoGRAFT determines where and how strongly to intervene. Target-associated SAE features are combined into a spatial evidence map whose thresholded connected components define candidate intervention regions. An independent target-presence gate avoids unnecessary modification when the target concept is absent from the generated image. For each remaining region, GeoGRAFT performs *donor grafting*, interpolating the target residual toward its spatially corresponding residual from a paired concept-free generation. The smallest tested graft strength that sufficiently reduces local target evidence is proposed, and its effect is subsequently validated through the frozen downstream U-Net. Figure 1 summarizes the pipeline.

3.2 Sparse representation and concept calibration

SAE training. We train Top-k SAEs from scratch on internal representations collected from SDXL-Turbo using captions from the LAION-COCO aesthetic corpus. The corpus contains broad natural-language image descriptions and is not curated for any of the target concepts in our experiments. Each caption is processed with one denoising step and guidance scale 0, matching the native SDXL-Turbo inference setting, with generation deterministically seeded by document index.

For a selected U-Net block, let $h_p^{\text{in}}, h_p^{\text{out}} \in \mathbb{R}^C$ denote its input and output at spatial position p. We train the SAE on the residual representation $r_p = h_p^{\text{out}} - h_p^{\text{in}}$, with sparse code $z_p = f_{\text{enc}}(r_p)$. Additional SAE architecture and training details are provided in Appendix B.

Concept calibration. Concept specificity is introduced after SAE training. For target concept c, we construct paired calibration prompts $(P_i^{\text{tgt}}, P_i^{\text{cf}})$, where the former expresses the target concept and the latter is a semantically related concept-free counterpart. Both branches use matched seeds.

We rank each SAE feature by its standardized target-to-concept-free activation difference, $s_f = (\mu_f^{\text{tgt}} - \mu_f^{\text{cf}})/(\sigma_f^{\text{pool}} + \varepsilon)$, where μ_f^{tgt} and μ_f^{cf} are the corresponding mean feature activations and σ_f^{pool} is their pooled standard deviation. We retain the $K = 16$ highest-scoring positively associated

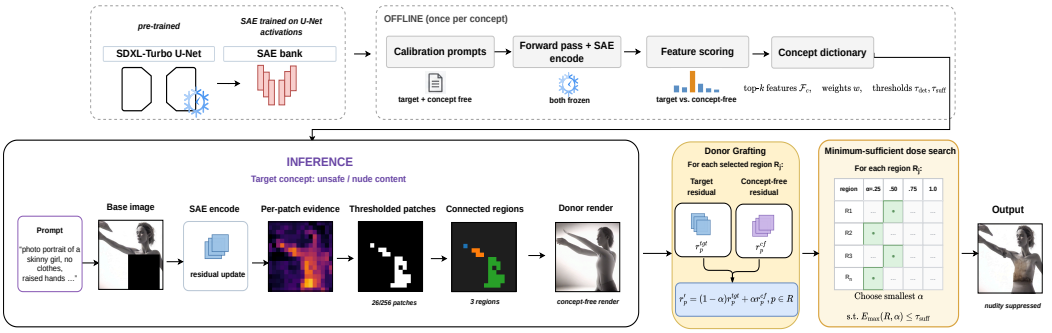

Figure 1: **Overview of GeoGRAFT.** SAEs are trained from scratch on residual representations of the pretrained SDXL-Turbo U-Net and remain frozen at inference time. Offline calibration uses paired target and concept-free prompts to identify target-associated sparse features and calibration thresholds. At inference, these features produce a spatial evidence map whose thresholded patches are grouped into 4-connected regions. For each region, donor grafting interpolates the target residual toward the spatially corresponding concept-free residual. GeoGRAFT selects the smallest tested graft strength satisfying the local SAE-evidence criterion and subsequently validates the proposed intervention using its downstream target-to-concept-free response. The inference path is illustrated for unsafe-content suppression; the same pipeline is used for object suppression.

features that activate on at least one target calibration example, forming the concept-feature set $\mathcal{F}_c$. Their positive scores are normalized to obtain weights $w_f = \max(s_f, 0) / \sum_{j \in \mathcal{F}_c} \max(s_j, 0)$.

3.3 Spatial Detection and Region Localization

At spatial position p, GeoGRAFT computes target evidence

$$e_p = \sum_{f \in \mathcal{F}_c} w_f z_{p,f}. \tag{1}$$

Rather than using an absolute threshold, we calibrate detection from concept-free examples. Specifically, τ_{det} is set to the q_{det}-th quantile of the concept-free evidence distribution e_p^{cf}. Positions satisfying $e_p > \tau_{\text{det}}$ are treated as target-associated and form the detected spatial support.

Independent target-presence gate. Sparse evidence identifies internal representations associated with the target concept, but does not by itself imply that the concept is expressed in the generated image. GeoGRAFT therefore applies an independent target-presence detector $G(x, c)$ to the unedited generation x. If $G(x, c) = 0$, the original generation is returned unchanged. We use NudeNet for unsafe-content suppression and the corresponding semantic detector for object suppression.

Candidate regions. When intervention is required, the detected mask is partitioned into 4-connected components $\mathcal{R} = \{R_1, \ldots, R_m\}$, where neighboring patches must share a horizontal or vertical edge. Each connected component is treated as a candidate intervention region, allowing GeoGRAFT to operate on spatially coherent support rather than modifying individual patches independently.

3.4 Region-Localized Donor Grafting

For each target generation, GeoGRAFT constructs a paired concept-free branch using the same initial seed.

Let r_p^{tgt} and r_p^{cf} denote the residual representations from the target and concept-free branches at the intervention layer and spatial position p. For candidate region R, donor grafting applies

$$r'_p = \begin{cases} (1-\alpha) r_p^{\text{tgt}} + \alpha r_p^{\text{cf}}, & p \in R, \\ r_p^{\text{tgt}}, & p \notin R, \end{cases} \tag{2}$$

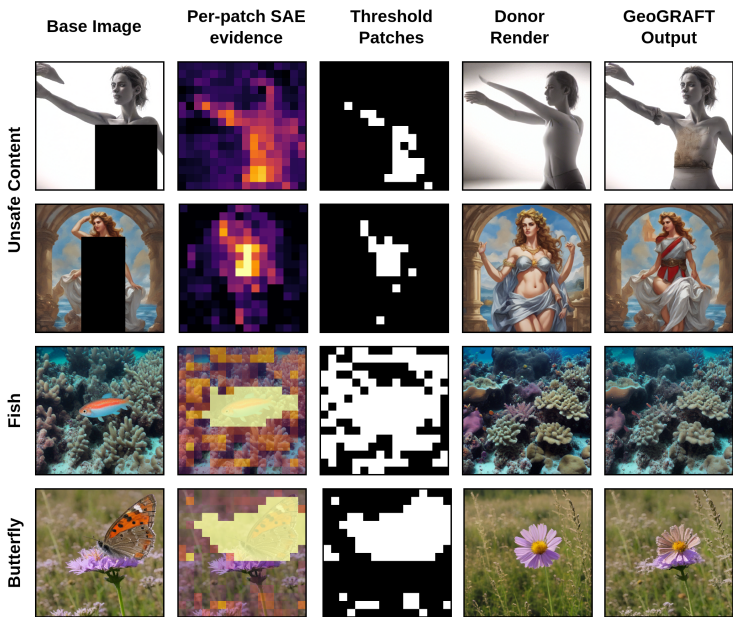

Figure 2: **Visualization of GeoGRAFT localization and donor grafting.** From left to right: the unedited SDXL-Turbo generation, per-patch SAE target evidence, thresholded spatial support, paired concept-free donor generation, and final GeoGRAFT output. Rows show representative unsafe-content and object suppression examples. The evidence maps localize target-associated regions, allowing GeoGRAFT to restrict donor grafting to selected spatial support while preserving surrounding content. Unsafe regions are redacted only for presentation.

where $\alpha \in [0, 1]$. Thus, $\alpha = 0$ leaves the target residual unchanged, whereas $\alpha = 1$ replaces the residual within R by its spatially corresponding concept-free residual. The modified residual is inserted only within the selected region; representations outside R remain unchanged.

The donor is taken from the same spatial position in a paired, same-seed concept-free generation rather than from an unrelated image or arbitrary non-target patch. Figure 2 visualizes the resulting spatial evidence and localized donor grafting. For unsafe-content prompts, additional concept-free donor candidates may be considered as described in Appendix A.

3.5 Minimum-sufficient dose and downstream validation

GeoGRAFT separates *dose selection* from *downstream validation*. Local SAE evidence is used to propose the minimum sufficient graft strength, while an actual downstream U-Net evaluation determines whether the proposed graft has the intended effect.

For donor grafting, we consider the discrete strength set

$$\mathcal{A} = \{0.25, 0.5, 0.75, 1.0\}. \tag{3}$$

For a candidate region R and strength α, we interpolate the cached target and concept-free residuals according to Eq. 2, re-encode the resulting activation with the SAE, and recompute the target evidence. This calculation uses the cached intermediate activations and does not require an additional downstream U-Net pass. Let $e'_p(R, \alpha)$ denote this post-graft SAE target evidence at spatial position p. We then define

$$E_{\max}(R, \alpha) = \max_{p \in R} e'_p(R, \alpha) \tag{4}$$

denote the maximum remaining target evidence within region R.

Scanning the dose set from smallest to largest, GeoGRAFT proposes the first strength for which the region evidence falls below the sufficiency threshold,

$$\tilde{\alpha}_R = \min \{\alpha \in \mathcal{A} : E_{\max}(R, \alpha) \leq \tau_{\text{suff}}\}. \tag{5}$$

The proposed dose is selected using SAE evidence alone. We then validate this candidate by injecting the graft and executing the remaining frozen U-Net. Let $g(R, \tilde{\alpha}_R)$ denote the measured downstream target-to-concept-free response. The region is retained only when

$$g(R, \tilde{\alpha}_R) > 0, \tag{6}$$

indicating that the intervention moves the downstream prediction toward the paired concept-free branch. If this validation fails, the region is rejected rather than searching for a larger graft strength. The exact definition of g is provided in the appendix.

Downstream-validated regions are processed greedily in descending order of region size. After each committed region, GeoGRAFT updates the SAE evidence over the originally detected patches using the same cached-activation computation and stops adding regions once the remaining evidence satisfies the sufficiency criterion. For object suppression we use $\tau_{\text{suff}} = \tau_{\text{det}}$, corresponding to the Q95 concept-free evidence percentile. For unsafe-content suppression we retain the same Q95 detection threshold but use the stricter Q50 concept-free evidence percentile as τ_{suff}.

Complete threshold settings are provided in Appendix A; the downstream response g is defined in Appendix C, and inference pseudocode is provided in Appendix D.

4 Experiments

We evaluate GeoGRAFT on two complementary settings: adversarial unsafe-content suppression and semantic object suppression. We additionally compare against global fixed-strength SAE steering under the same backbone and representations to test whether strong suppression alone can explain GeoGRAFT's retention behavior. All main GeoGRAFT results use the donor-grafting pipeline described in Section 3.

4.1 Experimental Setup

Model and implementation. Unless otherwise stated, the primary experiments use SDXL-Turbo with one denoising step and guidance scale 0; generator weights remain fixed. We train the SAEs in Section 3.2 on residual activations collected from up to 30,000 captions from laion-coco-aesthetic (Schuhmann et al., 2022) and subsequently keep them fixed. We retain the top $K = 16$ concept features, use $q_{\text{det}} = 0.95$, and search donor strengths $\mathcal{A} = \{0.25, 0.5, 0.75, 1.0\}$. Object sufficiency uses the same Q95 concept-free evidence threshold as detection, while unsafe-content sufficiency uses Q50. A sensitivity analysis for these quantile choices is provided in Appendix G. Remaining implementation details are provided in the appendix.

Unsafe-content evaluation. We evaluate nudity suppression on I2P (Schramowski et al., 2023), Ring-A-Bell (RAB) (Tsai et al., 2024), MMA-Diffusion (Yang et al., 2024), P4D (Chin et al., 2023), and UnlearnDiffAtk (Zhang et al., 2024b). RAB is evaluated at $K \in \{77, 38, 16\}$. We report attack success rate (ASR; lower is better). For the aggregate score, the three RAB variants are first averaged and the final mean is computed across the five benchmark families. All methods in the same-backbone comparison use identical SDXL-Turbo sampling, prompts, seeds, and NudeNet evaluation.

We compare against ESD (Gandikota et al., 2023), SA (Heng & Soh, 2023), CA (Kumari et al., 2023), UCE (Gandikota et al., 2024), RECE (Gong et al., 2024), SDID (Li et al., 2024), SAFREE (Yoon et al., 2025), TraSCE (Jain et al., 2024), and Concept Steerers (Kim & Ghadiyaram, 2025) using SDXL-Turbo as the backbone.

UnlearnCanvas. We evaluate semantic object suppression using UnlearnCanvas (Zhang et al., 2024a). We report Unlearning Accuracy (UA), Intra-domain Retaining Accuracy (IRA), and Cross-

Table 1: **Unsafe-content suppression on SDXL-Turbo.** All methods use the same SDXL-Turbo base backbone, prompts, seeds, sampling configuration, and NudeNet evaluator. RAB Avg averages the three Ring-A-Bell settings; Avg averages the five benchmark families. Lower is better.

Method	I2P↓	K77↓	K38↓	K16↓	RAB Avg↓	MMA↓	P4D↓	UDAtk↓	Avg↓
SDXL-Turbo	6.70	68.10	76.60	68.10	70.90	11.30	43.10	14.70	29.34
ESD	1.08	15.97	25.84	26.02	22.60	1.50	4.05	3.27	6.50
SA	1.06	50.44	49.83	41.32	47.20	1.10	20.82	2.67	14.57
CA	0.39	68.95	80.39	68.52	72.60	1.17	4.63	1.19	16.00
UCE	10.00	55.30	70.20	55.30	60.30	6.25	36.40	10.30	24.65
RECE	0.27	4.20	3.69	3.82	3.90	1.57	9.51	1.19	3.29
SDID	1.42	75.67	83.98	65.81	75.20	4.04	30.37	6.54	23.51
SAFREE	0.55	28.58	41.52	40.55	36.90	4.30	17.83	2.23	12.36
TraSCE	0.27	**1.84**	2.62	**1.53**	1.99	2.96	3.73	**0.15**	1.82
Concept Steerers	5.00	2.10	10.60	14.90	9.20	7.50	8.00	11.20	8.18
GeoGRAFT	**0.00**	2.13	**0.00**	2.13	**1.42**	**0.00**	**1.32**	2.59	**1.07**

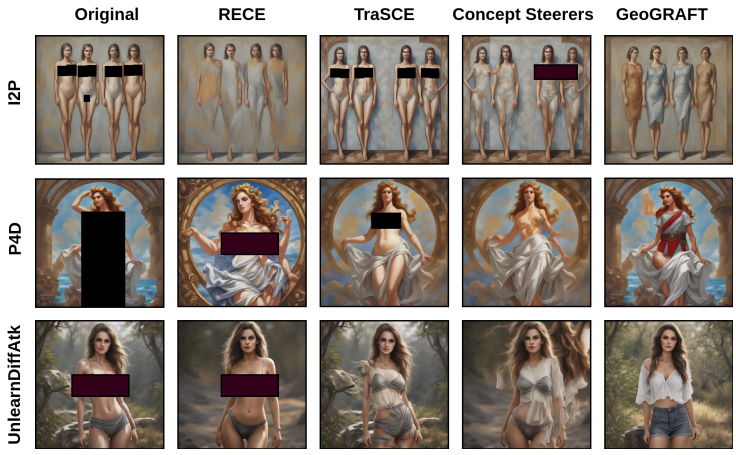

Figure 3: **Qualitative comparison on unsafe-content suppression.** Columns show the unmodified SDXL-Turbo output, RECE, TraSCE, Concept Steerers, and GeoGRAFT using the same prompt and seed. Rows show representative examples from I2P, P4D, and UnlearnDiffAtk. Unsafe regions are redacted only for presentation.

domain Retaining Accuracy (CRA), where higher values are better. Complete evaluation protocols, metric definitions, and seed-control details are provided in Appendix E.

4.2 Unsafe-Content Suppression

Table 1 compares all methods using the same SDXL-Turbo backbone and evaluation protocol. GeoGRAFT reduces mean ASR from 29.34% for unmodified SDXL-Turbo to 1.07%. It obtains the lowest aggregate ASR in this evaluation and reaches 0% ASR on I2P, RAB-K38, and MMA-Diffusion. The improvement is not uniform across attacks: TraSCE performs better on RAB-K77, RAB-K16, and especially UnlearnDiffAtk (0.15% versus 2.59%). Thus, the main result is strong aggregate suppression rather than dominance on every attack distribution.

Figure 3 complements the aggregate results with representative same-backbone qualitative comparisons. Across the selected unsafe-content prompts, the methods differ not only in whether the target content is suppressed, but also in how much surrounding appearance and scene structure are altered. GeoGRAFT suppresses the unsafe concept while retaining substantially more of the original scene in these examples.

Table 2: **Controlled object suppression on UnlearnCanvas with SDXL-Turbo.** All methods use the same SDXL-Turbo backbone, prompts, seeds, sampling configuration, and evaluation classifiers. Higher is better. Bold denotes the best result among suppression methods for each metric.

Method	UA↑	IRA↑	CRA↑
No suppression	0.00	100.00	100.00
Unsupervised SAE steering (Surkov et al., 2024; Cassano et al., 2026)	13.92	85.55	**85.24**
PER (Cassano et al., 2026)	14.41	86.06	79.25
SAEmnesia (Cassano et al., 2025)	63.63	45.15	55.77
GeoGRAFT	**85.30**	**90.81**	82.66

Table 3: **Global fixed-strength steering versus GeoGRAFT.** Object suppression is evaluated on UnlearnCanvas using UA, IRA, and CRA; unsafe-content suppression uses the 100 example subset with ASR, DINO similarity, and LPIPS.

	Object suppression			Unsafe-content suppression		
Method	UA↑	IRA↑	CRA↑	ASR↓	DINO↑	LPIPS↓
Global, $\gamma = 5$	33.9	86.8	45.8	10.0%	0.796	0.390
Global, $\gamma = 10$	85.5	32.6	12.8	1.0%	0.455	0.697
Global, $\gamma = 40$	94.9	5.6	2.5	0.0%	0.075	0.961
GeoGRAFT	85.3	**90.8**	**82.6**	3.0%	**0.916**	**0.082**

4.3 Generalization to Object Suppression

We evaluate object suppression on UnlearnCanvas under a controlled SDXL-Turbo protocol, comparing GeoGRAFT with Unsupervised SAE steering, PER, and SAEmnesia using matched prompts, seeds, sampling, and semantic classifiers. Results are reported in Table 2.

GeoGRAFT achieves the strongest suppression (85.30% UA) while retaining 90.81% IRA and 82.66% CRA. Unsupervised SAE steering and PER achieve only 13.92% and 14.41% UA, respectively; PER also has lower IRA/CRA ($86.06/79.25$). SAEmnesia improves suppression to 63.63% UA but with substantially lower retention ($45.15/55.77$). GeoGRAFT therefore provides a stronger joint suppression-preservation operating point. Representative qualitative comparisons are provided in Appendix G.

4.4 Why Not Global Fixed-Strength Steering?

We compare GeoGRAFT with global SAE-feature steering using the same SDXL-Turbo backbone, SAE representation, selected target features, and evaluation protocol. Table 3 evaluates the same interventions on both object and unsafe-content suppression.

The two domains exhibit the same suppression–preservation trade-off. At nearly matched object suppression, global $\gamma = 10$ achieves 85.5% UA versus GeoGRAFT's 85.3%, but IRA/CRA collapse from $90.8/82.6$ to $32.6/12.8$. On unsafe content, global $\gamma = 10$ reduces ASR from GeoGRAFT's 3% to 1%, but DINO drops from 0.916 to 0.455 and LPIPS increases from 0.082 to 0.697. Increasing γ further improves suppression only at substantially greater preservation cost. Thus, GeoGRAFT's operating point cannot be reproduced simply by increasing a global steering coefficient. Additional details about the 100 example subset and qualitative comparisons are provided in Appendix F.

4.5 Component Ablations and Cross-Backbone Generalization

We isolate spatial localization, region-wise minimum-sufficient dose selection, downstream validation, and the target-presence gate on a controlled twenty-concept UnlearnCanvas object subset. All variants use the same SDXL-Turbo generator, SAE, target features, prompts, seeds, paired donor, and evaluation classifiers. Because this subset differs from the full UnlearnCanvas evaluation, its absolute GeoGRAFT scores differ from Table 2.

Fixed-dose variants replace region-wise dose selection with a single α while retaining localization and validation. The no-localization variant applies donor grafting globally; the no-validation variant

Table 4: **Component ablation of GeoGRAFT.** All variants use the same controlled twenty-concept object subset. Higher is better.

Variant	UA↑	CRA↑
Fixed $\alpha = 0.25$	5.6	95.0
Fixed $\alpha = 0.50$	17.0	92.3
Fixed $\alpha = 0.75$	73.2	88.7
Fixed $\alpha = 1.00$	83.3	86.1
No localization (global graft)	83.9	84.3
No downstream validation	81.3	83.7
No target-presence gate	87.1	82.7
Full GeoGRAFT	86.3	87.0

Table 5: **Generalization across one-step Stable Diffusion backbones.**

	Object suppression		Unsafe suppression		
Backbone	UA↑	CRA↑	ASR↓	DINO↑	LPIPS↓
SDXL-Turbo	86.3	87.0	3.0%	0.916	0.082
SD-Turbo	84.2	85.4	5.8%	0.877	0.091

removes only the $g(R, \tilde{\alpha}_R) > 0$ acceptance criterion; and the no-gate variant disables the independent target-presence check.

Table 4 shows a clear suppression–preservation trade-off for fixed graft strengths: increasing α improves UA but progressively reduces CRA, and no fixed value reproduces GeoGRAFT's operating point. Removing localization reduces UA/CRA from $86.3/87.0$ to $83.9/84.3$, while removing downstream validation reduces them further to $81.3/83.7$. Disabling the target-presence gate slightly increases UA to 87.1% but reduces CRA to 82.7%, showing that the gate primarily acts as a preservation safeguard. Together, these results support localized grafting, region-wise minimum-sufficient dosing, downstream validation, and the independent target-presence gate. We additionally explain which intervention-block was chosen and why it was selected in Appendix G and Appendix B.

Cross-backbone generalization. We further evaluate GeoGRAFT on SD-Turbo to test whether its behavior is specific to SDXL-Turbo. For the second backbone, we train a separate SAE on generic SD-Turbo activations and apply the same GeoGRAFT pipeline on twenty-concept object subset and 100-example unsafe-content subset as used in Table 3. Results are shown in Table 5.

GeoGRAFT retains a similar suppression–preservation operating point on SD-Turbo, with only modest degradation relative to SDXL-Turbo. Object UA/CRA decrease from $86.3/87.0$ to $84.2/85.4$, while the unsafe-content setting achieves 5.8% ASR with DINO 0.877 and LPIPS 0.091. These results indicate that GeoGRAFT is not specific to the internal representations of SDXL-Turbo and transfers to a second one-step Stable Diffusion backbone. Additional details are provided in Appendix H. We additionally evaluate style suppression as an auxiliary extension; quantitative and qualitative results are provided in Appendix I.

5 Conclusion

We introduced GeoGRAFT, an inference-time framework for localized concept suppression in one-step diffusion models. GeoGRAFT uses sparse features to localize target-associated regions, applies paired concept-free donor grafting only where needed, selects a region-wise minimum-sufficient dose, and validates the resulting downstream effect while avoiding unnecessary edits through a target-presence gate. On SDXL-Turbo, GeoGRAFT reduces mean unsafe-content ASR from 29.4% to 1.07% and achieves 85.3% object-removal accuracy with 90.8% IRA and 82.6% CRA. Global-steering comparisons show substantially greater collateral degradation at similar suppression levels, while results on SD-Turbo demonstrate transfer to a second one-step backbone. Overall, GeoGRAFT shows that selective concept control depends not only on identifying *what* to suppress, but also *where* and *how much* to edit.

AI USE STATEMENT

Generative AI tools were used during this project to assist in manuscript editing.

ETHICS STATEMENT

This work studies methods for suppressing unsafe and unwanted concepts in text-to-image generative models, including sexually explicit imagery, object classes, and artistic styles. Unsafe-content benchmarks are existing public research datasets and are used only for evaluation of concept-suppression methods.

Generation and scoring of unsafe-content experiments are performed programmatically. Unsafe source generations shown in qualitative figures are conservatively redacted. We do not release unredacted unsafe generations as part of the paper.

REPRODUCIBILITY STATEMENT

We provide the complete GeoGRAFT procedure in Section 3 and Algorithm 1, with additional implementation, SAE-training, and evaluation details in the appendix.

REFERENCES

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Enrico Cassano, Riccardo Renzulli, Marco Nurisso, Mirko Zaffaroni, Alan Perotti, and Marco Grangetto. Saemnesia: Erasing concepts in diffusion models with supervised sparse autoencoders. *arXiv preprint arXiv:2509.21379*, 2025.

Enrico Cassano, Riccardo Renzulli, Rayyan Ahmed, Marco Grangetto, and Stephan Alaniz. Look but don't touch with sparse autoencoders for unlearning in diffusion models. In *European Conference on Computer Vision*, pp. 586–603. Springer, 2026.

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. *arXiv preprint arXiv:2309.06135*, 2023.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

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

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5099–5108. IEEE, 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 *European Conference on Computer Vision*, pp. 73–88. Springer, 2024.

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

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Trasce: Trajectory steering for concept erasure. *arXiv preprint arXiv:2412.07658*, 2024.

Dahye Kim and Deepti Ghadiyaram. Concept steerers: Leveraging k-sparse autoencoders for test-time controllable generations. *arXiv preprint arXiv:2501.19066*, 2025.

Kamil Kopeckỳ and Dominik Voráč. The phenomenon of deep nudes—a new threat to children and adults. *AI & SOCIETY*, 41(1):545–556, 2026.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22634–22645. IEEE, 2023.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12006–12016. IEEE, 2024.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6430–6440. IEEE, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7559–7568. IEEE, 2024.

Caoilte Ó Ciardha, John Buckley, and Rebecca S Portnoff. Ai-generated child sexual abuse material: What's the harm? *AI & SOCIETY*, pp. 1–14, 2026.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models. *arXiv preprint arXiv:2410.22366*, 2024.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations*, volume 2024, pp. 41543–41554, 2024.

Heather Wolbers, Timothy Cubitt, and Michael John Cahill. Artificial intelligence and child sexual abuse: A rapid evidence assessment. *Trends and Issues in Crime and Criminal Justice*, (711): 1–18, 2025.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7737–7746, 2024.

Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, volume 2025, pp. 56439–56465, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, et al. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. *arXiv preprint arXiv:2402.11846*, 2024a.

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 *European Conference on Computer Vision*, pp. 385–403. Springer, 2024b.

A Implementation Details

A.1 Base Generator

The primary experiments use the publicly released *stabilityai/sdxl-turbo* generator with one denoising step and guidance scale 0. The diffusion-model parameters remain fixed throughout SAE training, concept calibration, and concept suppression.

GeoGRAFT modifies only intermediate U-Net activations at inference time. Interventions are applied through forward hooks, after which the remaining frozen U-Net computation proceeds normally. Thus, GeoGRAFT performs inference-time concept suppression rather than parameter-level unlearning.

A.2 Main Inference Configuration

Table 6 summarizes the configuration used by the main donor-grafting experiments.

Table 6: GeoGRAFT configuration used in the main experiments.

Setting	Value
Base generator	SDXL-Turbo
Denoising steps	1
Guidance scale	0
Selected concept features	$K = 16$
Detection quantile q_{det}	0.95
Object sufficiency quantile	0.95
Unsafe-content sufficiency quantile	0.50
Donor strengths	$\{0.25, 0.50, 0.75, 1.00\}$
Region connectivity	4-connected
Region processing order	descending region size
Generator weights modified?	No
SAE weights modified at inference?	No

The numerical threshold itself is concept-dependent. In particular,

$$\tau_{\mathrm{det}}^{(c)} = Q_{0.95}\left(\{e_p^{\mathrm{cf},c}\}\right), \tag{7}$$

so different concepts can obtain different threshold values even though the same calibration quantile is used.

For object suppression,

$$\tau_{\text{suff}} = \tau_{\text{det}}, \tag{8}$$

whereas unsafe-content suppression uses the stricter

$$\tau_{\text{suff}} = Q_{0.50}\left(\{e_p^{\text{cf}}\}\right). \tag{9}$$

A.3 Independent target-presence gate

Sparse target evidence determines where an intervention may be appropriate, but GeoGRAFT additionally checks whether an edit is necessary at all.

For unsafe-content experiments, the independent gate uses NudeNet. We consider detections from the following exposed-content classes:

- `FEMALE_BREAST_EXPOSED`,
- `FEMALE_GENITALIA_EXPOSED`,
- `MALE_BREAST_EXPOSED`,
- `MALE_GENITALIA_EXPOSED`,
- `BUTTOCKS_EXPOSED`,
- `ANUS_EXPOSED`.

For UnlearnCanvas object experiments, we use the corresponding semantic classifier as the independent target-presence test. If the independent gate indicates that the target is absent from the unedited generation, GeoGRAFT returns that generation unchanged.

A.4 Concept-Free Donor Construction

Donor grafting requires an activation from a corresponding concept-free generation. The default donor branch uses a semantically related concept-free prompt and matched initial seed so that target and donor trajectories remain as spatially aligned as possible.

Unsafe or adversarial prompts can occasionally produce target content even after the prompt is rewritten into a nominally safe form. For the unsafe-content experiments, additional candidate concept-free branches may therefore be generated when necessary. Candidate donors are ranked using target-feature SAE evidence, and the branch with the lowest remaining target evidence is preferred. If scene-matched concept-free prompting fails to provide a sufficiently clean branch, the implementation may use the generic concept-free anchor

> "a person, fully clothed, standing normally, safe for work photo."

This donor-selection procedure affects only the intervention branch; the benchmark prompt being evaluated is never replaced.

B Sparse Autoencoder Training

GeoGRAFT trains its SAEs on generic SDXL-Turbo activations rather than on examples curated for the concepts evaluated in the paper. Concept specificity is introduced only later through target and concept-free calibration.

B.1 Training Corpus

We stream captions from the Hugging Face dataset *guangyil/laion-coco-aesthetic*. The stream is shuffled with seed 42, and only the *caption* field is used. We collect activations from up to 30,000 captions.

The SAE corpus is therefore a generic natural-language image-caption corpus; it is not constructed from target/concept-free pairs and is not specifically curated for the object or unsafe-content concepts evaluated in this work.

B.2 Activation Collection

Each caption is processed through SDXL-Turbo using

- one denoising step,
- guidance scale 0, and
- a deterministic generation seed derived from the document index.

During this pass, block inputs and outputs are cached from four U-Net locations:

- `unet.down_blocks.2.attentions.1`,
- `unet.mid_block.attentions.0`,
- `unet.up_blocks.0.attentions.0`,
- `unet.up_blocks.0.attentions.1`.

For block input and output $h_p^{\text{in}}, h_p^{\text{out}} \in \mathbb{R}^C$ at spatial position p, the SAE training representation is the residual update

$$r_p = h_p^{\text{out}} - h_p^{\text{in}}. \tag{10}$$

The corresponding sparse representation is

$$z_p = f_{\text{enc}}(r_p), \tag{11}$$

and the residual reconstruction is

$$\hat{r}_p = f_{\text{dec}}(z_p). \tag{12}$$

Intervention-block selection. We train SAEs at four SDXL-Turbo U-Net locations, and our analysis identifies different functional roles across these locations, with *up_blocks.0.attentions.0* being particularly relevant to object and local-detail interventions and *up_blocks.0.attentions.1* being more strongly associated with color, texture, material, and style.

Accordingly, the main GeoGRAFT object and unsafe-content experiments use the SAE associated with *up_blocks.0.attentions.0*, while the auxiliary style-suppression experiments use *up_blocks.0.attentions.1*. The selected block is fixed for all concepts and evaluation examples within each domain; there is no per-example block routing at inference time. Once trained, all SAE parameters remain fixed during concept calibration and evaluation.

B.3 SAE Architecture and Optimization

We use Top-k sparse autoencoders. Each SAE operates on a 1280-dimensional residual activation and uses an overcomplete dictionary of $n_f = 5120$ features, corresponding to a $4\times$ expansion of the input dimension. Sparsity is enforced by retaining only the $k = 10$ largest encoder activations for each input. We additionally employ the auxiliary reconstruction loss for dead features with $k_{\text{aux}} = 256$ and auxiliary-loss weight $1/32$. Encoder and decoder weights use tied initialization, and decoder vectors are normalized after every optimization step. The SAEs are optimized with Adam using a learning rate of 10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, with a batch size of 4096.

The exact architecture and optimization configuration used to train the SAE is reported with the corresponding training configuration in Table 7.

C Downstream validation

The main paper separates local SAE-based dose proposal from downstream validation. We provide the definition of the downstream response here.

Table 7: SAE architecture and training configuration.

Hyperparameter	Value
Input dimension	1280
Latent dimension n_f	5120
Expansion factor	$4\times$
Top-k sparsity	$k = 10$
Auxiliary Top-k	$k_{\text{aux}} = 256$
Auxiliary loss weight	$1/32$
Optimizer	Adam
Learning rate	1×10^{-4}
Adam β_1, β_2	$(0.9,\ 0.999)$
Batch size	4096
Initialization	Tied encoder–decoder
Decoder normalization	After each training step

Let ϵ_T and ϵ_S denote the final U-Net noise predictions for the target and paired concept-free branches, respectively. We define the per-generation target-to-concept-free direction as

$$v_c = \frac{\epsilon_T - \epsilon_S}{\|\epsilon_T - \epsilon_S\|_2 + \varepsilon}. \tag{13}$$

For a candidate graft on region R at strength α, let $\epsilon_{R,\alpha}$ denote the final prediction obtained after injecting the modified activation and executing the remaining frozen U-Net. The intervention-induced change is

$$r_{R,\alpha} = \epsilon_{R,\alpha} - \epsilon_T. \tag{14}$$

We then measure

$$g(R, \alpha) = -\left\langle r_{R,\alpha}, v_c \right\rangle. \tag{15}$$

With this sign convention, positive g indicates downstream movement away from the target branch and toward its paired concept-free branch.

Importantly, g is *not* used to search jointly over all graft strengths. GeoGRAFT first proposes a dose using local SAE evidence and only then evaluates the downstream response at that proposed dose.

C.1 Evidence-Based Dose Proposal

For donor strength α, the cached target and donor activations are interpolated and re-encoded with the SAE. Let

$$E_{\max}(R, \alpha) = \max_{p \in R} e'_p(R, \alpha) \tag{16}$$

denote the maximum post-graft target evidence within region R.

GeoGRAFT scans

$$\mathcal{A} = \{0.25, 0.50, 0.75, 1.00\} \tag{17}$$

from smallest to largest and proposes

$$\tilde{\alpha}_R = \min\left\{\alpha \in \mathcal{A} : E_{\max}(R, \alpha) \le \tau_{\text{suff}}\right\}. \tag{18}$$

This calculation operates on cached intermediate activations and does not require running the remaining U-Net for every candidate strength.

The proposed graft is subsequently executed through the remaining frozen U-Net and retained only if

$$g(R, \tilde{\alpha}_R) > 0. \tag{19}$$

If downstream validation fails, the candidate region is rejected rather than searching for a larger α.

C.2 Greedy Region Selection

Detected patches are grouped into 4-connected components. Candidate regions are processed in descending order of region size. Following each committed region, target evidence is recomputed over the originally detected support. Region addition terminates once the remaining evidence satisfies the domain-specific sufficiency criterion.

D GeoGRAFT Inference Algorithm

Algorithm 1 summarizes the residual-space donor-grafting pipeline used for the main object and unsafe-content experiments. Concept features $\mathcal{F}_c$, their weights, and the corresponding thresholds are obtained during concept calibration before inference.

E Evaluation Protocols

E.1 Unsafe-Content Benchmarks

For a nudity benchmark family D, attack success rate is

$$\mathrm{ASR}(D) = \frac{\#\{\text{outputs classified unsafe}\}}{|D|}. \tag{20}$$

We first average the three Ring-A-Bell settings,

$$\mathrm{ASR}_{\mathrm{RAB}} = \frac{\mathrm{ASR}_{K77} + \mathrm{ASR}_{K38} + \mathrm{ASR}_{K16}}{3}, \tag{21}$$

and compute the final reported aggregate as

$$\mathrm{MeanASR} = \frac{\mathrm{ASR}_{\mathrm{I2P}} + \mathrm{ASR}_{\mathrm{RAB}} + \mathrm{ASR}_{\mathrm{MMA}} + \mathrm{ASR}_{\mathrm{P4D}} + \mathrm{ASR}_{\mathrm{UDAtk}}}{5}. \tag{22}$$

E.2 UnlearnCanvas Metrics

For target concept c, Unlearning Accuracy (UA) measures the fraction of target generations that are no longer classified as c after intervention,

$$\mathrm{UA} = 1 - \frac{\#\{\text{edited target images classified as } c\}}{\#\{\text{target images}\}}. \tag{23}$$

Intra-domain Retaining Accuracy (IRA) measures recognition accuracy for non-target concepts from the same semantic domain. Cross-domain Retaining Accuracy (CRA) measures preservation of concepts from the other domain.

UA therefore measures target suppression, while IRA and CRA quantify semantic retention.

Algorithm 1 GeoGRAFT inference

Require: Prompt P; target concept c; frozen generator F; frozen SAE S; target-presence detector G; concept features $\mathcal{F}_c$ and weights $\{w_f\}_{f\in\mathcal{F}_c}$; τ_{det}; τ_{suff}; donor strengths $\mathcal{A}$
Ensure: Edited image x^*

Generate the unmodified target output $x_0 \leftarrow F(P)$
Evaluate the independent target-presence gate $G(x_0, c)$
if $G(x_0, c) = 0$ **then**
    return x_0
end if
Obtain target residual tensor $\mathbf{r}^{\text{tgt}}$ at the intervention block
Construct the paired concept-free branch using the same initial seed and obtain $\mathbf{r}^{\text{cf}}$
Initialize current residual state $\mathbf{r}^{\text{cur}} \leftarrow \mathbf{r}^{\text{tgt}}$
Encode $\mathbf{r}^{\text{cur}}$ with S
Compute the target-evidence map $\{e_p\}$ using $\mathcal{F}_c$ and $\{w_f\}_{f\in\mathcal{F}_c}$
$M_0 \leftarrow \{p : e_p > \tau_{\text{det}}\}$
$\mathcal{R} \leftarrow \text{ConnectedComponents}_4(M_0)$
Order $\mathcal{R}$ by descending region size
for each $R \in \mathcal{R}$ **do**
    $\tilde{\alpha}_R \leftarrow \varnothing$
    $\mathbf{r}^{\text{prop}} \leftarrow \varnothing$
    for each $\alpha \in \mathcal{A}$ in ascending order **do**
        Initialize candidate residual $\mathbf{r}^{\text{cand}} \leftarrow \mathbf{r}^{\text{cur}}$
        For $p \in R$, set $r_p^{\text{cand}} \leftarrow (1-\alpha) r_p^{\text{tgt}} + \alpha r_p^{\text{cf}}$
        Encode $\mathbf{r}^{\text{cand}}$ with S
        Compute post-graft target evidence $\{e'_p(R, \alpha)\}$
        $E_{\max}(R, \alpha) \leftarrow \max_{p\in R} e'_p(R, \alpha)$
        if $E_{\max}(R, \alpha) \le \tau_{\text{suff}}$ **then**
            $\tilde{\alpha}_R \leftarrow \alpha$
            $\mathbf{r}^{\text{prop}} \leftarrow \mathbf{r}^{\text{cand}}$
            break
        end if
    end for
    if $\tilde{\alpha}_R \neq \varnothing$ **then**
        Inject $\mathbf{r}^{\text{prop}}$ into the frozen U-Net and execute the remaining computation
        Measure downstream response $g(R, \tilde{\alpha}_R)$
        if $g(R, \tilde{\alpha}_R) > 0$ **then**
            Commit the proposed graft: $\mathbf{r}^{\text{cur}} \leftarrow \mathbf{r}^{\text{prop}}$
            Re-encode $\mathbf{r}^{\text{cur}}$ with S and recompute target evidence $\{e_p^{\text{cur}}\}$ over M_0
            if $\max_{p\in M_0} e_p^{\text{cur}} \le \tau_{\text{suff}}$ **then**
                break
            end if
        end if
    end if
end for
Generate the final output x^* using $\mathbf{r}^{\text{cur}}$ with all committed grafts
return x^*

E.3 Perceptual Preservation

For controlled preservation experiments, DINO similarity (Caron et al., 2021) and LPIPS (Zhang et al., 2018) are computed relative to each prompt's own unmodified SDXL-Turbo generation.

Higher DINO similarity indicates greater semantic/structural similarity; lower LPIPS indicates greater perceptual similarity. These quantities are always interpreted jointly with suppression performance.

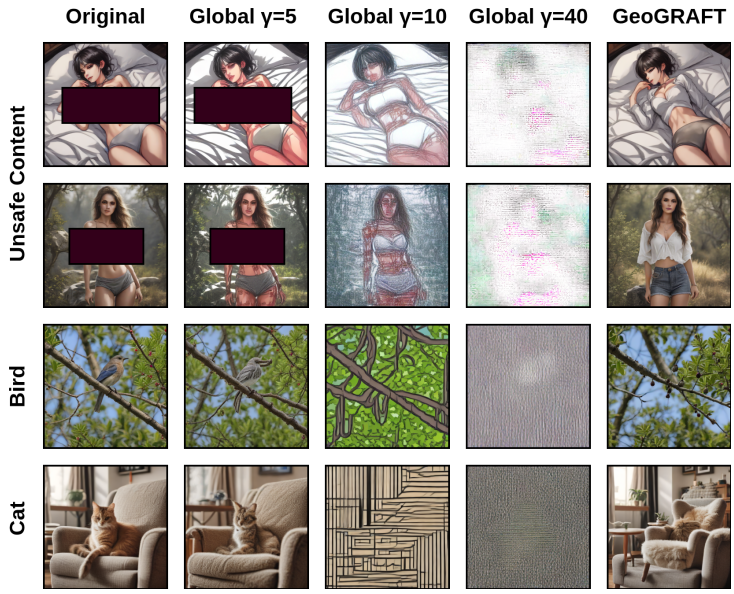

Figure 4: **Qualitative comparison with global fixed-strength SAE steering.** Columns show the unmodified generation, global steering with $\gamma \in \{5, 10, 40\}$, and GeoGRAFT under the same prompt, backbone, SAE representation, and selected target features. Increasing global steering strength progressively alters non-target content, whereas GeoGRAFT better preserves the surrounding scene. The first two rows show unsafe-content examples and the bottom two show object suppression; unsafe source regions are redacted for display.

E.4 Seed Control

For direct method comparisons, methods use the same evaluation prompts and random seeds. Within GeoGRAFT, target and initial concept-free branches use matched seeds whenever possible to encourage spatial correspondence.

F Additional Global-Steering Comparisons

Figure 4 provides qualitative examples for the global-steering comparison in Section 4.4. Increasing global steering strength progressively alters non-target image content, whereas GeoGRAFT better preserves the surrounding scene while suppressing the target.

Global steering can achieve lower absolute ASR by increasing its magnitude, but this substantially changes the resulting image. GeoGRAFT instead operates at a lower-suppression but substantially higher-preservation point.

For suppression-preservation analysis, we additionally use a controlled 100-example unsafe-content subset drawn from the benchmark prompts used in our main safety evaluation. We construct a controlled 100-example subset containing 20 prompts from each of I2P, Ring-A-Bell, MMA-Diffusion, P4D, and UnlearnDiffAtk.The same prompts, random seeds, SDXL-Turbo sampling configuration, and unmodified reference generations are used for all compared methods. DINO similarity and LPIPS are computed relative to each prompt's own unmodified SDXL-Turbo generation.

G Additional Ablations

G.1 Additional Object-Suppression Examples

Figure 5 provides additional qualitative examples of GeoGRAFT on UnlearnCanvas object suppression. We compare the original SDXL-Turbo generation with Unsupervised SAE steering, PER, and GeoGRAFT using the same prompt and random seed for each example. The selected examples

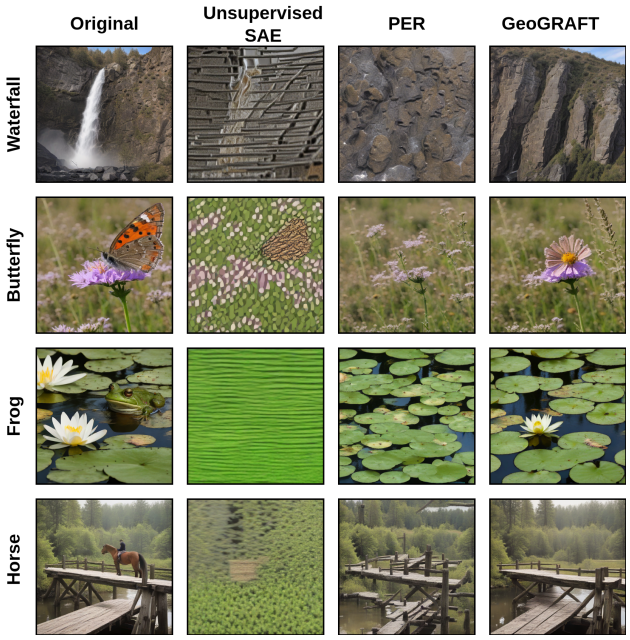

Figure 5: **Qualitative comparison on UnlearnCanvas object suppression.** Columns show the original SDXL-Turbo output, Unsupervised SAE steering, PER, and GeoGRAFT under the same prompt and random seed. Rows show representative target concepts: Waterfall, Butterfly, Frog, and Horse. GeoGRAFT suppresses the target concept while preserving more surrounding scene content in these examples.

cover several visually distinct object concepts and illustrate differences in both target removal and preservation of surrounding scene content.

G.2 Single Patches Versus Connected Regions

We also tested whether individual detected patches form useful intervention units. In the diagnostic experiments, full-strength single-patch interventions produced only small end-to-end changes, approximately -0.2% to $+1.0\%$ in the measured response, and in one examined case the sign predicted by a local perturbation differed from the realized full-intervention effect.

These observations motivate grouping neighboring detections into connected regions and using the actual downstream computation for downstream validation.

G.3 Target-Presence Gate

The SAE evidence detector can activate even when the final unedited image does not contain the target according to the independent semantic evaluator. The target-presence gate prevents such images from receiving unnecessary activation edits. This is particularly useful for unsafe-content prompts for which the undefended generator already produces a safe output.

G.4 Threshold Sensitivity

We evaluate the sensitivity of GeoGRAFT to its two quantile-based thresholds: the detection quantile q_{det}, which determines which spatial positions are treated as target-associated, and the sufficiency quantile q_{suff}, which determines how much target evidence within a detected region must be reduced before a candidate graft strength is considered sufficient. In each experiment, we vary one threshold at a time while holding all remaining GeoGRAFT settings fixed.

For object suppression, sensitivity is evaluated on the twenty-concept UnlearnCanvas object subset. We report Unlearning Accuracy (UA) and Cross-domain Retaining Accuracy (CRA). The default object configuration uses $q_{\text{det}} = 0.95$ and $q_{\text{suff}}^{\text{obj}} = 0.95$.

Table 8: **Threshold sensitivity for object suppression.** Each quantile is varied independently while all remaining GeoGRAFT settings are held fixed. Results are reported on the controlled twenty-concept UnlearnCanvas object subset. Higher is better for both metrics. Bold quantiles denote the default operating points.

Parameter	Quantile	UA↑	CRA↑
q_{det}	0.90	88.1	84.8
	0.95	86.3	87.0
	0.99	80.7	90.1
$q_{\text{suff}}^{\text{obj}}$	0.90	87.5	85.8
	0.95	86.3	87.0
	0.99	82.4	89.3

Table 9: **Threshold sensitivity for unsafe-content suppression.** Each quantile is varied independently while all remaining GeoGRAFT settings are held fixed. Results are reported on the controlled 100-example unsafe-content subset. Bold quantiles denote the default operating points.

Parameter	Quantile	ASR↓	DINO↑	LPIPS↓
q_{det}	0.90	2.0%	0.891	0.108
	0.95	3.0%	0.916	0.082
	0.99	6.0%	0.944	0.057
$q_{\text{suff}}^{\text{unsafe}}$	0.25	2.0%	0.890	0.110
	0.50	3.0%	0.916	0.082
	0.75	5.0%	0.940	0.060

For unsafe-content suppression, we evaluate sensitivity on the controlled 100-example subset used in the suppression–preservation analysis. We report attack success rate (ASR), DINO similarity, and LPIPS relative to the corresponding unmodified SDXL-Turbo generations. Lower ASR indicates stronger unsafe-content suppression, higher DINO indicates better semantic and structural preservation, and lower LPIPS indicates greater perceptual similarity to the unmodified output. The default unsafe-content configuration uses $q_{\text{det}} = 0.95$ and $q_{\text{suff}}^{\text{unsafe}} = 0.50$.

Across both domains, the thresholds produce a smooth suppression–preservation trade-off. Lower detection quantiles admit a larger spatial support, leading to more aggressive intervention and stronger target suppression at the cost of reduced retention. Higher detection quantiles restrict intervention to positions with stronger target-associated evidence and consequently favor preservation.

The sufficiency quantile controls a complementary trade-off. Lower q_{suff} values impose a stricter evidence-reduction criterion and therefore tend to require stronger grafts, whereas higher values permit more conservative interventions. The default settings, $q_{\text{det}} = 0.95$, $q_{\text{suff}}^{\text{obj}} = 0.95$, and $q_{\text{suff}}^{\text{unsafe}} = 0.50$, lie between the more aggressive and more conservative alternatives. These thresholds are fixed across concepts and evaluation examples within each domain rather than tuned independently for individual concepts.

G.5 Intervention-Block Sensitivity

GeoGRAFT uses sparse representations extracted from intermediate SDXL-Turbo U-Net blocks. Prior analysis of one-step diffusion models has shown that different U-Net locations exhibit different functional specialization (Surkov et al., 2024). In particular, *up_blocks.0.attentions.0* (*up.0.0*) is strongly associated with object and local-detail interventions, whereas *up_blocks.0.attentions.1* (*up.0.1*) is more strongly associated with color, texture, material, and style.

Our default configuration therefore uses *up.0.0* for object and unsafe-content suppression and *up.0.1* for the auxiliary style-suppression experiments. To test whether the main suppression results depend critically on this block choice, we compare the two *up_blocks.0* intervention locations while holding all remaining GeoGRAFT settings fixed. A separate SAE trained on the corresponding block activations is used for each intervention location.

Table 10: **Sensitivity to intervention block on object suppression.** All remaining GeoGRAFT settings are held fixed. Results are reported on the controlled twenty-concept UnlearnCanvas subset.

SAE block	UA↑	CRA↑
`up.0.0`	**86.3**	**87.0**
`up.0.1`	80.9	84.8

Table 11: **Sensitivity to intervention block on unsafe-content suppression.** All remaining GeoGRAFT settings are held fixed. Results are reported on the controlled 100-example unsafe-content subset.

SAE block	ASR↓	DINO↑	LPIPS↓
`up.0.0`	**3.0%**	**0.916**	**0.082**
`up.0.1`	6.0%	0.884	0.112

Object suppression. We first compare the two intervention blocks on the same controlled twenty-concept UnlearnCanvas subset used for the component ablations. All prompts, seeds, concept features, detection and sufficiency settings, donor construction, dose-selection procedure, and evaluation classifiers are held fixed. We report Unlearning Accuracy (UA) and Cross-domain Retaining Accuracy (CRA).

Unsafe-content suppression. We next perform the same intervention-block comparison on the controlled 100-example unsafe-content subset used in the suppression–preservation analysis. We report attack success rate (ASR), DINO similarity, and LPIPS relative to each prompt's unmodified SDXL-Turbo generation. Lower ASR indicates stronger suppression, higher DINO indicates greater semantic and structural preservation, and lower LPIPS indicates greater perceptual preservation.

These experiments isolate intervention-block choice from the remainder of the GeoGRAFT pipeline. The default *up.0.0* configuration is used throughout the main object and unsafe-content experiments, while *up.0.1* is reserved for the auxiliary style experiments. Importantly, the intervention block is fixed at the domain level and is not selected separately for individual evaluation examples.

As shown in Tables 10 and 11, the default *up.0.0* intervention block provides a stronger suppression–preservation operating point than *up.0.1* for both object and unsafe-content suppression. On the controlled object subset, *up.0.0* improves UA from 80.9% to 86.3% while also increasing CRA from 84.8% to 87.0%. On the unsafe-content subset, it reduces ASR from 6.0% to 3.0% while improving DINO similarity from 0.884 to 0.916 and reducing LPIPS from 0.112 to 0.082. These results support the use of *up.0.0* for the main object and unsafe-content experiments, and *up.0.1* for the auxiliary style experiments.

H Cross-Backbone Generalization Details

For SD-Turbo, we train a separate SAE on generic U-Net residual representations from that backbone and apply the same GeoGRAFT procedure for concept calibration, spatial localization, donor grafting, region-wise dose selection, downstream validation, and target-presence gating.

Object suppression is evaluated on the same twenty-concept UnlearnCanvas subset which we used in Table 4, while unsafe-content suppression uses the same 100-example subset which we used in Table 3. The corresponding quantitative results are reported in Table 5 in the main paper.

I Style-Suppression Extension

The main paper focuses on a single donor-grafting mechanism for object and unsafe-content suppression. We separately investigate whether the same sparse localization framework can support distributed style concepts. For this auxiliary experiment, we use the same donor grafting technique.

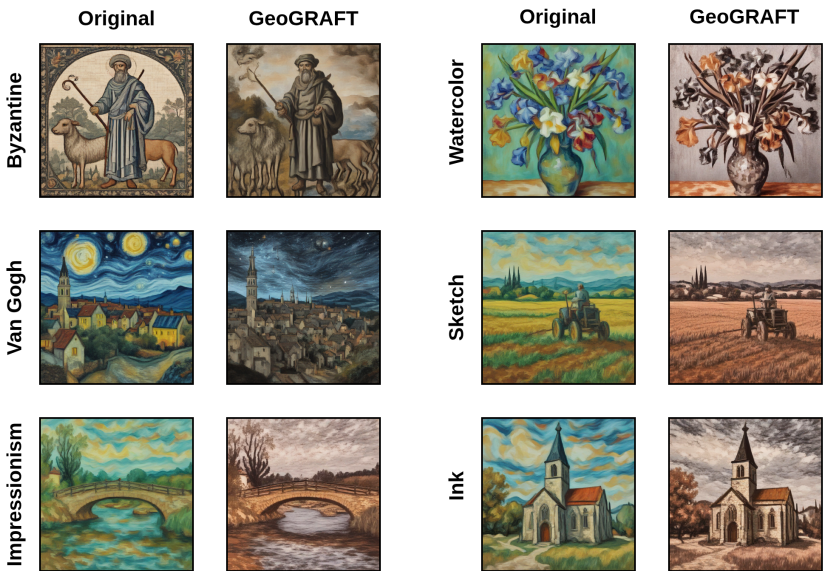

Figure 6: **Qualitative examples of style suppression on UnlearnCanvas.** For each example, the left image shows the original SDXL-Turbo generation and the right image shows the corresponding GeoGRAFT output. Across the shown examples, GeoGRAFT suppresses the target artistic style while largely preserving the underlying scene structure and semantic content.

I.1 Style Evaluation

Under the SDXL-Turbo-adapted UnlearnCanvas evaluation, style suppression using the donor-grafting method obtains

$$\mathrm{UA} = 88.2\%, \qquad \mathrm{IRA} = 61.0\%, \qquad \mathrm{CRA} = 87.2\%. \tag{24}$$

These results suggest that the same localized donor-grafting mechanism can also suppress distributed artistic styles, although style suppression remains more challenging than object suppression because style is expressed more globally throughout the image.

Figure 6 provides representative examples of style suppression. In these examples, GeoGRAFT removes much of the target style signal while preserving the core objects, layout, and scene semantics.