A CONCEPT IS MORE THAN THE SUM OF ITS PARTS: CONCEPT ERASURE BY EXCISING THE SURPLUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept erasure aims to remove undesired concepts from text-to-image diffusion models without harming their ability to generate other concepts. Recent training-free methods tackle the problem by suppressing a direction that represents the target concept. However, this direction also carries parts the concept shares with similar ones (e.g., the wheels a garbage truck shares with cars), damaging these nearby concepts as well. Prior works mitigate this issue by protecting a retain set of related concepts, yet their performance is sensitive to which neighbors are chosen. Instead of searching for better neighbors outside the concept, we look inside the concept itself and let the parts of the concept decide what to keep. Since the same parts can form different wholes, a concept is more than the sum of its parts. We call this extra component the concept surplus: the portion of its features that its parts cannot explain. Thus, erasing only this surplus prevents the model from assembling the concept. Meanwhile, its parts remain intact, and so does every concept that shares them. Based on this motivation, we propose to **Ex**tract the **C**oncept **I**dentity as its **S**urplus over its parts, and **E**rase it. With the same principle, our method, named **EXCISE**, can erase a single part or a subset thereof, such as nudity, while preserving the rest of the person. The experimental results validate that EXCISE achieves strong and robust erasure performance across unlearning tasks while better preserving related concepts, without hand-picked or costly retrieved external concept lists.

1 INTRODUCTION

Trained on large-scale image–text data collected from the internet (Schuhmann et al., 2022), text-to-image diffusion models (Rombach et al., 2022; Podell et al., 2024; Black Forest Labs, 2024) can now generate images that are often indistinguishable from real photographs. However, the same data also contains explicit content, copyrighted characters, and the likeness of real people, which the models readily learn to reproduce (Schramowski et al., 2023). Since naïvely retraining a model on a filtered dataset is prohibitively expensive, concept erasure has emerged to remove a target concept from a pretrained model while preserving its ability to generate other concepts. One of the main challenges in concept erasure is how to remove a target concept without harming model utility on unrelated concepts and specifically nearby concepts.

To tackle the problem, early approaches fine-tune the model to suppress the target concept (Gandikota et al., 2023; Kumari et al., 2023; Lu et al., 2024), often at a high cost and at the expense of generation quality. Consequently, training-free methods have gained popularity, either by editing the cross-attention projections in closed form (Gandikota et al., 2024; Gong et al., 2024; Li et al., 2026; Sun et al., 2026) or by intervening during generation (Schramowski et al., 2023; Wang et al., 2025; Gaintseva et al., 2026a). Despite their differences, these methods share a common recipe: they identify a direction that represents the target concept and suppress it.

However, a concept direction rarely represents the concept alone. For instance, the direction for "garbage truck" also encodes wheels and windows that it shares with cars. Thus, suppressing it damages nearby concepts, a problem known as over-erasure; erasing nudity, for example, can prevent the model from generating a person at all (Kim et al., 2026; Saha et al., 2025).

To mitigate over-erasure, recent works constrain the target direction with a retain set of related concepts, which is predefined (Gandikota et al., 2024; Lu et al., 2024; Sun et al., 2026; Lee et al.,

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

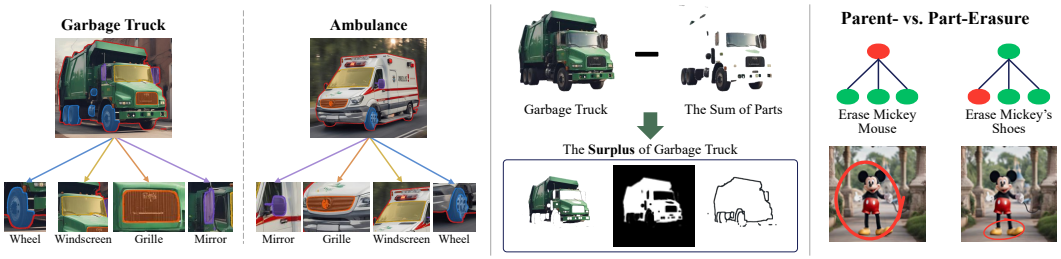


Figure 1: **Concept surplus and hierarchical erasure. Left:** Garbage trucks and ambulances share visual components, motivating preservation of shared structure during erasure. Middle: The schematic subtraction illustrates the concept surplus: the portion of the target representation unexplained by a selected low-rank subspace of component representations. **Right:** Parent erasure targets the whole concept (Mickey Mouse), using its components to guide preservation. Part erasure targets a selected component (the shoes), using non-target siblings to guide preservation of the remaining parent content. Red marks the erasure target; green marks content intended to be preserved.

2025; Vardhana & Biswas, 2026), filtered from candidate prompts (Li et al., 2026), collected in a bank of anchors (Upman et al., 2026), or retrieved from a token vocabulary (Kim et al., 2026; Saha et al., 2026; Shi et al., 2026). Yet the outcome depends on which concepts end up in the retain set, and retained concepts that closely resemble the target absorb the very component to be erased (Upman et al., 2026; Shi et al., 2026). More importantly, the retain set is always composed of *other* concepts outside the target; even when extracted from images of the target, it consists of text tokens of co-occurring concepts rather than the visual parts that compose the target itself (Kim et al., 2026).

In this work, we take a different perspective: rather than searching *outside* the target concept for what should be preserved, we look *inside* the concept itself. Our key intuition is that a concept is more than the sum of its parts. Just as the same letters spell both anagrams "listen" and "silent," the same wheels, cab, and box-shaped body make up both a garbage truck and a delivery van. Hence, what should be removed is not the parts but the component that makes their composition specifically a target concept (e.g., garbage truck), which we call the *surplus*: the portion of the concept's activation that the span of its parts cannot explain. Thus, erasing only the surplus prevents the model from assembling the target concept, while other concepts that share its parts remains intact.

Moreover, the same principle naturally extends from whole-concept erasure to finer-grained removal. Erasing an object or a character removes the surplus of the whole over its parts (the *root cut*). On the other hand, erasing nudity removes the surplus of one part over its siblings (the *sibling cut*), preserving what the explicit regions share with the rest of the person. In this way, whole-concept and part-level erasure arise from the same operation at different levels of the part–whole hierarchy, rather than requiring separate mechanisms.

To this end, we propose a new training-free framework, dubbed **ExCISE** (**Ex**tract the **C**oncept Identity as its **S**urplus over its parts, and **E**rase it), which computes the surplus from the model's own generations. Specifically, ExCISE localizes a concept and its parts with the model's internal features and isolates the surplus by projecting the concept's activation away from those of its parts. Importantly, ExCISE projects against the concept's own parts, rather than other concepts. As the resulting surplus is specific to the concept, ExCISE steers generation without a gating mechanism (Lee et al., 2025; Gaintseva et al., 2026a) and does not require hand-picked retain sets.

Our contributions are summarized as follows:

- We introduce the notion of *surplus*, the portion of concept's activation that its parts cannot explain, and show that erasing only the surplus removes the concept while preserving its parts as well as nearby concepts. This unifies whole-concept and part-level erasure as the root and sibling cuts of a part–whole hierarchy.
- We propose ExCISE, a training-free method that computes the surplus from the model's own generations, localizing a concept and its parts with the model's internal features. Ex-

CISE requires no gating mechanism or carefully tuned/retrieved external concept retain sets.

- We demonstrate through experiments on object, character, and nudity erasure that EXCISE achieves strong erasure performance while improving the preservation of nearby concepts across tasks.

2 RELATED WORK

Concept erasure. Concept erasure removes target concepts from pretrained diffusion models while preserving their remaining generation capabilities. Early approaches rely on fine-tuning, including ESD (Gandikota et al., 2023), Concept Ablation (Kumari et al., 2023), Selective Amnesia (Heng & Soh, 2023), SPM (Lyu et al., 2024), and MACE (Lu et al., 2024). More recent approaches explore reducing the cost of retraining through closed-form model editing or inference-time intervention. UCE (Gandikota et al., 2024) and RECE (Gong et al., 2024) modify cross-attention projections, while CURE (Biswas et al., 2026) and OCE (Sun et al., 2026) directly manipulate model parameters or their subspaces. Meanwhile, AdaVD (Wang et al., 2025) extends this principle to inference time by removing target-aligned components in cross-attention value space. Several recent methods exploit orthogonal complements or low-dimensional subspaces to separate target information from information that should remain Wang et al. (2025); Biswas et al. (2026); Saha et al. (2026); Upman et al. (2026); Sun et al. (2026). EXCISE builds on this general mechanism, but differs in how the direction to be erased is constructed, as discussed below.

Preserving related concepts. Strong erasure can unintentionally suppress concepts related to the target, a problem known as over-erasure (Saha et al., 2026; Kim et al., 2026). To mitigate this, recent works explicitly model information that should be retained. FADE (Thakral et al., 2025) explicitly preserves a sampled semantic neighborhood, ReCARE (Kim et al., 2026) constructs a set of benign co-occurring concepts from target images, and PARSE (Saha et al., 2026) dynamically retrieves erase and retain concepts from the model vocabulary. GenErase (Vardhana & Biswas, 2026) protects a preservation subspace during inference-time editing, while NLCE (Shi et al., 2026) explicitly stabilizes neighboring concepts before localized erasure. CARE (Upman et al., 2026) similarly refines the target direction using a bank of retained-concept anchors. While differing in how these concepts are obtained, these methods define retained structure through additional concepts outside the target—whether predefined, sampled, retrieved, or collected as anchors. Consequently, the resulting retain space depends on which neighboring concepts are selected. This dependence is particularly important for concepts close to the target, which may themselves contain the component intended for removal. In this work, we take a different view. Rather than searching outside the target for what should be preserved, we look inside the target itself. Its visual components define the retained structure, and the portion of the target unexplained by these components defines the surplus to erase. Preservation therefore requires neither a predefined nor a retrieved concept-level retain set.

Spatial localization. Spatial information has commonly been used to control *where* an erasure operation acts. MACE (Lu et al., 2024) uses Grounded-SAM masks to supervise attention-localized suppression during refinement. GLoCE (Lee et al., 2025) applies its concept intervention through a spatial gate, and NLCE (Shi et al., 2026) similarly uses an attention-guided gate before localized hard erasure. In these methods, the target representation is first defined and spatial information subsequently determines *where* the intervention should be applied. By contrast, our method EXCISE instead uses spatial structure to determine *what to remove, not where*. The masks select the activations of the target and its parts, from which the surplus is computed. Since the surplus is specific to the target, EXCISE does not require a gate at inference.

Representation steering. Some inference-time methods steer the outputs of cross-attention layers directly. CASteer (Gaintseva et al., 2026a) derives training-free steering vectors from cross-attention outputs and applies them directly during generation, while MidSteer (Gaintseva et al., 2026b) generalizes such interventions through a minimal-disturbance affine formulation. Our method EXCISE operates in a related representation space, but first decomposes the target into shared component structure and target-specific surplus before steering. Importantly, EXCISE focuses on refining what is steered, rather than introducing another steering operator. More detailed comparisons on the requirements can be found in Table 1.

Table 1: Requirements, erasure performance and robustness. Retain mechanism and data describe what each method needs before erasure; performance is rated strong, mid or weak per concept group; a mark denotes erasure holding under the stated condition.

	Requirements			Erasure performance			Robustness		
Method	**Retain mech.**	**Data Req.**	**Gates**	**Nudity**	**Identity**	**Object**	**Advers.**	**Paraphr.**	**Co-occ.**
CASteer	text-site	contrastive prompts	1	**strong**	mid	mid	✓	✓	✗
NLCE	text-side	retrieved neighbours	2	weak	**strong**	mid	✗	✓	✓
AdaVD	text-side	none	1	weak	mid	**strong**	✗	✗	✗
Ours	**image-side**	**none**	**0**	**strong**	**strong**	**strong**	✓	✓	✓

3 Problem Formulation

Given a pretrained text-to-image diffusion model and a target concept c_{e}, we aim to suppress c_{e} while preserving other content, including concepts that share its visual components (or parts) with the target concept. In this work, we consider training-free interventions on intermediate activations with model parameters held fixed.

Here, c_{e} denotes the semantic concept (e.g., "garbage truck"), whereas $\boldsymbol{x}$ denotes its pooled feature representation. The indexed representation $\boldsymbol{x}_{\ell,t}^{(i)}$ varies with reference image i, layer ℓ, and sampling step t; we omit these indices in this formulation for the sake of simplicity.

Let $\boldsymbol{p}$ denote the representation pooled from a parent region (i.e., the whole object) and $\mathcal{R}(\boldsymbol{p}) = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m\}$ its observed component representations. We define the target representation $\boldsymbol{x}$ and retain set $\mathcal{R}_{\mathrm{keep}}$ through two settings: :

$$(\boldsymbol{x}, \mathcal{R}_{\mathrm{keep}}) = \begin{cases} (\boldsymbol{p}, \mathcal{R}(\boldsymbol{p})), & \text{root cut,} \\ (\boldsymbol{r}_a, \{\boldsymbol{r}_j : j \neq a\}), & \text{sibling cut.} \end{cases} \tag{1}$$

A *root cut* (or *parent erasure*) erases the whole target concept identity while retaining component structure reusable by other concepts. A *sibling cut* (or *part erasure*) erases a selected component while retaining its non-target siblings. The latter excludes the parent from the retain set because it also contains the target. Both settings require refining the target representation against structure shared with the retained components.

4 Method

In this work, we propose to **Ex**tract the **C**oncept **I**dentity as its **S**urplus over its parts, and **E**rase it. To this end, we introduce **ExCISE**, a training-free method that constructs an erasure direction from the target's *surplus*: the portion of its representation unexplained by a low-rank subspace of retained components. ExCISE (1) pools cross-attention output activations within segmented parent and component regions (Section 4.1), (2) constructs a shared Gram retain span from the target's own spatial parts across generated instances, and (3) orthogonalizes the parent representation against this shared-part span in whitened space to extract the concept (Section 4.2). The resulting surplus defines a refined target direction (i.e., steering direction) directly from the model's own representations, without selecting a separate set of neighboring concepts (Section 4.3). The overall framework is illustrated in Figure 2.

4.1 Region Representations

For each target concept, we generate n reference images and cache cross-attention output matrices $\boldsymbol{O}_{\ell,t}^{(i)} \in \mathbb{R}^{N_\ell \times d_\ell}$. Here, N_ℓ is the number of spatial positions; d_ℓ is the feature dimension; and $t = 0$ denotes the first executed sampling step. Each output row corresponds to a spatial query position, enabling region-level pooling (Appendix A.1).

Attention saliency selects a bounding region for each reference image, and SAM2 (Ravi et al., 2025) produces parent and component masks. We resize and flatten each mask into a vector $\boldsymbol{m} = (m_s)_{s=1}^{N_\ell}$.

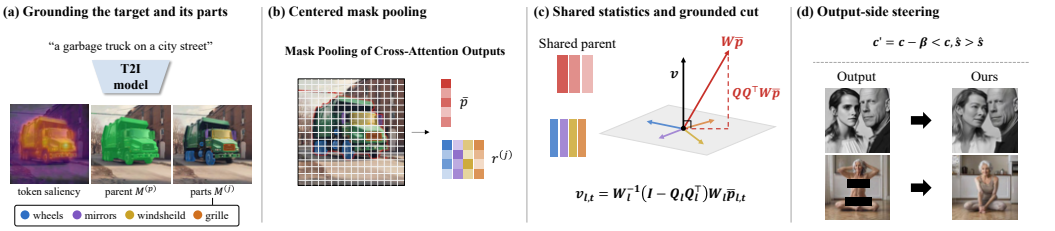


Figure 2: **Overview of ExCISE: concept erasure by excising the surplus.** (a) Parent and component regions in reference generations are identified. (b) Pooling spatially centered cross-attention outputs within these regions yields their representations. (c) We aggregate representations across images and remove the target's projection onto a shared low-rank retain subspace in whitened space. The remaining surplus captures what this subspace cannot explain and is mapped back to activation space. (d) The resulting direction steers cross-attention outputs throughout denoising to suppress the target. The same construction supports parent and part erasure: the parent's components define the retain subspace in the former, and the selected part's non-target siblings in the latter.

For $\sum_s m_s > 0$, the spatially centered region representation is

$$\boldsymbol{f}_{\ell,t}^{(i)}(\boldsymbol{m}) = \frac{\sum_s m_s \boldsymbol{o}_{\ell,t}^{(i)}(s)}{\sum_s m_s} - \frac{1}{N_\ell} \sum_s \boldsymbol{o}_{\ell,t}^{(i)}(s), \tag{2}$$

where $\boldsymbol{o}_{\ell,t}^{(i)}(s) \in \mathbb{R}^{d_\ell}$ is the spatial feature at position s, expressed as the transpose of row s of $\boldsymbol{O}_{\ell,t}^{(i)}$. Centering measures each region relative to the image-wide mean. Pooling parent and component masks yields the target and retain representations in Equation 1.

4.2 Surplus over the retain subspace

Normalizing representation geometry. High-variance activation directions can dominate target–component alignment. We apply a shared, invertible layer-specific zero-phase component analysis (ZCA) whitening transform (Kessy et al., 2018), $\boldsymbol{W}_\ell \in \mathbb{R}^{d_\ell \times d_\ell}$, estimated from cached activations. Whitening rescales covariance directions by their inverse standard deviations, placing features on comparable scales. Applying the same transform to the target and retained components makes their alignment relative to the observed activation variability, consistent with covariance-aware analyses of concept representations (Nava & Wyart, 2026; Upman et al., 2026). We construct the retain subspace and compute the surplus in this shared geometry, then map the residual back for steering. Appendix A.2 gives the metric interpretation, and Figure A.1 visualizes its effect on parent–component alignment.

Shared retain subspace. To aggregate component structure across instances, we stack the whitened retain representations as columns and average the whitened target representations:

$$\boldsymbol{R}_{\ell,t} = \left[\boldsymbol{W}_\ell \boldsymbol{r}_{\ell,t}^{(i,j)}\right]_{(i,j)\in\mathcal{K}}, \qquad \bar{\boldsymbol{x}}_{\ell,t} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_\ell \boldsymbol{x}_{\ell,t}^{(i)}, \tag{3}$$

where $\mathcal{K}$ indexes the retained component records. Let $\boldsymbol{B}_{\ell,t} \in \mathbb{R}^{d_\ell \times k}$ contain the leading k left singular vectors of $\boldsymbol{R}_{\ell,t}$, with $\boldsymbol{B}_{\ell,t}^\top \boldsymbol{B}_{\ell,t} = \boldsymbol{I}_k$. Its columns define a compact retain subspace shared across reference images.

We subtract the target's projection onto this subspace and map the residual back to activation space:

$$\boldsymbol{v}_{\ell,t} = \boldsymbol{W}_\ell^{-1}\left(\boldsymbol{I}_{d_\ell} - \boldsymbol{B}_{\ell,t}\boldsymbol{B}_{\ell,t}^\top\right)\bar{\boldsymbol{x}}_{\ell,t}. \tag{4}$$

The same operation handles both root cut and sibling cut through the target–retain assignments in Equation equation 1. By construction, $\boldsymbol{B}_{\ell,t}^\top \boldsymbol{W}_\ell \boldsymbol{v}_{\ell,t} = \boldsymbol{0}$: the direction is orthogonal to the selected retain subspace. We empirically assess preservation in generated images through retained-identity accuracy and neighboring-concept CLIP score (Sections 5.2–5.4).

The surplus depends on the retained components, rank, and metric. In particular, when disjoint component masks partition a parent, linear pooling places the mean parent in the full span of the

Table 2: Nudity detection rate on I2P. The final column reports attack success rate (ASR, best of three) on Ring-A-Bell. The best value per column is bold and the second best underlined; tied values share a mark. Rows marked ‡ are trained on separate GPUs.

Method	Armpits	Belly	Buttocks	Feet	F. Breasts	F. Genitalia	M. Breasts	M. Genitalia	Any	ASR
Original	97	144	21	18	172	8	25	5	323	95
AdaVD	94	144	21	18	166	9	21	5	314	95
NLCE	72	112	10	11	136	7	18	7	244	94
UCE	58	87	9	16	87	5	12	<u>2</u>	185	90
ReCARE ‡	<u>8</u>	23	3	9	13	2	6	<u>2</u>	54	<u>38</u>
ESD-u ‡	10	15	**0**	6	<u>7</u>	<u>1</u>	3	**0**	33	47
CASteer	**0**	<u>1</u>	<u>1</u>	<u>2</u>	<u>7</u>	**0**	**0**	<u>2</u>	<u>13</u>	**0**
ExCISE (Ours) ($\alpha = 2.5$)	**0**	**0**	<u>1</u>	**0**	**3**	2	<u>1</u>	5	**12**	**0**

collected components. The low-rank constraint defines which component modes are retained and which target information remains available for erasure. Figure 3 examines the empirical effect of rank; Appendix A.5 discusses its role in the decomposition.

4.3 Steering generation

For each layer, we construct $\boldsymbol{v}_{\ell,0}$ at the first sampling step and reuse it at every generation step without an input-dependent spatial or token gate. Thus, direction construction is performed once; its application continues throughout denoising. The masks are needed only during construction, and inference stores one direction per edited layer. For a fixed erasure target, this amortizes the construction cost across generations, limiting the inference-time intervention to applying one cached direction per edited layer.

This extraction step is motivated by the Taylor Swift analysis in Appendix A.4: the retain span explains a larger fraction of the target representation at the first step than mid-trajectory. We use $n = 20$ reference generations, retain rank $k = 8$, and steering strength $\alpha = 1.5$ for root cuts and $\alpha = 2.5$ for sibling cuts. These settings control the empirical erasure–preservation trade-off.

5 Experiments

We evaluate whether ExCISE suppresses target concepts while preserving related content, and whether deriving the retain subspace from visual components improves this trade-off. Our experiments cover nudity erasure, feature removal, celebrity erasure, object and character erasure on SD v1.4 (Rombach et al., 2022) and SDXL (Podell et al., 2024). Additional results over tasks on Flux.1 dev [1] can be found in Appendix C.

Experimental setup. Our comparisons include ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), AdaVD (Wang et al., 2025), NLCE (Shi et al., 2026), ReCARE (Kim et al., 2026), and CASteer (Gaintseva et al., 2026a). We use each baseline's published task settings when available and otherwise its repository defaults. ExCISE uses the configuration in Section 4: $n = 20$ reference generations, rank $k = 8$, and steering strengths $\alpha = 2.5$ for *sibling cuts* (a.k.a. *part erasure*) and $\alpha = 1.5$ for root cuts (a.k.a. *parent erasure*), following the analysis of differences of angles between target direction and retain span and our rank ablations in Section 5.4. Additional results on multi-concept and art style erasure as well as concept switching can be found in Appendices D, E, and F.

5.1 Part erasure

Nudity erasure tests whether explicit regions can be suppressed while retaining the surrounding person. Hence, nudity erasure is the key example of *part erasure*. Exposed body parts need to be

[1] https://huggingface.co/black-forest-labs/FLUX.1-dev

Table 3: Celebrity erasure on SDXL, averaged over four erased celebrities. GIPHY celebrity detector, identified at `is_known` ≥ 0.4. Acc_e is the erased celebrity still identified (lower is better), Acc_s a preserved celebrity identified (higher is better), and H the mean of the per-celebrity harmonic means of $1-\text{Acc}_e$ and Acc_s. Best per column in bold, second best underlined; AdaVD is excluded from the ranking because it leaves prompts that do not name the target byte-identical to the unedited model.

	Single-celebrity			Pairwise		
Method	**Acc**$_e$ ↓	**Acc**$_s$ ↑	H ↑	**Acc**$_e$ ↓	**Acc**$_s$ ↑	H ↑
Orig	85.10	86.32	21.44	50.55	53.35	50.76
ESD	0.40	74.09	84.82	1.00	20.30	33.16
UCE	10.60	81.84	85.25	1.15	61.80	75.95
ReCARE	<u>0.20</u>	9.42	16.90	0.30	12.30	21.72
AdaVD	0.00	86.32	92.66	12.60	57.75	69.51
NLCE	**0.10**	80.65	89.25	**0.00**	<u>66.15</u>	<u>79.60</u>
CASteer	13.20	**82.77**	84.25	<u>0.05</u>	15.05	25.61
NLCE + Ours	1.30	<u>82.19</u>	<u>89.69</u>	0.20	**67.20**	**80.30**
Ours	0.40	81.55	**89.70**	0.50	54.55	70.34

erased while the parent (*person*) and sibling components (e.g., head, hands) should be retained. We evaluate our method on the full I2P (Schramowski et al., 2023) dataset and report the Top-3 ASR on the Ring-A-Bell (Tsai et al., 2024) nudity prompt set for SD v1.4[2]. We use NudeNet (Praneeth, 2019) with a detection threshold of 0.6 for detecting nudity in generated images. Following previous evaluations on nudity erasure (Gaintseva et al., 2026a; Shi et al., 2026) and evaluate on SD v1.4. Table 2 reports category-level counts and the number of images with any detected nudity. ExCISE reduces the latter from 323 for the unedited model to 12, obtaining the lowest aggregate nudity detection count among the evaluated methods.

Further results on Ring-A-Bell nudity prompts under the best-of-three attack protocol show that ExCISE obtains an observed attack success rate of zero. The specific Ring-A-Bell nudity prompt set contains prompts optimized on SD v1.4 to recover erased. These results establish strong suppression by ExCISE on the evaluated prompts; preservation is assessed separately through retained-content comparisons. To additionally evaluate robustness and enable a direct comparison with CASteer, we run UnlearnDiffAtk (Zhang et al., 2024) against CASteer (Gaintseva et al., 2026a), ReCARE (Kim et al., 2026), and ExCISE. UnlearnDiffAtk iteratively optimizes adversarial prompt perturbations to recover erased concepts. While ReCare breaks with 39.0% and CASteer sits at 19.5%, ExCISE shows the strongest robustness with an attack success rate of 10.2%.

Our selective mechanism further allows individual nudity concepts to be erased separately (e.g., *female breasts*). While other nudity erasure methods often remove a set of visual components related to nudity, such as begin exposed arms and legs, our proposed method erases female breasts while retaining other nudity-related body components, demonstrating the flexibility and precision of our method across models. Figure 5a demonstrates qualitative results for selective female-breast erasure on FLUX.1 dev and SDXL. Even fine-grained image properties such as pose, identity, and lighting remain preserved after the edit.

Our *part erasure* also allows for excision of fine-grained visual features (e.g., "a cowboy hat"). The cowboy-hat example in Figure 4 illustrates removal of the selected component ("a cowboy hat") while retaining the person and other hats, using the same design as nudity erasure. This changes the target region while retaining the same direction-construction procedure.

[2]https://huggingface.co/datasets/Chia15/RingABell-Nudity

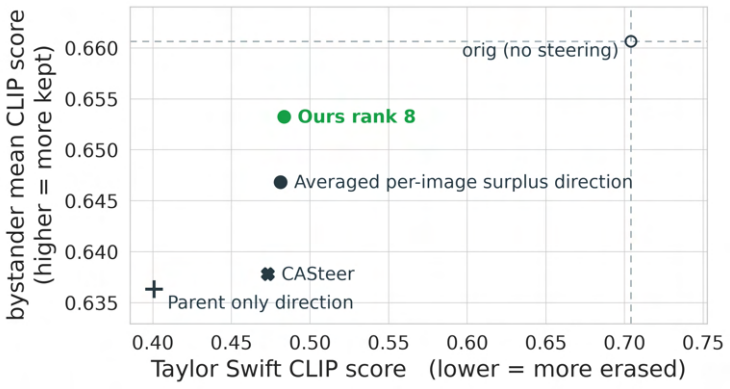


Figure 3: Component ablation and rank probe on Taylor Swift across a set of 8 related concepts.

5.2 Identity erasure

For SDXL, we follow the MACE (Lu et al., 2024) celebrity benchmark and average results over four erased identities. Following their evaluation protocol, we use the GIPHY detector with a threshold of 0.4. Table 3 reports target identification, $\mathrm{Acc_e}$, and retained-identity identification, $\mathrm{Acc_s}$; lower values of the former and higher values of the latter are preferred. We evaluate single-celebrity prompts and pairwise prompts containing erased and retained identities in both name orders.

For single-celebrity prompts, ExCISE achieves $\mathrm{Acc_e} = 0.40\%$ and $\mathrm{Acc_s} = 81.55\%$, compared with 13.20% and 82.77% for CASteer. Thus, target recognition decreases substantially while retained-identity recognition remains similar. On pairwise prompts, ExCISE obtains 0.50% target identification and 54.55% retained-identity identification. The latter is close to the unedited model's reported 53.35% and exceeds CASteer's 15.05%, while remaining below NLCE's 66.15%.

On FLUX.1 dev, we evaluate three identities recognized in more than 60% of their unedited generations. ExCISE obtains 1.33% target identification and 56.11% retained-identity identification. Full results can be found in Appendix C.

Interestingly using NLCE's (Shi et al., 2026) GLOCE-style localised gating (Lee et al., 2025) orchestrated by our surplus direction instead of their retrieved-vocabulary retained direction, we can improve their performance on both single-instance and pairwise prompts. This shows the strength of our direction within modular gating approaches.

Results across nudity and identity erasure empathise the wide strength of our method. ExCISE is strong across all erasure tasks and robust against co-occurence, in-the-wild and adversarially trained prompts as well as iterative adversarial attacks. We have summarized the strengths of our method in Table 1.

5.3 Neighbor preservation and ablations

Figure 3 examines Taylor Swift erasure alongside eight related concepts. These concepts include *a blonde girl, Donald Trump, Justin Bieber, Miley Cyrus, pop star, Selena Gomez, singer, and woman.*

Target and mean neighboring-concept CLIP scores are normalized by their respective unedited baselines. Lower target scores indicate greater suppression, while higher neighboring-concept scores indicate better preservation under this metric.

The shared retain-subspace construction preserves more neighboring-concept similarity than averaging independently constructed per-image surplus directions at a similar target score. It also improves preservation relative to CASteer at a similar target score. In contrast, the parent-only direction yields stronger suppression but a larger decrease in neighboring-concept similarity, illustrating the trade-off incurred by removing the unrefined parent representation.

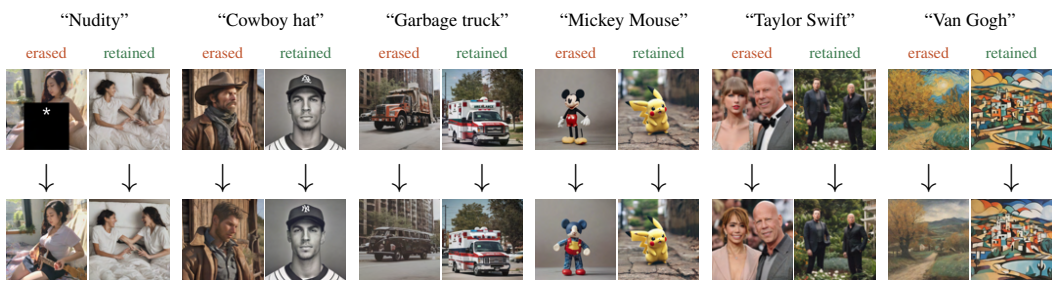
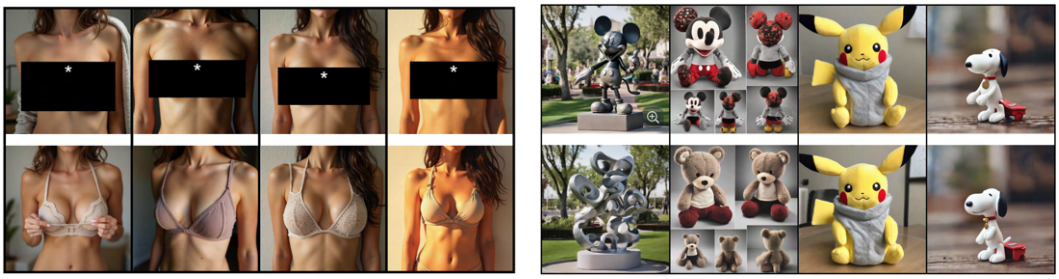


Figure 4: Qualitative results on SDXL across six erasure settings. In each group the left column is the erased target and the right column a retained neighbour; the top row is the original model and the bottom row the edited model under the same seed.



(a) (b)

Figure 5: Overall caption. (a) Nudity erasure on FLUX.1-dev. (b) Mickey Mouse erasure on SDXL (first two columns) and nearby character preservation (last two columns).

5.4 Ablation results

Figure 3 and B.1 evaluate Taylor Swift erasure alongside eight related concepts, varying the components used and the rank. Target and mean neighboring-concept CLIP scores are normalized by their respective unedited baselines. Lower target scores indicate greater suppression under this metric, while higher neighboring-concept scores indicate greater preservation of text–image similarity. All three retain more neighboring-concept similarity than the parent-only direction. Rank 4 yields stronger target suppression but lower preservation than ranks eight and twelve. These comparisons support jointly estimating the retained component structure and treating rank as a control on the erasure–preservation balance.

6 Conclusion

In this work, we introduce ExCISE, a training-free method built on the notion of *surplus*: the portion of a target representation unexplained by a shared low-rank subspace of retained components. By extracting this surplus in whitened space, ExCISE derives erasure directions from the target's own visual structure and unifies parent and part erasure without an external retain set. Experiments show effective nudity suppression on SD v1.4 and identity suppression on SDXL and FLUX.1 dev. The direction-construction ablation further supports sharing component structure across reference images to improve neighboring-concept preservation at comparable target suppression. These results are consistent with our intuition that a concept is *more than the sum of its parts*: its surplus relative to the selected low-rank component subspace provides a useful direction for selective erasure.

AI USE STATEMENT

We did not use generative AI tools to formulate our hypotheses, problem definition, or research ideas, nor to design the proposed methodology or experiments, derive theoretical or mathematical claims, or interpret the experimental results. Tasks related to proofs, translation, and qualitative data analysis were not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shristi Das Biswas, Arani Roy, and Kaushik Roy. Cure: Concept unlearning via orthogonal representation editing in diffusion models. In *NeurIPS*, 2026.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. CASteer: Cross-attention steering for controllable concept erasure. In *ICLR*, 2026a.

Tatiana Gaintseva, Andrew Stepanov, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Midsteer: Optimal affine framework for steering generative models. In *ICML*, 2026b.

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

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, 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.

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

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

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 2018.

Miso Kim, Georu Lee, Yunji Kim, Hoki Kim, Jinseong Park, and Woojin Lee. Co-occurring associated REtained concepts in diffusion unlearning. In *ICLR*, 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 *ICCV*, 2023.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *CVPR*, 2025.

Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. SPEED: Scalable, precise, and efficient concept erasure for diffusion models. In *ICLR*, 2026.

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

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications. In *CVPR*, 2024.

Andres Nava and Matthieu Wyart. Hierarchical concept geometry in language models emerges from word co-occurrence. *arXiv preprint arXiv:2605.23821*, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

Bedapudi Praneeth. NudeNet: Neural nets for nudity classification, detection and selective censoring. `https://github.com/notAI-tech/NudeNet`, 2019.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2025.

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

Shaswati Saha, Sourajit Saha, Manas Gaur, and Tejas Gokhale. Side effects of erasing concepts from diffusion models. In *EMNLP*, 2025.

Shaswati Saha, Rajasekhar Anguluri, and Manas Gaur. To erase, or not to erase: Robust training-free concept erasure with preservation aware adaptive ranked subspace expansion. *arXiv preprint arXiv:2607.23492*, 2026.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Zhuan Shi, Alireza Dehghanpour Farashah, Rik de Vries, and Golnoosh Farnadi. Neighbor-aware localized concept erasure in text-to-image diffusion models. In *CVPR*, 2026.

Yuhao Sun, Lingyun Yu, Haoxiang Xu, Fengyuan Miao, Zhuoer Xu, and Hongtao Xie. Orthogonal concept erasure for diffusion models. *arXiv preprint arXiv:2605.28902*, 2026.

Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *CVPR*, 2025.

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

Parth Upman, Nishita Jain, and Shreyank N. Gowda. Erasing without collateral damage: Precise concept removal in diffusion models. *arXiv preprint arXiv:2607.05274*, 2026.

Korada Sri Vardhana and Soma Biswas. Generase: Generalizable and semantically-aware concept erasure in diffusion models. In *CVPR*, 2026.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *CVPR*, 2025.

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.

A REPRESENTATION CONSTRUCTION AND GEOMETRY

We detail the spatial representations used by ExCISE and examine how whitening and the sampling step affect the parent–retain decomposition. Figures A.1 and A.2 provide the geometric diagnostics supporting Section 4.

A.1 SPATIAL STRUCTURE OF CROSS-ATTENTION OUTPUTS

Let $\boldsymbol{H}_{\ell,t} \in \mathbb{R}^{N_\ell \times d_\ell}$ be the spatial features and $\boldsymbol{C} \in \mathbb{R}^{L \times d_c}$ the text conditioning. For one attention head, $\boldsymbol{Q} = \boldsymbol{H}_{\ell,t}\boldsymbol{W}_Q$, $\boldsymbol{K} = \boldsymbol{C}\boldsymbol{W}_K$, and $\boldsymbol{V} = \boldsymbol{C}\boldsymbol{W}_V$. Its output is

$$\boldsymbol{O}_{\ell,t} = \boldsymbol{A}_{\ell,t}\boldsymbol{V}\boldsymbol{W}_O, \qquad \boldsymbol{A}_{\ell,t} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_h}}\right), \tag{A.1}$$

where d_h is the head dimension. Multi-head attention concatenates the head outputs before the output projection. The N_ℓ query positions remain spatially indexed, allowing a resized image mask to select the corresponding output features for pooling in Equation 2.

A.2 WHITENING AS A CHANGE OF METRIC

For a positive-definite covariance estimate $\boldsymbol{\Sigma}_\ell = \boldsymbol{U}_\ell \boldsymbol{\Lambda}_\ell \boldsymbol{U}_\ell^\top$, with orthonormal eigenvectors in $\boldsymbol{U}_\ell$ and positive eigenvalues in the diagonal matrix $\boldsymbol{\Lambda}_\ell$, ZCA whitening is the symmetric inverse square root (Kessy et al., 2018):

$$\boldsymbol{W}_\ell = \boldsymbol{U}_\ell \boldsymbol{\Lambda}_\ell^{-1/2} \boldsymbol{U}_\ell^\top, \qquad \boldsymbol{W}_\ell \boldsymbol{\Sigma}_\ell \boldsymbol{W}_\ell^\top = \boldsymbol{I}_{d_\ell}. \tag{A.2}$$

The transform rescales covariance eigen-directions by their inverse standard deviations. For target and component representations $\boldsymbol{x}$ and $\boldsymbol{r}$, the resulting inner product is

$$(\boldsymbol{W}_\ell \boldsymbol{x})^\top (\boldsymbol{W}_\ell \boldsymbol{r}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}_\ell^{-1} \boldsymbol{r}. \tag{A.3}$$

Whitening therefore changes the metric used for retain-subspace selection and projection. The inverse transform in Equation 4 returns the residual to activation coordinates for steering. Spatial centering has a separate role: it subtracts the image-wide mean before regional pooling.

Figure A.1 visualizes parent and component representations at five sampled U-Net layers. The merged parent is close to the retain span in raw coordinates, while whitening increases its angle to the span in these examples. This diagnostic illustrates why measuring alignment relative to activation variability changes the surplus construction.

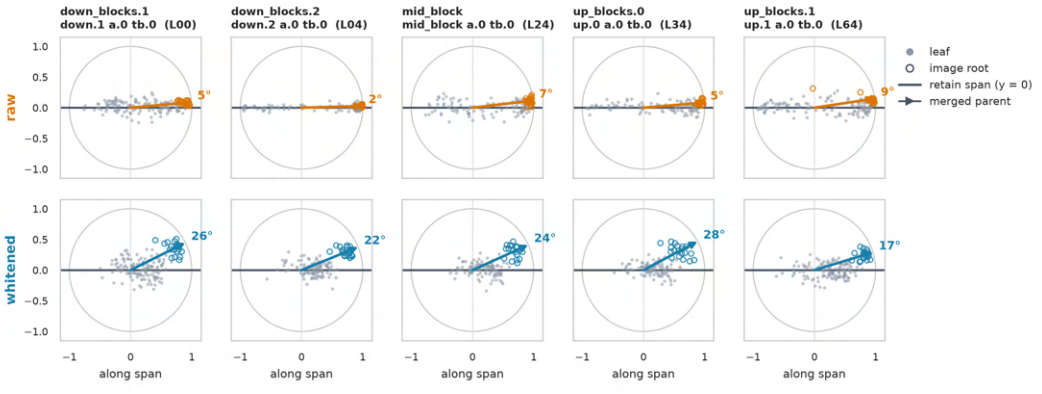
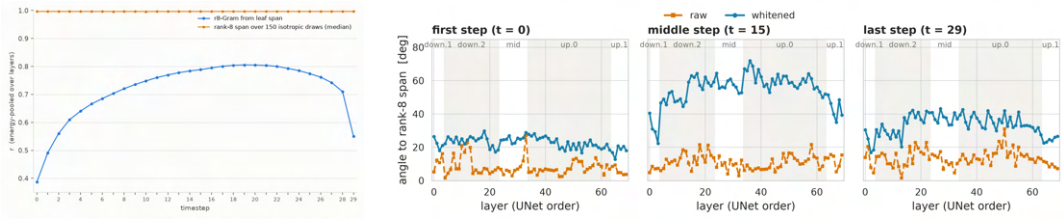


Figure A.1: **Effect of whitening on parent–component geometry.** Columns show five sampled U-Net layers; rows show raw and whitened representations. Filled points denote components, open circles denote per-image parents, and arrows denote the merged parent. The horizontal axis indicates the retain span in the displayed projection.



(a) (b)

Figure A.2: **Parent–retain-span geometry for Taylor Swift.** (a) Residual fraction pooled over layers at each sampling step, compared with the median over 150 matched-rank isotropic subspaces. The curve labeled r corresponds to ρ in Equation A.4. (b) Angle between the merged parent and the rank-eight retain span across layers, before and after whitening, at the first, middle, and last sampling steps.

A.3 Structured overlap with retained components

For a nonzero mean whitened target $\bar{\boldsymbol{x}}_{\ell,t}$, define the residual fraction

$$\rho_{\ell,t} = \frac{\left\| \left(\boldsymbol{I}_{d_\ell} - \boldsymbol{B}_{\ell,t} \boldsymbol{B}_{\ell,t}^\top \right) \bar{\boldsymbol{x}}_{\ell,t} \right\|_2}{\left\| \bar{\boldsymbol{x}}_{\ell,t} \right\|_2}, \qquad \rho_{\ell,t} \in [0,1]. \tag{A.4}$$

The retain span explains a fraction $1 - \rho_{\ell,t}^2$ of this representation's squared norm. For a uniformly random rank-k subspace in d_ℓ dimensions and a fixed target, rotational symmetry gives

$$\mathbb{E}[\rho_{\ell,t}^2] = 1 - \frac{k}{d_\ell}, \tag{A.5}$$

with root-mean-square residual $\sqrt{1 - k/d_\ell}$.

Figure A.2(a) compares the Taylor Swift component span with the median over 150 matched-rank isotropic subspaces. The component-derived span yields pooled residuals of 0.385–0.805, compared with 0.996–0.998 for the random subspaces; the theoretical root-mean-square reference is approximately 0.997. The lower residuals indicate structured parent–component overlap beyond what the subspace dimension alone would explain. The plotted median and theoretical root-mean-square reference are distinct summaries.

A.4 Motivation for first-step extraction

In Figure A.2(a), the pooled residual is 0.385 at the first sampling step and 0.805 mid-trajectory. Figure A.2(b) also shows a narrower range of whitened parent–span angles at the first step ($13°$–$29°$, versus $22°$–$72°$ mid-trajectory). Thus, the first-step retain span explains more of the parent representation in this case study, motivating the extraction step used by ExCISE.

The reported cross-image agreement of the parent representation is also higher at the first step (0.72, versus 0.45 mid-trajectory). This complements the geometric comparison, but is a separate statistic from the angles and residuals shown in Figure A.2.

A.5 Role of the retain rank

The surplus depends on how much component structure the selected subspace retains. If disjoint component masks fully partition the parent, mean pooling expresses the parent as an area-weighted combination of its components. Shared spatial centering and invertible whitening preserve this linear relation, placing the mean parent in the full span of the collected components. Rank truncation therefore determines which component modes constrain the target and which remain in the surplus.

The reported component effective ranks of 15–28 motivate retaining a subset of this variation with $k = 8$. Section 5.4 examines the empirical erasure–preservation trade-off using the rank comparison in Figure B.1. The residual is defined relative to this selected subspace; semantic preservation is measured on generated images.

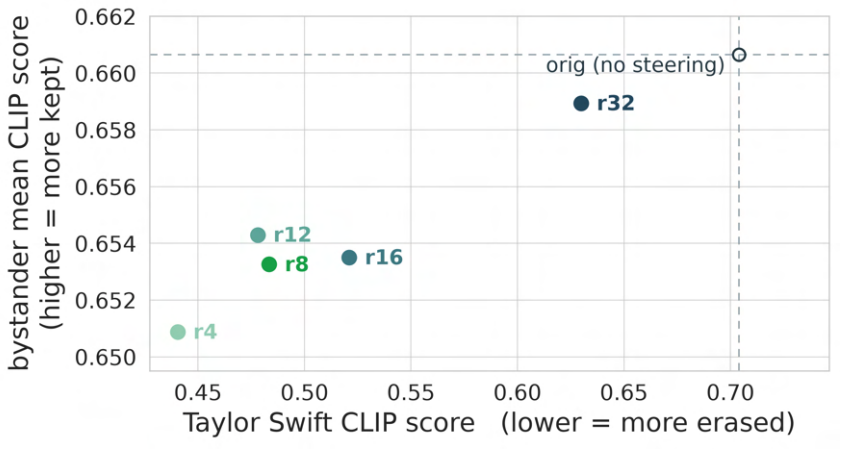


Figure B.1: **Direction-construction and rank comparison for Taylor Swift erasure.** The horizontal axis gives normalized target CLIP score; the vertical axis gives mean normalized CLIP score over eight related concepts. The comparison includes the parent-only direction, CASteer, averaged per-image surplus directions, and shared retain spans of ranks four, eight, and twelve.

B Direction construction and rank analysis

Figure B.1 evaluates Taylor Swift erasure alongside eight related concepts. Target and mean neighboring-concept CLIP scores are normalized by their respective unedited baselines. Lower target scores indicate greater suppression under this metric, while higher neighboring-concept scores indicate greater preservation of text–image similarity.

Sharing component structure across instances. The shared retain-span construction attains higher neighboring-concept similarity than averaging independently constructed per-image surplus directions at a similar target score. It also improves neighboring-concept similarity relative to CASteer at a similar target score. The parent-only direction suppresses the target more strongly but causes a larger decrease in neighboring-concept similarity. These comparisons support refining the target against component structure shared across reference images.

Choosing the retain rank. All three evaluated ranks retain more neighboring-concept similarity than the parent-only direction. Rank four gives stronger target suppression but lower preservation than ranks eight and twelve. The results show that rank changes the erasure–preservation balance; they do not establish rank eight as universally optimal. We use rank eight as the fixed setting in the reported SD experiments.

C Concept Erasure on Flux.1 dev

C.1 Flux implementation details

ExCISE collects and steers activations in the cross-attention output space. Flux.1-dev and other models in its family (Black Forest Labs, 2024) do not contain such cross-attention layers. Flux-specific erasure strategies have therefore focused on identifying where in FLUX one can obtain similar erasure control to that enabled by cross-attention editing and steering in SD v1.4 (Rombach et al., 2022) and SDXL (Podell et al., 2024). EraseAnything (Vardhana & Biswas, 2026) finds that FLUX exhibits cross-attention-like linear relationships between text embeddings and attention maps. EraseAnything edits W_{attn} within a fine-tuning framework. In contrast, we act on $W_{\text{attn}}V_{\text{text}}$, the image-side output of the double-stream blocks, where spatial positions are preserved across layers. This makes our method transferable. Following EraseAnything (Gao et al., 2025), we steer only the last double-attention block, but we do so at all timesteps for smoother guidance. For objects and characters, we use one-step steering, as the required drift is larger.

Table C.1: Celebrity erasure on FLUX.1 dev, averaged over 3 erasure concepts (Barack Obama, Donald Trump and Queen Elizabeth II) that show over 60% giphy detection on the unsteered images generated by Flux.1 dev. Acc_e giphy identification rate over erasure concepts, Acc_s over preserved concepts, and H their Harmonic mean. the best value per column is in bold and the second best underlined.

Method	**Acc**$_e$ ↓	**Acc**$_s$ ↑	H ↑
Orig (unsteered)	86.40	71.67	17.51
EraseAnything	32.40	19.44	15.66
ESD	<u>15.73</u>	32.78	44.02
OCE	21.73	**68.89**	<u>70.30</u>
Ours	**1.33**	<u>56.11</u>	**71.32**

C.2 Flux on Identity

Following the evaluations we conducted on SDXL (Podell et al., 2024) we evaluate our erasure of celebrities in Flux.1 dev (Black Forest Labs, 2024) across a subset of MACE celebrity benchmark (Lu et al., 2024). As Flux models are purposefully not trained on large sets of celebrity data to avoid misuse of identity, we only evaluate on a subset of celebrities Flux can generate reliably. Table C.1 shows quantitaive results across EraseAnything (Gao et al., 2025), ESD (Gandikota et al., 2023) and OCE (Sun et al., 2026). Most inference-time methods we based our evaluation on for SDXL do not have a publicly available implementation for erasure on Flux, making a matched comparison not possible.

Qualitative results on identity erasure under regular, pairwise, and rephrased prompts can be found in Figure C.1. Additional Qulaitative results across erasure tasks are presented in Figure C.2 and C.3.

D Multi-Concept Erasure

While multi-concept erasure in many inference-time methods introduces additional cost through gating, our gate-free pipeline allows multiple vectors to be stacked and applied jointly at inference time. Rather than simply stacking independently constructed per-concept vectors, we generate one vector for each erased concept by orthogonalizing it against a combined retain span constructed from the leaves of all concepts, effectively expanding the retain span with each added concept.

D.1 Implementation

We test whether adding additional erasure concepts harms the erasure quality of the initially erased concept. We compare a setting with one target concept erased against settings with two or three erased concepts. Consistent with our previous analysis, the performance of different rank settings is determined by the growth of the retain span. With two or more erased concepts, rank 8 lacks the capacity to effectively refine the parent directions. Moreover, as the retain span grows, the remaining surplus per concept shrinks, resulting in the need for increased erasure strength. The evaluation is presented in Table D.1

E Concept Switching

An additional benefit of the implementation of our surplus direction as cross-attention output steering is the ability to switch a target concept to another. To directly map our direction onto the concept we want to generate under the target prompt after switching, we collect the same pooled-parent-mask representations without the parts for the switch concept. In our framework, concept switching is performed by removing the source-specific component while adding the substitute-specific component

Table D.1: Multi-concept erasure. Acc_e is the erased identity still identified (lower is better), Acc_s a preserved identity identified (higher is better), and H the harmonic mean of $1-\text{Acc}_e$ and Acc_s. Rank is the dimension of the retain span. Erasure is complete in every configuration; the retain span is shared across the jointly erased identities.

Erased	**Rank**	**Acc**$_e$ ↓	**Acc**$_s$ ↑	H ↑
none (unsteered)	—	96.33	95.00	7.07
Taylor	8	0.00	92.00	**95.83**
Taylor + Emma	8	0.00	87.00	93.05
Taylor + Emma	16	0.00	91.00	**95.29**
Taylor + Emma + Elon	8	0.00	87.00	93.05
Taylor + Emma + Elon	16	0.00	89.00	94.18
Taylor + Emma + Elon	24	0.00	88.00	93.62
Taylor + Emma + Elon	32	0.00	90.00	**94.74**

within the same source retain geometry:

$$v_{\text{switch}} = w_{\text{sub}} \left(p_{\text{sub}} - P_{S_{\text{src}}} p_{\text{sub}}\right) - \left(p_{\text{src}} - P_{S_{\text{src}}} p_{\text{src}}\right), \tag{E.1}$$

where S_{src} denotes the retain span constructed from the source concept's visual components. The substitute weight w_{sub} controls how strongly the representation is steered toward the substitute concept relative to the removal of the source concept.

F Style Erasure

Art-style erasure does not fall into either of the two erasure settings. Style is not defined by a spatial hierarchy, but rather by attributes distributed across image regions, including texture, brush strokes, and color palette. To erase a style with our method, we generate a set of randomly selected objects ("goldfish" "sea anemone" "banana" "acoustic guitar" "lighthouse" "hot air balloon" "espresso maker" "mountain bike" "sundial" "teapot" "throne" "umbrella" "vase" "wooden spoon" "pineapple" "grand piano" "sax" "steam locomotive" "castle" "broccoli") once without and once with the target style, following CASteer's contrastive prompting. We then pool the activations over the object region for both runs and use the non-styled representation span as our retain span. As the two representations share less component structure, leaving a larger residual, we reduce the steering strength to 1.0. Visual results on SDXL (Podell et al., 2024) are presented in Figure F.1. *Van Gogh style* brush strokes, swirls and color palettes are removed while the composition of the image stays intact and neighboring concepts are preserved.



Figure C.1: Erasure results on different identities that are reliably generateable with with Flux.1 dev. *Top*: erasure strength of *Donald Trump* under a prompt containing "The President of the United States of America". *Middle*: Within-Image preservation under two seeds with *Donald Trump* erased and two with *Queen Elizabeth*. *Bottom*: Single identity erasure of *Barack Obama* and *Taylor Swift*



Figure C.2: Qualitative results for part-erasure and garbage truck object erasure on Flux.1 dev as well as car preservation performance under Garbage truck erasure



Figure C.3: Additional Qualitative results on erasure of *Mickey Mouse*, *a chainsaw*, *Van Gogh style*, and *Pikachu*.

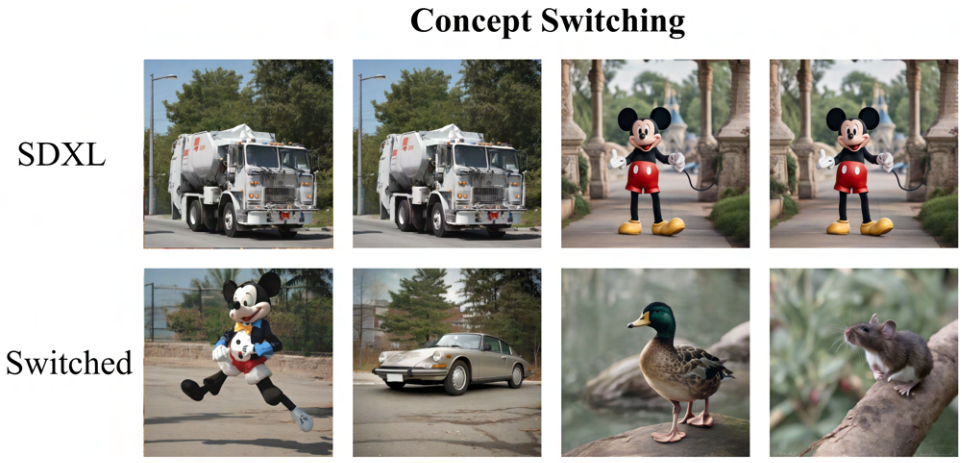


Figure E.1: Qualitative results on concept switching on SDXL. *garbage truck* is switched to *Mickey Mouse* and *car*. *Mickey Mouse* is switched to *duck* and *mouse*. Each four images are generated under the same prompt and seed.



Figure F.1: Qualitative results on style erasure with *Van Gogh* erased (first two columns) and Monet retained.