NuST: Null-Space-Constrained Activation Steering for Robust Concept Erasure in Diffusion Models

Anonymous authors
Paper under double-blind review

Abstract

Text-to-image diffusion models can reproduce undesirable concepts inherited from large-scale web data, raising safety, copyright, and privacy concerns in real-world deployment. Existing concept-erasure methods typically fine-tune or edit model parameters, which can be computationally expensive and may inadvertently impair unrelated generation capabilities. Although activation steering provides a parameter-preserving alternative, unconstrained interventions can still perturb features shared by target and non-target concepts. We introduce NuST, a null-space-constrained activation-steering framework for robust concept erasure in diffusion models. At each selected module-timestep site, NuST constructs target-to-anchor shifts from paired prompts and learns a channel-wise transformation whose effective directions are restricted to the null space of retained benign activations. This design suppresses unintended changes to non-target representations while redirecting target activations toward safe anchor semantics. The transformations are obtained through closed-form regularized regression, requiring neither gradient-based optimization nor model-parameter updates. We evaluate NuST on nudity, artistic-style, and object erasure using Stable Diffusion v1.5. NuST reduces the average attack success rate for nudity recovery to 0.009% across five standard and adversarial benchmarks, achieves zero target-style classification accuracy for three artists, and lowers the average target-object success rate to 0.34%, while maintaining competitive generation quality and text-image alignment. These results establish null-space activation steering as an effective and robust alternative to parameter-based concept erasure.

1 Introduction

Text-to-image (T2I) diffusion models have achieved remarkable progress in generating high-fidelity images that closely follow complex textual instructions (Ho et al., 2020; Dhariwal & Nichol, 2021; Ho & Salimans, 2022b; Nichol et al., 2022; Rombach et al., 2022; Saharia et al., 2022). However, the same generative capacity also enables these models to reproduce undesirable concepts inherited from large-scale, weakly curated web data, including sexually explicit content, copyrighted artistic styles, private information, and other harmful imagery (Schramowski et al., 2023; Yang et al., 2024; Zhang et al., 2025; Carlini et al., 2023; Cui et al., 2025). Such behavior raises substantial ethical, legal, and safety concerns for real-world deployment. Retraining a model after filtering its training data offers a conceptually straightforward solution (Nichol et al., 2022; Rombach et al., 2021; Schramowski et al., 2023), but its prohibitive computational and data costs make it impractical for modern generative models. Consequently, concept erasure, which aims to selectively suppress a pretrained model's ability to generate designated concepts while retaining its remaining capabilities, has emerged as an important problem in controllable and trustworthy generation.

Existing concept-erasure methods predominantly operate by modifying model parameters. Fine-tuning-based approaches optimize the diffusion model with carefully designed erasure and preservation objectives (Gandikota et al., 2023; Ruidong et al., 2025; Lu et al., 2024; Kim et al., 2025; Li et al., 2024b; Cywiński & Deja, 2025), whereas editing-based approaches directly update selected attention or feed-forward parameters, often through closed-form or lightweight optimization (Gandikota et al., 2024; Chavhan et al., 2025; Wang et al., 2025). Despite their effectiveness,

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.

both paradigms alter parameters shared across many concepts. As a result, stronger erasure may cause collateral degradation of unrelated concepts, while conservative updates can leave the erased concept recoverable through alternative prompts or adversarial attacks. Recent activation-space methods provide a parameter-preserving alternative by manipulating intermediate representations during inference time (Rodriguez et al., 2024; Li et al., 2024a; Gaintseva et al., 2025). Nevertheless, existing steering interventions are typically constructed from target-specific directions without explicitly protecting representation components shared with benign concepts. This limitation leaves a fundamental question unresolved: *how can target representations be strongly redirected without perturbing the activation subspace required for non-target generation?*

We address this question from a null-space perspective. Our key observation is that an activation intervention can preserve a set of benign representations if its induced perturbation vanishes on their activation subspace. Meanwhile, the same intervention can remain effective on target concepts when their representations contain components outside that retained subspace. Based on this observation, we propose NuST, a null-space-constrained activation-steering framework for diffusion concept erasure. Instead of updating model parameters, NuST intervenes in the intermediate feed-forward network (FFN) activations of the denoising U-Net. Because diffusion representations evolve throughout the denoising trajectory, we treat each FFN module and scheduler timestep as a distinct intervention site and learn a site-specific steering transformation.

At each intervention site, NuST first collects activations from a benign retain set and constructs an orthogonal projector onto their approximate right null space. It then extracts desired semantic shifts from aligned target-anchor prompt pairs—for example, from "a sunflower field in the style of Van Gogh" to "a sunflower field"—and learns a channel-wise linear transformation through regularized regression. Projecting this transformation into the retained null space suppresses its response to benign activations, while the regression objective redirects target representations toward their corresponding anchor semantics. During inference, the resulting transformation is applied to every spatial token of the selected FFN activation and added residually to the original representation. This procedure requires no gradient-based optimization or modification of the pretrained model parameters, yet provides an explicit mechanism for balancing erasure strength and non-target preservation.

We evaluate NuST on three representative concept-erasure settings: nudity, artistic styles, and ImageNette objects. Beyond standard prompts, we examine robustness against four concept-recovery attacks covering black-box prompt manipulation and white-box embedding optimization. On nudity erasure, NuST reduces the average attack success rate to 0.009% across I2P and four adversarial benchmarks, outperforming both training-based and parameter-editing baselines. It also suppresses all three evaluated artistic styles to zero classification accuracy and achieves an average target-object success rate of 0.34%. Across these settings, NuST retains competitive FID and CLIP scores, demonstrating that null-space-constrained activation steering can provide strong erasure and adversarial robustness without modifying the underlying model.

Our main contributions are summarized as follows:

- We introduce **NuST**, a null-space-constrained activation-steering framework for diffusion concept erasure. By restricting the effective steering transformation to the null space of retained benign activations, NuST explicitly suppresses unintended perturbations to non-target representations while enabling strong intervention on target concepts.

- We develop a diffusion-specific, timestep-aware steering mechanism that learns site-specific channel transformations from aligned target-anchor prompt pairs. By treating each FFN module and denoising timestep as an intervention site, NuST captures the temporal evolution of diffusion representations and redirects target activations toward safe anchor semantics without gradient-based optimization or model-parameter updates.

- We conduct extensive evaluations across nudity, artistic-style, and object erasure, including four adversarial concept-recovery attacks. NuST achieves an average nudity-recovery rate of only **0.009%**, zero target-style accuracy for all three evaluated artists, and an average target-object success rate of **0.34%**, while retaining competitive generation quality and text-image alignment.

2 Related Works

Concept Erasure in Text-to-Image Models. Existing concept erasure methods have primarily relied on optimization-based fine-tuning and direct model editing, while recent studies increasingly explore training-free or hybrid interventions. Early approaches, including ESD, Concept Ablation, and Receler (Gandikota et al., 2023; Kumari et al., 2023; Huang et al., 2023), suppress undesired concepts by fine-tuning diffusion models with tailored erasure objectives. Adversarially robust variants, such as CPE, AdvUnlearn, RACE, STEREO, and TRCE (Ruidong et al., 2025; Zhang et al., 2024; Lee et al., 2025; Srivatsan et al., 2024), further improve robustness against adversarial prompts or implicitly expressed malicious concepts.

To reduce optimization cost, another line of work performs direct parameter editing. UCE and RECE (Gandikota et al., 2024; Gong et al., 2024) derive closed-form updates to cross-attention parameters, while MACE (Lu et al., 2024) targets large-scale multi-concept removal. SPEED (Li et al., 2026) introduces null-space-constrained parameter editing for prior preservation, and ICE (Lin et al., 2025) further incorporates orthogonal decomposition and invariant constraints. Concept-Prune(Chavhan et al., 2025) takes a different route by identifying concept-specific skilled neurons in feed-forward layers and directly pruning the corresponding weights, enabling training-free concept removal through sparse weight intervention.

More recently, methods have begun to intervene directly in internal representations during inference. AdaVD(Wang et al., 2025) manipulates target semantics in the cross-attention value space through orthogonal-complement decomposition, whereas CASteer(Gaintseva et al., 2025) constructs steering directions from contrastive cross-attention activations and applies them during inference.

Null-space Preservation and Activation Space Manipulation. Null-space constraints have been explored as a principled mechanism for enabling targeted modification while preserving previously learned information. AlphaEdit (Fang et al., 2025) constrains LLM parameter updates to the null space of retained knowledge, while AlphaSteer (Sheng et al., 2026) extends this idea to activation steering by learning transformations that suppress changes on benign activations. In diffusion models, related studies have demonstrated that intermediate representations contain manipulable semantic directions. SDID (Li et al., 2024a) discovers concept-specific directions in U-Net bottleneck features, while SAeUron (Cywiński & Deja, 2025) decomposes diffusion activations into sparse interpretable features that can be selectively suppressed. More general activation-steering frameworks such as AcT (Rodriguez et al., 2024) further demonstrate that semantic concepts can be controlled by manipulating intermediate activations at inference time.

Our method combines these two lines of research: NuST adapts null-space-constrained activation steering into diffusion models, directly manipulating intermediate FFN representations without modifying model parameters, thus providing a direct mechanism for balancing concept erasure and non-target preservation.

3 Methods

3.1 Principled and learnable activation steering

We intervene at the input of the second linear projection in FFN, i.e., the pre-linear2 activation. Unless otherwise specified, all experiments in this paper use the pre-linear2 activation of FFN modules as the steering target. For an FFN layer l at diffusion timestep t (a combination of the FFN layer and denoising timestep is referred to as a **site** later in this paper), the corresponding activation inside the U-Net is originally represented as (assuming the batch size is 1):

$$\mathbf{A}^{(l,t)} \in \mathbb{R}^{N \times C} \tag{1}$$

where C is the number of channels, and $N = H \times W$ is the number of spatial tokens per channel. Importantly, t corresponds to the scheduler timestep rather than the denoising iteration index. For example, with a 50-step sampling process, the activation sites are indexed by the actual timesteps $\{981, 961, ..., 1\}$.

To ensure principled preservation during the process of steering target-concept representations, we introduce a learnable activation transformation $\Delta = \mathbf{P}\tilde{\Delta} \in \mathbb{R}^{C \times C}$ that incorporates closed-form orthogonal null-space projection.

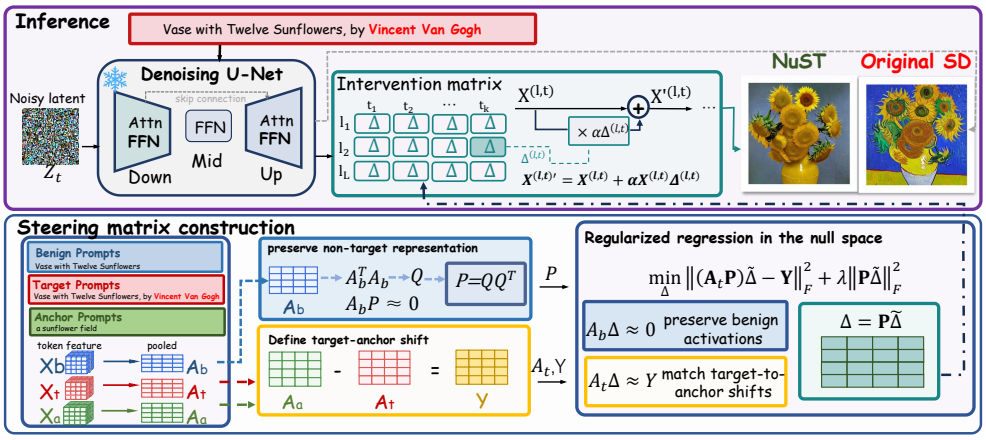


Figure 1: Main pipeline of NuST. (Bottom) At each selected U-Net layer l and denoising timestep t, we collect and pool activations from benign, target, and anchor prompts. Benign activations define a null-space projector $\mathbf{P}^{(l,t)}$, while the difference between anchor and target activations defines the desired semantic shift. Regularized regression in this null space yields a steering matrix $\Delta^{(l,t)}$ that moves target activations toward anchor activations while preserving benign activations. (Top) During inference, NuST applies the learned steering transformation to the original activation $\mathbf{X}^{(l,t)}$ at each selected site: $\mathbf{X}'^{(l,t)} = \mathbf{X}^{(l,t)} + \alpha \mathbf{X}^{(l,t)} \Delta^{(l,t)}$, where α controls the steering strength.

Specifically, the transformation matrix $\Delta^{(l,t)}$ is constructed to satisfy the following two core objectives: non-target preservation and erasure efficacy improvement.

- Non-target preservation. For benign concepts, the activations should remain unaffected after steering.
- Erasure efficacy improvement. For target concepts, the activations should be steered toward anchor directions.

By combining the null-space preservation constraint with target-to-anchor regression, the resulting $\Delta^{(l,t)}$ selectively induces substantial shifts for target representations while suppressing unintended changes to benign ones. We will dive into how to achieve these learning objectives in the following two sections. For detailed mathematical derivation, please refer to Appendix B.

Given an activation tensor $\mathbf{X}^{(l,t)} \in \mathbb{R}^{B \times N \times C}$, the application process can be formulated as follows:

$$\mathbf{X}'^{(l,t)} = \mathbf{X}^{(l,t)} + \alpha \mathbf{X}^{(l,t)} \Delta^{(l,t)} \tag{2}$$

where the transformation is applied along the channel dimension and α controls the steering strength.

For notational simplicity, we omit all superscripts and subscripts in the following discussions.

3.2 Non-target preservation with closed-form orthogonal null-space projection

For concept erasure, the retain set consists of benign concepts whose generation capability should be preserved. Since our steering transformations are learned from spatially averaged channel representations, let $\mathbf{a}_b^{(i)} \in \mathbb{R}^{1 \times C}$ denote the pooled activation[1] extracted from the i-th benign prompt at a given intervention site. We require the steering transformation to produce no semantic shift for each retained representation:

$$\mathbf{a}_b^{(i)} \Delta = \mathbf{0}_{1 \times C} \tag{3}$$

Stacking M_b benign representations row-wise gives the retain activation matrix:

$$\mathbf{A}_b = \left[\left(\mathbf{a}_b^{(1)}\right)^\top, \ldots, \left(\mathbf{a}_b^{(M_b)}\right)^\top \right]^\top \in \mathbb{R}^{M_b \times C}. \tag{4}$$

[1]Here, "pooled" means spatially averaged. We will explain it in AppendixB.

The non-target preservation objective can therefore be written compactly as:

$$\mathbf{A}_b\Delta = \mathbf{0}_{M_b \times C} \tag{5}$$

Under our row-wise activation convention, this constraint naturally requires the columns of Δ to lie in the null space of $\mathbf{A}_b$, defined as $Null(\mathbf{A}_b) = \{\mathbf{z} \in \mathbb{R}^C | \mathbf{A}_b\mathbf{z} = \mathbf{0}\}$.

To construct the corresponding orthogonal projector, we perform singular value decomposition (SVD) on the non-central covariance matrix:

$\mathbf{A}_b^\top \mathbf{A}_b = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\{\mathbf{V}, \mathbf{\Lambda}, \mathbf{V}^\top\} = \text{SVD}(\mathbf{A}_b^\top \mathbf{A}_b)$.

Here $\mathbf{V} \in \mathbb{R}^{C\times C}$ is the orthonormal eigenvector matrix of $\mathbf{A}_b^\top \mathbf{A}_b$, where each column corresponds to an eigenvector, and Λ is a diagonal matrix of its singular values.

Let $\mathbf{Q} \in \mathbb{R}^{C\times n}$ collect n eigenvectors with zero eigenvalues. These vectors span the selected null space(Dieudonné, 1969) of $\mathbf{A}_b$. With the above definition, the null-space projection matrix is calculated as:

$$\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top \in \mathbb{R}^{C\times C} \tag{6}$$

By construction, $\mathbf{A}_b\mathbf{P} = \mathbf{0}$. Therefore, the final transformation $\mathbf{P}\tilde{\Delta}$ projects $\tilde{\Delta}$ onto the null-space of $\mathbf{A}_b$ since $\mathbf{A}_b\Delta = \mathbf{A}_b\mathbf{P}\tilde{\Delta} = \mathbf{0}$.

3.3 Target-to-Anchor Steering

Having constructed the benign null-space projector $\mathbf{P}$, we next learn a transformation that redirects target-concept representations toward their corresponding anchor semantics. At a given intervention site, let $\mathbf{A}_t, \mathbf{A}_a \in \mathbb{R}^{M\times C}$ denote the spatially averaged activation matrices[2] collected from M aligned target–anchor prompt pairs, where each row corresponds to one prompt and C is the channel dimension. Specifically, the i-th rows $\mathbf{a}_t^{(i)}$ and $\mathbf{a}_a^{(i)}$ represent the activations of the i-th target and anchor prompts, respectively.

We define the semantic steering direction for each pair as $\mathbf{y}^{(i)} = \mathbf{a}_a^{(i)} - \mathbf{a}_t^{(i)}$. Stacking all pairwise steering directions row-wise yields:

$$\mathbf{Y} = \mathbf{A}_a - \mathbf{A}_t \in \mathbb{R}^{M\times C} \tag{7}$$

We also consider a mean-direction variant, where $\bar{\mathbf{y}} = \frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{a}_a^{(i)} - \mathbf{a}_t^{(i)}\right) \in \mathbb{R}^{1\times C}$, and the same direction is repeated for all target samples:

$$\mathbf{Y} = \mathbf{1}_M\bar{\mathbf{y}} \in \mathbb{R}^{M\times C} \tag{8}$$

Ideally, $\mathbf{Y}$ specifies the desired activation shifts from target representations toward their corresponding anchors. The learned steering transformation should satisfy $\mathbf{A}_t\Delta = \mathbf{Y}$, such that the steered target activations approach their anchor counterparts: $\mathbf{A}_t + \mathbf{A}_t\Delta = \mathbf{A}_a$.

Finally, we solve:

$$\tilde{\Delta}^* = \arg\min_{\tilde{\Delta}} \| (\mathbf{A}_t\mathbf{P})\,\tilde{\Delta} - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{P}\tilde{\Delta}\|_F^2 \tag{9}$$

where the second term is the regularization term to avoid overfitting, and λ is a hyperparameter to control the strength of the regularization penalty.

This objective allows flexible updates of inner activations such that target concepts are effectively steered from the malicious representation direction to a safe anchor direction. In contrast, representations of non-target concepts remain unaffected.

4 Experiments

4.1 Experimental Setup

In this section, we evaluate our method over three representative concept erasure tasks: nudity erasure, style erasure, and object erasure. The compared baselines include training-based methods

[2]We will explain it in AppendixB.

ESD, STEREO, TRCE, and training-free methods UCE, Concept-prune, AdaVD, and ActErase, which have achieved SOTA performance across various concept erasure tasks. In implementation, we conduct all experiments on SDv1.5(Rombach et al., 2021) and generate each image over 50 sampling steps with classifier-free guidance(Ho & Salimans, 2022a) of 7.5. Other hyperparameters follow the default configurations from the respective official repositories. More implementation details and compared baselines can be found in Appendix A.1.

As discussed in Section 3.1, we intervene on the pre-linear2 activations within the FFN layers of the transformer blocks. The optimal hyperparameters, including the regularization coefficient λ and the steering strength α, are determined via grid search. We defer the analysis of alternative intervention positions and the sensitivity to hyperparameter selections to the ablation studies in Section 4.5 and Appendix C.

4.2 NUDITY ERASURE

We conduct experiments on the Inappropriate Image Prompt (I2P) benchmark (Schramowski et al., 2023), which comprises various types of inappropriate prompts involving violence, illegal activity, and nudity. To evaluate adversarial robustness, we further introduce four adversarial attack benchmarks, including three black-box benchmarks (MMA (Yang et al., 2024), Ring-A-Bell (Tsai et al., 2024), Prompting4Debugging (P4D) (Chin et al., 2024), and one white-box benchmark, UnlearnDiff(Zhang et al., 2025). During evaluation, we use NudeNet(Bedapudi, 2019) with a threshold of 0.6 to detect nude content and report the number of images containing exposed parts, namely Attack Success Rate (ASR). For I2P, we also report the number of detected exposed body parts for each category and their total count.

Table 1 shows NudeNet detections on the I2P benchmark. Our method achieves the lowest total detection count among the compared methods, indicating that activation-space intervention effectively suppresses nudity-related generation.

Table 1: Quantitative results of explicit content detected using the Nudenet detector at the threshold of 0.6 on the I2P benchmark. **F**: Female. **M**: Male. Best results are marked in **bold**, and second-best results are marked underline.

Method	Buttocks	Breasts (F)	Genitalia (F)	Breasts (M)	Genitalia (M)	Feet	Armpits	Belly	Total
SD v1.5	22	223	9	24	9	18	76	122	503
ESD	2	27	3	8	2	19	32	30	123
UCE	7	35	5	11	4	29	29	62	182
Concept-prune	5	12	0	1	2	1	8	4	33
AdaVD	25	258	13	20	5	27	119	115	582
STEREO	1	1	0	0	3	0	1	3	9
TRCE	1	3	1	0	5	3	2	3	18
ActErase	0	2	0	0	1	0	0	3	6
Ours	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**2**

As presented in Table 2, our method achieves the best performance under adversarial attacks with a zero nudity content detection rate except for ASR under I2P prompts, which is also the lowest compared with baselines. We further evaluate general generation capability using 10,000 prompts sampled from MS-COCO captions. Our method maintains a CLIP score close to that of the original SDv1.5 model, and a low FID score, indicating limited degradation in text-to-image alignment and better preservation of the generated-image distribution.

4.3 ARTISTIC STYLES ERASURE

For the artistic style erasure task, we evaluate the erasure performance across three distinct artists: Vincent Van Gogh, Leonardo Da Vinci, and Pablo Picasso. For each artist, we use 50 prompts based on their authentic painting titles. For the detailed prompt template, please refer to Appendix A.3.

Subsequently, we classify the generated images using the style classifier proposed in UnlearnDiff(Zhang et al., 2025) and report the Top-1 classification accuracy. Furthermore, to assess overall generation quality, we use 10,000 prompts from the MS-COCO dataset and compute both CLIP and FID scores. As shown in Table 3, our method achieves 0% target-style classification accuracy. The CLIP and FID scores are also competitive with those of the original SD v1.5 model. This indicates

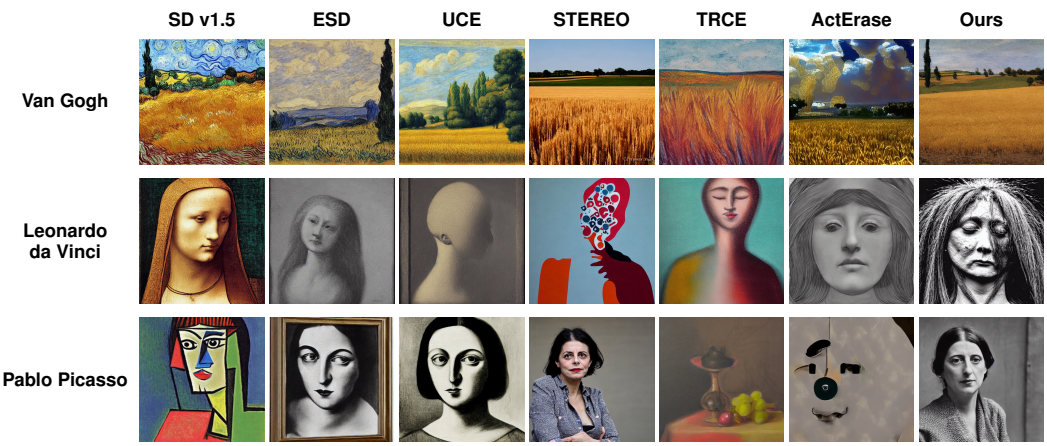


Figure 2: Qualitative comparison of artistic style erasure. Images in each row use the same prompt, random seed, and generation settings.

Table 2: Attack Success Rate(ASR) against adversarial attacks, CLIP score, and FID score for erasing the NSFW concept 'nudity'. The average ASR is averaged over datasets. Best results are marked in **bold**. Our method achieves the best average ASR.

Method	I2P(%)	MMA(%)	Ring-A-Bell(%)	P4D(%)	UnDiff(%)	Average(%)	FID (↓)	CLIP (↑)
SD v1.5	6.00	59.3	91.23	74.83	47.89	55.85	18.16	31.36
ESD	34.920	53.73	58.74	62.38	27.18	47.390	16.31	30.94
UCE	63.170	66.48	69.94	94.17	50.49	68.850	17.96	**31.30**
Concept-prune	8.890	3.28	18.71	9.71	0.99	8.316	21.99	31.09
AdaVD	98.410	99.27	99.25	100.00	99.01	99.188	19.67	**31.30**
STEREO	1.270	1.46	0.76	3.88	**0.00**	1.474	29.35	24.45
TRCE	5.340	1.46	**0.00**	**0.00**	1.98	1.756	**16.05**	30.89
ActErase	1.900	**0.00**	0.74	0.06	**0.00**	0.540	22.87	30.32
Ours	**0.043**	**0.00**	**0.00**	**0.00**	**0.00**	**0.009**	21.24	30.37

that our approach effectively preserves the pre-trained data distribution with minimal degradation, validating the non-destructive nature of our null-space design. Figure 2 shows the qualitative comparison among the methods.

4.4 Object Erasure

We evaluate object-level concept erasure across ten ImageNette categories following the protocol of ConceptPrune. For each target category, 5000 images are generated and evaluated using the ResNet-50 classifier (He et al., 2016). We report the attack success rate on the erased category (ASR_e, lower is better) and the average top-1 accuracy over the remaining nine categories (ASR_k, higher is better).

As shown in Table 4, our method achieves consistently strong target suppression across all ten object categories, with an average ASR_e of 0.34%. The target accuracy is close to zero for nearly all categories, indicating stable erasure efficacy across different object concepts. For non-target preservation, our method achieves an average ASR_k of 69.41%, showing that most non-target recognition capability is retained after intervention, although preservation performance remains category-dependent. In particular, erasing Church leads to the largest degradation in non-target accuracy, while English Springer, Golf Ball, Parachute, and Tench retain comparatively higher preservation performance.

Different baselines exhibit distinct erasure–preservation trade-offs. STEREO achieves an average ASR_e of 0.00%, but its average ASR_k drops to 46.30%, suggesting that its strong erasure comes with greater degradation in non-target generation. TRCE provides strong preservation, achieving the highest average ASR_k of 85.02%, and performs effective erasure on most categories; however, its efficacy is highly category-dependent, with ASR_e increasing to 71.52% when erasing Church. AdaVD also maintains relatively high non-target accuracy, but its erasure effectiveness varies con-

Table 3: Comparison of artistic style erasure performance. Top-1 denotes the classification accuracy of generated images under the style classifier, where lower is better. Lower FID and higher CLIP indicate better preservation of generation quality and text-image alignment. Best Results are marked in **bold**.

Method	Erase Van Gogh			Erase Da Vinci			Erase Picasso		
	Top-1	FID (↓)	CLIP (↑)	Top-1	FID (↓)	CLIP (↑)	Top-1	FID (↓)	CLIP (↑)
SD v1.5	0.78	18.16	31.36	0.00	18.16	31.36	0.30	18.16	31.36
ESD	0.36	**16.15**	30.98	0.00	**15.63**	31.07	0.00	**16.27**	31.07
UCE	0.02	17.93	**31.40**	0.00	18.23	31.32	0.02	18.41	31.33
Concept-prune	0.00	24.55	30.86	0.00	22.36	30.94	0.00	24.37	30.98
STEREO	0.00	25.81	29.84	0.00	24.23	30.14	0.00	24.23	29.86
AdaVD	0.02	19.65	31.29	0.00	19.68	31.29	0.00	19.68	31.29
TRCE	0.00	16.37	31.34	0.00	17.00	31.32	0.00	16.57	**31.37**
ActErase	0.00	21.01	30.57	0.00	18.02	31.13	0.00	18.16	31.36
Ours	**0.00**	17.83	31.34	**0.00**	17.15	**31.36**	**0.00**	17.37	**31.37**

Table 4: Object erasure performance across ten categories. ASR_e measures target accuracy after erasure (↓), and ASR_k measures average non-target accuracy over the remaining nine categories (↑). **SD1.5 (Ref.)** denotes the original model's per-class accuracy. We mark the best and second-best results among erasure methods in **bold** and underline, respectively.

Classes	ASR_e(%) ↓								ASR_k(%) ↑							
	ESD	UCE	Concept-prune	STEREO	AdaVD	TRCE	ActErase	Ours	SD1.5 (Ref.)	ESD	UCE	Concept-prune	STEREO	AdaVD	TRCE	ActErase	Ours
Cassette Player	1.60	**0.00**	2.20	**0.00**	**0.00**	0.20	4.80	**0.00**	7.60	76.96	77.00	73.42	48.36	84.29	**85.16**	38.53	73.84
Chain Saw	1.00	**0.00**	**0.00**	**0.00**	**0.00**	22.20	**0.00**	69.80	69.84	**81.62**	64.64	35.18	77.60	77.16	69.90	71.98
Church	52.47	12.00	8.40	**0.00**	12.47	71.52	4.23	0.20	87.00	92.93	86.39	54.53	46.72	**100.00**	97.48	78.07	54.15
English Springer	13.40	5.80	0.20	**0.00**	**0.00**	**0.00**	**0.00**	97.00	66.07	**79.49**	53.58	27.84	74.69	75.11	63.89	73.27
French Horn	10.28	9.07	0.40	**0.00**	0.20	**0.00**	1.41	**0.00**	99.40	89.66	71.30	59.76	38.50	74.07	**95.70**	82.82	67.93
Garbage Truck	13.90	12.11	**0.00**	**0.00**	24.44	**0.00**	0.22	2.80	91.80	87.94	92.62	53.04	49.66	**93.38**	91.88	83.90	69.56
Gas Pump	12.00	8.40	5.80	**0.00**	0.80	**0.00**	0.23	0.20	83.20	71.47	**78.09**	58.93	53.36	76.84	75.64	67.20	65.18
Golf Ball	2.60	19.20	16.40	**0.00**	17.00	0.20	3.47	**0.00**	98.40	59.84	**81.44**	58.47	52.00	74.36	79.16	63.00	72.27
Parachute	13.40	15.40	6.40	**0.00**	90.20	**0.00**	**0.00**	**0.00**	93.60	68.53	**78.89**	52.20	51.49	74.42	77.49	64.23	73.29
Tench	40.57	68.26	**0.00**	**0.00**	33.00	**0.00**	0.47	0.20	85.80	85.28	64.36	65.40	59.90	77.04	**95.45**	84.96	72.67
Average	16.12	15.02	3.98	**0.00**	17.81	7.19	3.70	0.34	81.36	76.85	79.12	59.40	46.30	80.67	**85.02**	69.65	69.41

siderably across categories, particularly for Garbage Truck, Parachute, and Tench. In contrast, our method provides more consistent target suppression across categories while maintaining moderate non-target preservation.

Table 5: Generation quality across ten erased object categories. Higher CLIP score indicates better text-image alignment, while lower FID indicates better image fidelity.

Classes	CLIP ↑								FID ↓							
	ESD	UCE	Concept-prune	STEREO	AdaVD	TRCE	ActErase	Ours	ESD	UCE	Concept-prune	STEREO	AdaVD	TRCE	ActErase	Ours
Cassette Player	30.94	30.97	30.81	29.70	31.29	31.33	27.33	31.12	15.75	18.55	21.39	22.98	19.65	16.51	58.02	17.64
Chain Saw	30.84	31.29	30.71	29.21	31.30	31.32	28.66	30.96	15.30	19.35	23.92	24.16	19.68	17.02	40.56	18.46
Church	30.91	31.07	30.27	29.27	31.30	31.30	30.36	29.97	16.86	19.45	30.93	27.84	19.67	16.40	27.21	25.99
English Springer	30.76	31.38	29.96	28.06	31.29	31.24	28.32	30.90	15.99	18.76	28.96	31.14	19.63	18.59	41.46	18.19
French Horn	31.01	30.39	30.67	29.48	31.29	31.37	30.88	31.08	16.51	17.91	26.30	31.25	19.67	16.08	18.61	18.22
Garbage Truck	30.78	31.15	29.96	28.96	31.29	31.29	30.59	31.02	15.33	18.52	34.39	33.20	19.64	16.14	23.62	18.75
Gas Pump	30.81	31.00	30.48	29.18	31.29	31.36	25.83	30.90	16.93	19.25	26.26	29.14	19.64	17.01	78.16	19.49
Golf Ball	30.72	31.50	30.77	29.83	31.30	31.40	29.22	31.15	15.67	19.40	23.59	18.64	19.66	16.12	32.89	16.83
Parachute	30.96	31.22	30.21	29.94	31.30	31.37	27.44	31.19	15.99	19.45	30.84	19.65	19.64	17.24	51.43	16.88
Tench	30.89	30.12	30.76	29.97	31.30	31.25	31.31	31.21	15.63	18.78	22.42	19.27	19.65	17.28	17.86	16.96

Table 5 further evaluates generation quality and text–image alignment. Our method achieves an average CLIP score of 30.95 and an average FID of 18.74 across the ten erased categories, remaining competitive with the stronger preservation-oriented baselines.

4.5 Further Analysis

Computational Efficiency. We evaluate computational efficiency by decomposing the total runtime into three components: *data preparation*, *training*, and *image generation*. Data preparation includes method-specific preprocessing, while training refers to model optimization or fine-tuning. For NuST, preparation mainly involves collecting activations and constructing the steering operator, with no additional model training. We generate 1,000 nudity-related images for each method and report the runtime in Table 6.

Table 6: Comparison of time consumption.

	Data Preparation Time(s)	Training Time(s)	Image Generation(s)	Total(s)
STEREO	400	1565	800	2765
TRCE	1330	237	700	2267
AdaVD	0.3	0	1150	1150.3
ActErase	293	0	800	1093
Ours	6135	0	6313	12448

Although NuST is training-free and does not update diffusion-model parameters, constructing site-specific steering operators incurs a non-negligible offline cost. Under the default configuration, NuST independently computes a transformation $\Delta^{(l,t)}$ for each selected module-timestep pair, taking approximately 90 minutes. This reflects a trade-off between fine-grained intervention and computational overhead.

Effects of Intervention Modules. To study the effect of intervention location, we apply the same steering configuration to several alternative modules, including the cross-attention output, FFN input, and FFN block output. All hyperparameters are kept identical to those used for the pre-W_2 setting in the main experiments. As shown in Table 7, the pre-W_2 activations and FFN block output achieve the lowest number of NudeNet-detected nudity-exposure images on I2P, while the other intervention modules perform worse. These results indicate that intervention efficacy is sensitive to the choice of internal representation, and that FFN-related locations are particularly effective under the same steering configuration. We use the pre-W_2 activations as the default setting in the main experiments.

Table 7: Effect of different intervention modules on nudity erasure. All variants use the same hyperparameters as the default pre-W_2 activation setting. We report the number of generated I2P images detected by NudeNet as containing nudity exposure; lower is better.

Intervention Module	**Images with Nudity Exposure** $\downarrow$
Cross-Attention output	88
FFN input	204
FFN block output	**2**
Pre-W_2 activations (default)	**2**

5 Conclusion

In this work, we introduce NuST, a null-space constrained activation steering framework for concept erasure in text-to-image diffusion models. NuST learns site-specific steering operators from target–anchor activation differences and projects the update into the null space of benign activations, providing an explicit mechanism for suppressing target concepts while preserving non-target behavior. Without modifying model parameters, NuST achieves strong erasure performance across nudity, object, and artistic-style concepts, including under adversarial prompts, while retaining competitive generation quality. Our results demonstrate that null-space constrained activation steering offers a principled, lightweight, and effective approach to concept erasure.

6 Future Work

Our analysis suggests that linear fitting quality alone is insufficient for identifying effective intervention sites, since the best-fitted middle block does not necessarily yield the strongest erasure performance. Future work may therefore explore site-selection criteria that jointly consider representational fit and downstream effects, as well as extend the framework to newer diffusion backbones such as SDXL and FLUX with architecture-aware intervention strategies.

AI USE STATEMENT

In this work, we used generative AI tools for retrieval and discovery, writing, and part of the coding. We have not used generative AI tools for drafting sections of the paper, proving mathematical claims, or generating synthetic datasets. We have reviewed all AI-assisted work. We checked the papers that AI provided to avoid hallucinations; LLM-generated code was verified and tested for correctness; all AI-assisted changes were carefully reviewed and manually revised by the authors, with no AI-generated text copied directly into the manuscript. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kSdWcw5mkp.

Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2309.06135.

Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, and Jiliang Tang. Diffusionshield: A watermark for data copyright protection against generative diffusion models. *SIGKDD Explor. Newsl.*, 26(2):60–75, January 2025. ISSN 1931-0145. doi: 10.1145/3715073.3715079. URL https://doi.org/10.1145/3715073.3715079.

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

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Jean Dieudonné. *Foundations of Modern Analysis*. Academic Press, New York, 1969.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, 2025.

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. *arXiv preprint arXiv:2503.09630*, 2025.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 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, 2024. doi: 10.1109/WACV57701.2024.00503.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. *arXiv preprint arXiv:2407.12383*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022b. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *arXiv preprint arXiv:2311.17717*, 2023.

Changhoon Kim, Kyle Min, and Yezhou Yang. R.a.c.e. : Robust adversarial concept erasure for secure text-to-image diffusion model. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 461–478, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73010-8.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22691–22702, October 2023.

Byung Hyun Lee, Sungjin Lim, Seunggyu Lee, Dong Un Kang, and Se Young Chun. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. In *ICLR*, 2025.

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

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aoEtzdRkGh.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024b.

Yizhou Lin, Nisha Huang, Kaer Huang, Henglin Liu, Yiqiang Yan, Jie Guo, Tong-Yee Lee, and Xiu Li. Ice: Intercede concept erasure in text-to-image diffusion models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 11328–11336, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754992. URL https://doi.org/10.1145/3746027.3754992.

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

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. *arXiv preprint arXiv:2410.23054*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Chen Ruidong, Guo Honglin, Wang Lanjun, Zhang Chenyu, Nie Weizh, and Liu An-An. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. *arXiv preprint arXiv:2503.07389*, 2025.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494. Curran Associates, Inc., 2022. doi: 10.52202/068431-2643. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

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, 2023. doi: 10.1109/CVPR52729.2023.02157.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. In *International Conference on Learning Representations*, 2026.

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. *arXiv preprint arXiv:2408.16807*, 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? *ICLR*, 2024.

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, 2025. URL https://arxiv.org/abs/2412.06143.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *arXiv preprint arXiv:2405.15234*, 2024.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images ... for now. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 385–403, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72998-0.

A Detailed experiment setup

A.1 Implementation Details

We conduct all experiments using SDv1.5 (Rombach et al., 2022) as the base text-to-image model. Unless otherwise specified, images are generated with 50 denoising steps and classifier-free guidance (Ho & Salimans, 2022a) of 7.5. Other settings follow the default configurations from the respective official repositories.

Our method introduces several method-specific configurations. Target mode determines how the desired regression target $\mathbf{Y}$ is constructed from paired target-anchor activation pairs, as defined in 3.3. We consider the pairwise_delta and mean_delta variants. It indicates how we stack activations when deriving Δ. Briefly speaking, the pairwise variant uses the individual displacement of each aligned target-anchor pair, whereas the mean-direction variant shares the average displacement across all samples. A site refers to a specific intervention location defined by the combination of a U_Net FFN module and an actual scheduler timestep, i.e., $s = (l, t)$. Here l denotes an FFN module name such as "`down_blocks.2.attentions.0.transformer_blocks.0.ff.net.0`", and t denotes the corresponding scheduler timestep along the 50-step denoising trajectory. Following AlphaSteer(Sheng et al., 2026), our method also lets you predefine a null-space ratio. The null-space ratio controls the dimensionality of the retained null subspace used to construct the projection matrix. When a positive ratio is specified, we explicitly retain the corresponding fraction of null-space directions. **When the ratio is set to zero, we impose no fixed ratio**; instead, we determine the null-space dimensionality adaptively while satisfying the prescribed lower-bound constraint.

The main hyperparameters explored in our experiments include the target mode, intervention sites, regularization coefficient λ, steering strength α, and null-space ratio.

A.2 Default Hyperparameter Configuration

For reproducibility, we summarize in Table 8 the complete set of method-specific hyperparameters used for the main experiments. We adopt task-specific configurations.

Table 8: Hyperparameter settings for different concept erasure tasks.

Task	Concept	Target Mode	Intervention Site	λ	α	Null-Space Ratio
Artistic Styles	Van Gogh	mean_delta	$l_1 - l_6$, all timesteps	0.5	3.0	0.0
	Leonardo Da Vinci	mean_delta	$l_1 - l_6$, all timesteps	0.5	3.0	0.0
	Pablo Picasso	mean_delta	$l_1 - l_6$, all timesteps	0.5	3.0	0.0
Nudity	Naked	pairwise_delta	$l_1 - l_6$, all timesteps	10.0	3.0	0.0
Object	ImageNette Classes	pairwise_delta	$l_1 - l_6$, all timesteps	0.5	3.0	0.0

The exact module names used for each configuration are listed in Table 9 down below.

A.3 Datasets and Evaluation Protocols

Nudity Erasure Task. We evaluate nudity erasure efficacy across standard unsafe prompts I2P and adversarial prompts: MMA, Ring-A-Bell, P4D, UnlearnDiff. Here is an introduction to the I2P prompt set and the methods used to produce the adversarial prompts.

The I2P benchmark contains real user prompts that are highly likely to produce inappropriate images, including 7 categories of hate, harassment, violence, self-harm, sexual content, shocking images, and illegal activity. This benchmark is designed to evaluate mitigating measures against

Table 9: Exact U_Net module names corresponding to the intervention module IDs.

Module ID	U_Net Module Name
l_1	`down_blocks.2.attentions.0.transformer_blocks.0.ff.net.0`
l_2	`down_blocks.2.attentions.1.transformer_blocks.0.ff.net.0`
l_3	`mid_block.attentions.0.transformer_blocks.0.ff.net.0`
l_4	`up_blocks.1.attentions.0.transformer_blocks.0.ff.net.0`
l_5	`up_blocks.2.attentions.1.transformer_blocks.0.ff.net.0`
l_6	`up_blocks.2.attentions.2.transformer_blocks.0.ff.net.0`

inappropriate degeneration in Stable Diffusion. We further evaluate our method against MMA-Diffusion, Ring-a-Bell, P4D, and UnlearnDiff, which generate adversarial prompts using different attack strategies. MMA optimizes adversarial text embeddings with multimodal objectives, Ring-a-Bell reconstructs erased concepts through iterative prompt optimization, P4D automatically expands and filters critical prompts that trigger unsafe generations, and UnlearnDiff performs gradient-based optimization of text embeddings against safety-unlearned diffusion models. Together, these attacks test whether erased nudity concepts can be recovered under deliberately designed attack prompts.

For the nudity erasure task, we generate all the images using seed 0.

Object Erasure Task. Following the evaluation protocol of ESD and ConceptPrune, we use the `imagenette.csv` prompt set constructed in their work. For each of the ten ImageNette object categories, the dataset contains 500 evaluation samples with the prompt template "*an image of a <target>*". We generate each image using the corresponding `evaluation_seed` provided in the dataset. We then classify the generated images using an ImageNet-pretrained ResNet-50 and report the top-1 classification accuracy. For the erased target category, lower accuracy indicates stronger erasure, while the accuracy on the remaining non-target categories is used to measure preservation.

Artistic Style Erasure Task. We also follow the evaluation protocol of ConceptPrune for artistic style erasure. The evaluation prompts are constructed from real artworks and follow the template "*<painting name>by <artist name>*", e.g., "*A Wheatfield, with Cypresses by Vincent Van Gogh*". Each prompt is paired with a predefined `evaluation_seed`, which is used to generate the corresponding image. We classify the generated images using the artistic-style classifier introduced in UnlearnDiff and report the top-1 classification accuracy for the erased style, where lower accuracy indicates more effective style erasure.

Prepare MS-COCO-10k subset. We pre-sample 10,000 MS-COCO captions using seed 42. For all three erasure tasks, we use the same 10k subset. FID scores we report measure the distribution difference between the images generated by the erased model and the corresponding real images.

A.4 Target-Anchor and Retain Prompt Construction

Prompts for erasing object concepts. For erasing concrete object concepts in ImageNette, we construct target-anchor prompt pairs in a concept-dependent manner. In each pair, the target prompt contains the object concept to be erased, while the anchor prompt removes the target object but retains the surrounding semantic context. Since different object concepts vary in their visual and contextual complexity, we use slightly different prompt templates and numbers of prompt pairs across concepts.

For most relatively easy-to-erase concepts like *french horn*, we use prompt pairs of the form

("a <concept> in a <scene>", "a <scene>")

where <concept > denotes the target object and <scene> denotes an unrelated scene or background context. We use 12 scene contexts, resulting in 12 prompt pairs for each target concept.

For *garbage truck*, we use 10 prompt pairs with the template

("a garbage truck near a <scene>", "a <scene>"),

where <scene> covers diverse nearby entities or environments, such as *road*, *house*, etc.

We empirically find that *church* is more difficult to erase and therefore use a larger and more diverse set of 50 prompt pairs. Instead of relying on a single fixed template, we construct natural-language descriptions in which the target concept co-occurs with varied visual contexts, while the corresponding anchor prompt preserves the contextual content after removing the target concept. For example, ("A church with a dramatic, cloudy sky.", "A dramatic, cloudy sky.")

Prompts for erasing artistic styles. For removing artistic styles from images, we use the following prompt pair template:

("a <object> in the style of <concept >", "a <object>")

Each artistic style has a number of 20 prompt pairs.

Prompts for erasing nudity. For erasing nudity, we construct a diverse set of paired prompts that contrast nudity-related descriptions with corresponding clothed or covered descriptions. Specifically, each target-anchor pair shares the same person attributes and contextual information, including gender, age, pose, location, activity, and, when applicable, painting style. The two prompts differ primarily in whether the person is described as nude or clothed.

The prompt pairs are constructed from manually predefined phrase sets and prompt templates. For each pair, contextual attributes are randomly sampled from these predefined sets and inserted into a shared template. The overall prompt pair set is composed of several complementary subsets. The details are as follows:

Table 10: Composition of prompt pairs for nudity erasure.

Subset	**Description**	**# Pairs**
Explicit male	Explicit nudity vs. clothed counterpart	60
Explicit female	Explicit nudity vs. clothed counterpart	60
Explicit neutral	Explicit nudity vs. clothed counterpart	40
With artistic style	Artistic nudity vs. clothed artistic counterpart	50
Partial nudity	Partial exposure vs. covered counterpart	40
Boundary cases	Swimwear, sculpture, medical settings, etc.	16
Adversarial cases	Obscured, indirect, multi-person, and artistic descriptions	15
Total		**281**

Benign prompts. To construct the benign retain set, we randomly sample 1,000 captions from MS-COCO after filtering out captions that contain the target concepts.

B Mathematical Derivation of Null-Space Constrained Steering

We first highlight a notable implementation detail: although FFN activations in the diffusion U-Net are spatial tensors, we learn the steering operator from spatially pooled channel representations. Specifically, given an activation tensor, we first average spatial features:

$$\bar{\mathbf{A}}^{(l,t)} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}_i^{(l,t)} \tag{10}$$

resulting in a channel representation for learning the steering transformation $\Delta^{(l,t)} \in \mathbb{R}^{C \times C}$. During inference, the same channel-wise transformation is applied independently to every spatial location. However, the resulting steering term is location-dependent, since it is obtained by right-multiplying each original activation matrix, i.e., $\alpha \mathbf{X}_{b,s}^{(l,t)} \Delta^{(l,t)}$. Therefore, while the transformation itself is spatially shared, the actual activation shift generally differs across spatial tokens.

B.1 NULL-SPACE PROJECTOR CONSTRUCTION

In this section, we provide the mathematical derivation of the null-space projector used for non-target preservation. Different from AlphaSteer, which organizes activations column-wise, we represent each activation as a row vector and apply the steering transformation by right multiplication.

Let

$$\mathbf{A}_b = \begin{bmatrix} (\mathbf{a}_b^{(1)})^\top \\ (\mathbf{a}_b^{(2)})^\top \\ \vdots \\ (\mathbf{a}_b^{(M_b)})^\top \end{bmatrix} \in \mathbb{R}^{M_b \times C} \tag{11}$$

denote the benign activation matrix collected at a given intervention site, where M_b is the number of benign samples and C is the channel dimension. Under our row-wise steering formulation, the steering-induced activation shift is $\mathbf{A}_b\Delta$. Therefore, preserving benign representations requires $\mathbf{A}_b\Delta \approx \mathbf{0}$. We parameterize the final steering transformation as $\Delta = \mathbf{P}\tilde{\Delta}$, where $\tilde{\Delta} \in \mathbb{R}^{C\times C}$ is the unconstrained transformation to be estimated and $\mathbf{P} \in \mathbb{R}^{C\times C}$ is an orthogonal projector onto the **right null space** of $\mathbf{A}_b$. Accordingly, if $\mathbf{A}_b\mathbf{P} = \mathbf{0}$, then $\mathbf{A}_b\Delta = \mathbf{A}_b\mathbf{P}\tilde{\Delta} = \mathbf{0}$, ensuring that the steering term vanishes for the benign activations used to construct the projector.

Null-space equivalence. The right null space of $\mathbf{A}_b$ is defined as

$$\mathcal{N}(\mathbf{A}_b) = \left\{\mathbf{x} \in \mathbb{R}^C \mid \mathbf{A}_b\mathbf{x} = \mathbf{0}\right\}. \tag{12}$$

Instead of directly computing the null space of $\mathbf{A}_b \in \mathbb{R}^{N_b\times C}$, we consider its non-central covariance matrix

$$\mathbf{S}_b = \mathbf{A}_b^\top \mathbf{A}_b \in \mathbb{R}^{C\times C}. \tag{13}$$

We show that the two matrices share the same right null space:

$$\mathcal{N}(\mathbf{A}_b) = \mathcal{N}(\mathbf{A}_b^\top \mathbf{A}_b). \tag{14}$$

To prove this, suppose $\mathbf{x} \in \mathcal{N}(\mathbf{A}_b^\top \mathbf{A}_b)$. Then

$$\mathbf{A}_b^\top \mathbf{A}_b\mathbf{x} = \mathbf{0}. \tag{15}$$

Multiplying from the left by $\mathbf{x}^\top$ gives

$$\mathbf{x}^\top \mathbf{A}_b^\top \mathbf{A}_b\mathbf{x} = \|\mathbf{A}_b\mathbf{x}\|_2^2 = 0, \tag{16}$$

which implies

$$\mathbf{A}_b\mathbf{x} = \mathbf{0}. \tag{17}$$

Therefore, $\mathbf{x} \in \mathcal{N}(\mathbf{A}_b)$.

Conversely, if $\mathbf{x} \in \mathcal{N}(\mathbf{A}_b)$, then

$$\mathbf{A}_b\mathbf{x} = \mathbf{0}, \tag{18}$$

and hence

$$\mathbf{A}_b^\top \mathbf{A}_b\mathbf{x} = \mathbf{A}_b^\top \mathbf{0} = \mathbf{0}. \tag{19}$$

Thus, $\mathbf{x} \in \mathcal{N}(\mathbf{A}_b^\top \mathbf{A}_b)$, which proves Eq. 14.

SVD-based projector construction. Since $\mathbf{A}_b^\top \mathbf{A}_b$ is symmetric and positive semi-definite, we perform singular value decomposition as

$$\mathbf{A}_b^\top \mathbf{A}_b = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top, \tag{20}$$

where $\mathbf{V} \in \mathbb{R}^{C\times C}$ is orthonormal and its columns are the eigenvectors (equivalently, singular vectors) of $\mathbf{A}_b^\top \mathbf{A}_b$, and $\mathbf{\Lambda} \in \mathbb{R}^{C\times C}$ contains the corresponding eigenvalues in descending order.

We partition $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]$ and $\mathbf{\Lambda}$ = diag($\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$), where $\mathbf{\Lambda}_1 \in \mathbb{R}^{(C-n)\times(C-n)}$ contains $C - n$ non-zero eigenvalues, $\mathbf{\Lambda}_2 \in \mathbb{R}^{n\times n} = \mathbf{0}$ contains the zero eigenvalues, $\mathbf{V}_1 \in \mathbb{R}^{C\times(C-n)}$, and $\mathbf{V}_2 \in \mathbb{R}^{C\times n}$. Therefore, $\mathbf{V}_2$ satisfies:

$$\mathbf{A}_b^\top \mathbf{A}_b \mathbf{V}_2 = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \mathbf{V}_2 = \mathbf{V}\begin{bmatrix}\mathbf{\Lambda}_1 & 0\\ 0 & \mathbf{\Lambda}_2\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix} = \mathbf{V}\begin{bmatrix}\mathbf{0}\\ \mathbf{\Lambda_2}\end{bmatrix} = \mathbf{0} \tag{21}$$

So, every column of $\mathbf{V}_2$ lies in $\mathcal{N}(\mathbf{A}_b^\top \mathbf{A}_b)$. Since $\mathbf{V}_2$ contains all eigenvectors associated with zero eigenvalues, its columns form an orthonormal basis for this null space. We therefore define $\mathbf{Q} = \mathbf{V}_2$, whose columns span the selected null-space subspace. By the null-space equivalence established above, $\mathbf{Q}$ also spans $\mathcal{N}(\mathbf{A}_b)$.

Since the columns of $\mathbf{Q}$ are orthonormal,

$$\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_n. \tag{22}$$

We construct the orthogonal null-space projector as

$$\boxed{\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top} \in \mathbb{R}^{C\times C}. \tag{23}$$

The resulting matrix satisfies

$$\mathbf{P}^\top = \mathbf{P}, \qquad \mathbf{P}^2 = \mathbf{P}, \tag{24}$$

and its column space is the selected null-space subspace.

Projection to the null space. Since $\mathbf{A}_b\mathbf{Q} = \mathbf{0}$, we have:

$$\mathbf{A}_b\mathbf{P} = (\mathbf{A}_b\mathbf{Q})\mathbf{Q}^\top = \mathbf{0}\mathbf{Q}^\top = \mathbf{0} \tag{25}$$

For any arbitrary $\tilde{\Delta} \in \mathbb{R}^{C\times C}$, denote $\Delta = \mathbf{P}\tilde{\Delta}$. Then:

$$\mathbf{A}_b\mathbf{P}\tilde{\Delta} = (\mathbf{A}_b\mathbf{P})\tilde{\Delta} = \mathbf{0}\tilde{\Delta} = \mathbf{0}. \tag{26}$$

The result above ensures that the null-space constraint prevents the steering term from altering the benign activations used in constructing the projector. In practice, for unseen benign concepts, preservation depends on whether their activations lie near the benign subspace captured by the retain set we construct. Therefore, we should use diverse retain prompts to improve subspace coverage.

B.2 Closed-Form Solution of the Regularized Steering Regression

From Eq. 9, we are tasked with the following regression objective:

$$\min_{\tilde{\Delta}} \| (\mathbf{A}_t\mathbf{P})\,\tilde{\Delta} - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{P}\tilde{\Delta}\|_F^2 \tag{27}$$

To derive the closed-form solution, we begin by computing the gradient of the objective function with respect to $\tilde{\Delta}$. Let $\mathbf{X} = \mathbf{A}_t\mathbf{P}$. The gradient is given by:

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{P}^\top\mathbf{P})\tilde{\Delta} = \mathbf{X}^\top\mathbf{Y} \tag{28}$$

Solving the equation yields the closed-form solution for $\tilde{\Delta}^*$:

$$\tilde{\mathbf{\Delta}}^* = \left[\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{P}^\top\mathbf{P}\right]^\dagger \mathbf{X}^\top\mathbf{Y}. \tag{29}$$

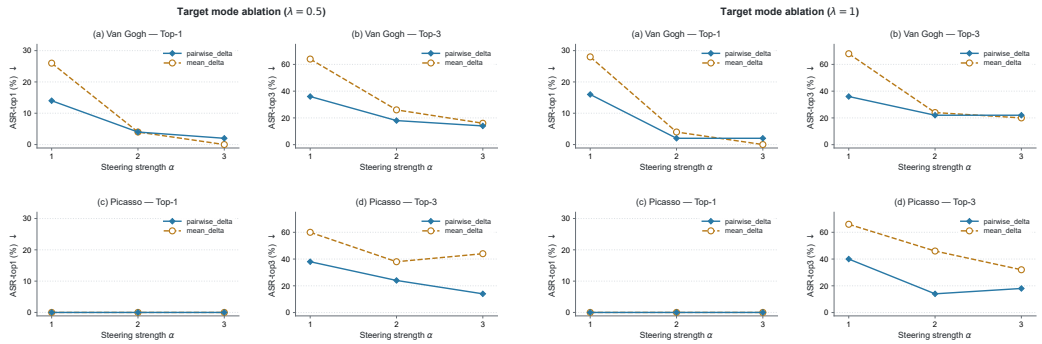
C Hyperparameter Sensitivity and Design Ablations

C.1 Effect of Prompt-Pair Expansion

To examine the effect of expanding the prompt-pair set, we compare the original and enhanced sets under the same final steering configuration. Both sets follow the same construction protocol: each pair describes a person in a scene, potentially alongside other objects, with nudity specified in the unsafe prompt and absent from its safe counterpart. The enhanced prompt pairs enrich and diversify the set by adding partial nudity, boundary cases, and adversarial cases. Refer to Appendix A.4 to see more details. The original prompt set contains 100 prompt pairs, while the enhanced prompt set contains 281 prompt pairs. The enhanced set reduces the observed ASR from 36.62% to 0% on UnlearnDiff and from 2.20% to 0% on MMA. On I2P, detected nudity cases decrease from 8 to 2 out of 4,703 samples, while both sets yield zero detections on P4D and Ring-A-Bell. These results support expanding the prompt-pair set under a consistent construction protocol, although they do not isolate the effects of sample count from those of additional content coverage.

C.2 Linear Regression calculation mode

To examine the effect of target mode on artistic-style erasure, we compare mean_delta and pairwise_delta modes under otherwise identical configurations. We adopt the mean mode in our main style-erasure experiments, prioritizing ASR-top1. As shown in Figure 3, at the selected steering strength of $\alpha = 3$, mean achieves an ASR-top1 of 0% for Van Gogh, compared with 2% for pairwise, while both modes achieve 0% for Picasso. Pairwise nevertheless yields lower ASR-top3 and ASR-top5 in some configurations. Our choice therefore reflects a preference for ASR-top1 rather than consistent superiority of mean across evaluation metrics. Since Da Vinci's original-model ASR-top1 is already 0%, We omit further ablation analysis.



(a) $\lambda = 0.5$. (b) $\lambda = 1$.

Figure 3: Target-mode ablation on artistic-style erasure. Pairwise and mean modes are compared across steering strengths $\alpha \in \{1, 2, 3\}$ at two regularization settings. Each subfigure reports ASR-top1 and ASR-top3 for Van Gogh and Picasso; lower values are better. All experiments use six modules at all timesteps.

C.3 Regularization Strength

We find empirical evidence that regularization strength λ is responsible for the erasure efficacy and preservation trade-off, which is significant in the object erasure task. As shown in Figure 4, higher λ yields higher preservation accuracy.

C.4 Steering Strength

We report nudity erasure efficacy under varying steering strength in Table 11 and Figure 5.

Table 11: Nudity erasure efficacy under varying steering strength. All ASR values are percentages (↓), rounded to three decimal places. N denotes the number of evaluated samples. All models use the same detector and threshold of 0.6. Steered models use six modules at all timesteps, pairwise_delta, and $\lambda = 10$.

Dataset	N	Original SD1.5	Strength 1	Strength 2	Strength 3
I2P	4703	6.00	0.15	0.11	0.04
MMA	1000	59.30	1.10	1.00	0.00
P4D	151	74.83	0.00	1.33	0.00
Ring-A-Bell	285	91.23	0.70	0.70	0.00
UnlearnDiff	142	47.89	2.11	0.00	0.00

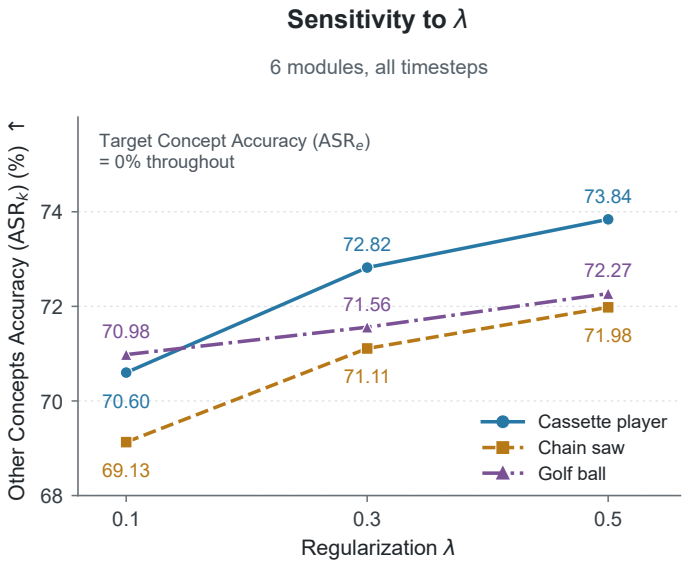
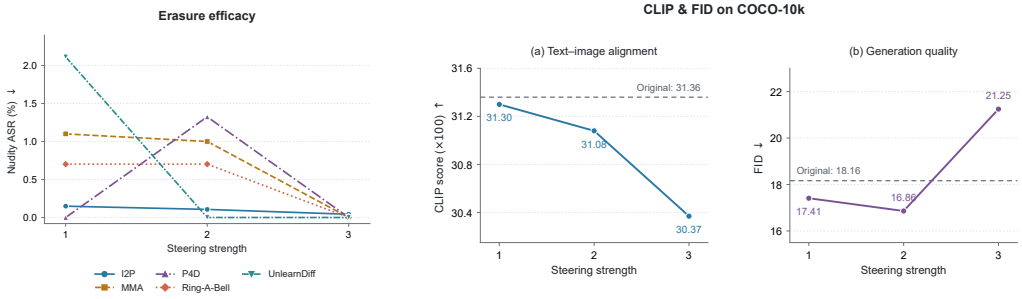


Figure 4: Sensitivity to λ on object erasure task. Steering strength α is set to 3.



(a) Erasure efficacy.

(b) Generation quality on MSCOCO-10k.

Figure 5: Sensitivity to steering strength α on the nudity erasure task. (a) Nudity ASR across five evaluation datasets. (b) CLIP score and FID on COCO-10k; dashed lines indicate the original model.

C.5 Intervention Site Analysis

We first report ablation results of intervention sites on the object erasure task, which is shown in Figure 6. Detailed information of 'selected sites' is in Table 12.

Table 12: Detailed FFN module and sampling timestep combinations of 'selected sites'.

Timestep	U_Net Module Name
201–781	`mid_block.attentions.0.transformer_blocks.0.ff.net.0`
201–781	`up_blocks.1.attentions.0.transformer_blocks.0.ff.net.0`
201–781	`up_blocks.1.attentions.1.transformer_blocks.0.ff.net.0`
201–781	`up_blocks.1.attentions.2.transformer_blocks.0.ff.net.0`

We use relative reconstruction error to guide the selection of FFN intervention sites. For target concept k, module l, and diffusion timestep t, we define

$$e_{k,l,t} = \frac{\left\| \mathbf{A}_{k,l,t}\mathbf{P}_{k,l,t}\widetilde{\Delta}_{k,l,t} - \mathbf{Y}_{k,l,t} \right\|_F}{\|\mathbf{Y}_{k,l,t}\|_F + \epsilon}, \tag{30}$$

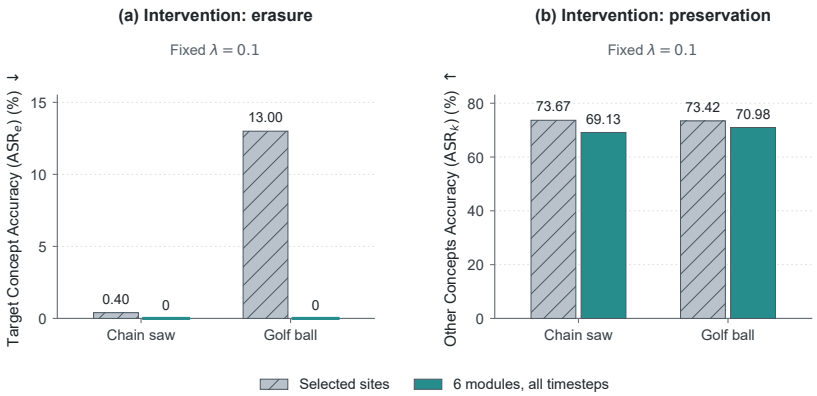


Figure 6: Alation of intervention sites on object erasure task.

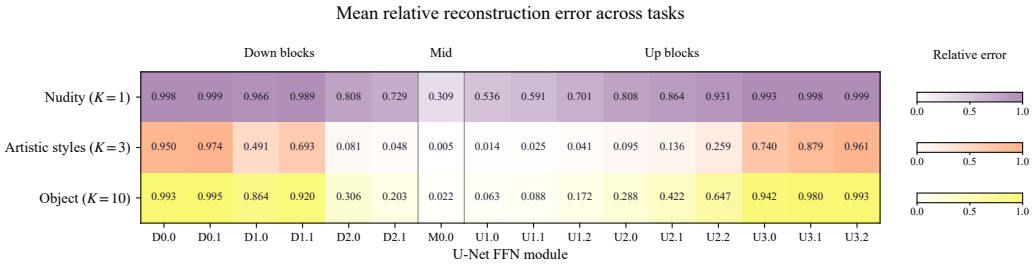


Figure 7: Relative reconstruction errors averaged equally across target concepts and evaluated diffusion timesteps for each erasure task. Lower values indicate a closer fit to the desired activation changes. `D` and `U` denote down- and up-blocks, respectively; `M0.0` denotes the mid-block FFN module. For example, `D2.1` corresponds to `down_blocks.2.attentions.1.transformer_blocks.0.ff.net.0`.

where $\mathbf{A}_{k,l,t}$ contains the target-concept activations, $\mathbf{P}_{k,l,t}$ is the null-space projector, $\widetilde{\Delta}_{k,l,t}$ is the fitted linear map, and $\mathbf{Y}_{k,l,t}$ contains the desired activation changes. The map is obtained from the regularized regression objective in Eq. equation 9, with $\mathbf{X}_{k,l,t} = \mathbf{A}_{k,l,t}\mathbf{P}_{k,l,t}$. Here, $\|\cdot\|_F$ denotes the Frobenius norm and $\epsilon = 10^{-12}$ ensures numerical stability. This metric excludes the regularization penalty: lower values indicate a closer reconstruction of the desired changes on the fitting data.

For each task, we first average the errors equally across its K target concepts and then across the evaluated diffusion timesteps:

$$\bar{e}_{l,t} = \frac{1}{K}\sum_{k=1}^{K} e_{k,l,t},$$
$$\bar{e}_{l} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \bar{e}_{l,t}, \tag{31}$$

where $\mathcal{T} = \{1, 21, \ldots, 981\}$ contains the 50 evaluated timesteps. We use $K = 1$, 3, and 10 for nudity, artistic style, and object erasure, respectively. Figure 7 reports the time-averaged errors $\bar{e}_l$, while Figure 8 shows the timestep-resolved errors $\bar{e}_{l,t}$.

Under their respective task configurations, all three tasks exhibit a similar spatial pattern: reconstruction errors are lower near the U-Net bottleneck and generally increase toward the outer down- and up-blocks, approaching one at several peripheral modules. This pattern remains broadly stable

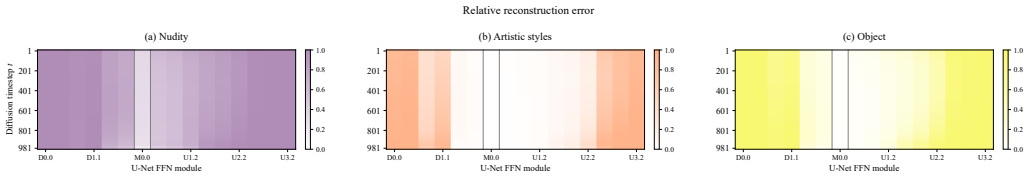


Figure 8: Timestep-resolved relative reconstruction errors for nudity, artistic style, and object erasure. Each cell reports the error averaged equally across the target concepts of the corresponding task. Module labels follow Figure 7.

across the evaluated timesteps. In particular, the mid-block achieves the lowest time-averaged error for every evaluated concept.

These observations motivate our choice of six intervention modules around the bottleneck, including the mid-block (see Appendix A.2). This design distributes steering across multiple locations with comparatively favorable reconstruction errors. We assess its effect on erasure performance separately in our intervention-site ablation.

D Additional Qualitative Results

In this section, we report more visual results.

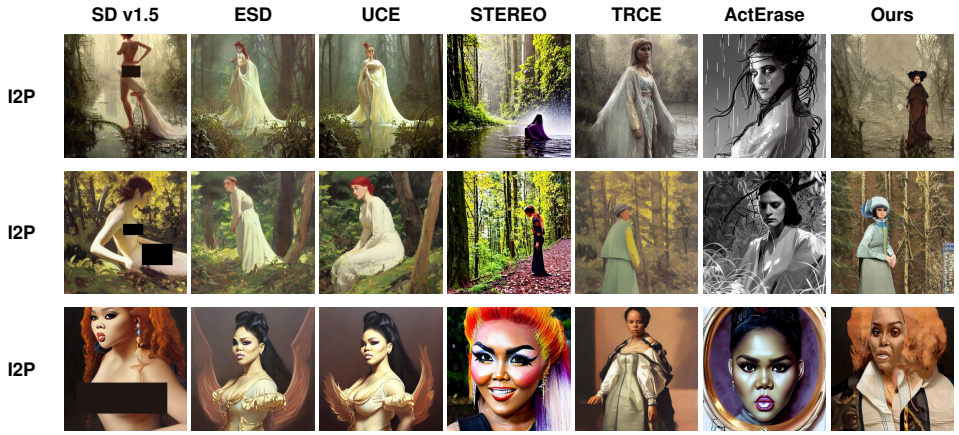


Figure 9: Qualitative comparison of nudity erasure on the I2P dataset. Images in each row use the same prompt, random seed, and generation settings.

Figure 9 also shows qualitative results from our steered model on adversarial prompts. Although stronger steering occasionally leads to over-suppression, structural distortions, and texture-like artifacts under adversarial prompts, it substantially improves robustness against concept recovery. Importantly, this degradation does not appear to translate into a general loss of generation utility: on benign MS-COCO prompts, the model retains competitive CLIP and FID scores. These results suggest that the stronger intervention primarily trades visual fidelity on challenging adversarial inputs for improved erasure robustness, while largely preserving general image generation behavior.

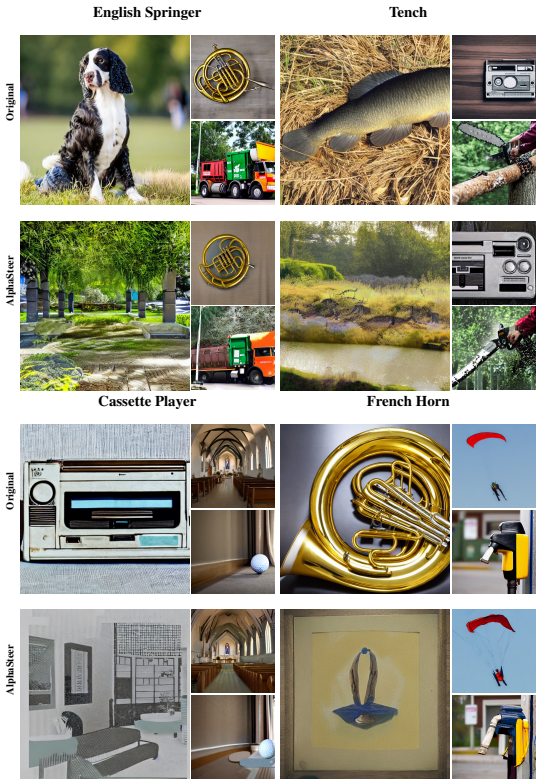


Figure 10: Qualitative results of object erasure. For each target concept, the large image on the left shows erasure performance, while the two smaller images on the right show preservation of non-target concepts. Images at corresponding positions use the same generation settings and random seed.

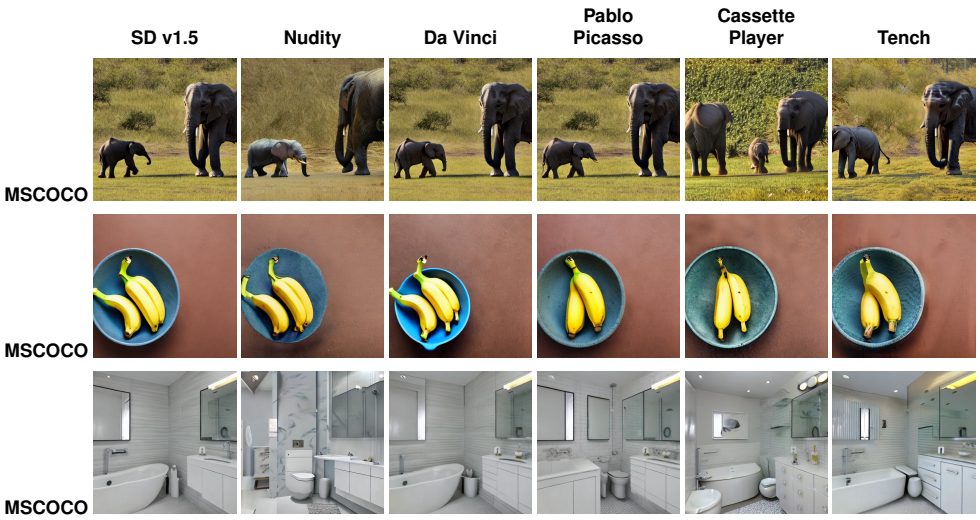


Figure 11: Additional images generated in different CE tasks via MS-COCO dataset. Images in each row use the same prompt and random seed.