CRET: TARGET-SELECTIVE SUBSPACE LEARNING FOR CONCEPT ERASURE IN TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models have demonstrated remarkable capabilities in visual generation, but their widespread use also raises concerns about the generation of unsafe content. Concept erasure aims to suppress designated undesirable concepts while preserving non-target semantics. Projection-based methods provide interpretable intervention by removing target-related representation components. Their projection subspaces are commonly constructed directly from selected embeddings. However, for complex concepts, target-related and benign representations can share certain components, making it difficult for such subspaces to separate the information to be edited from that to be preserved, resulting in incomplete target-concept erasure or suppression of benign information. To overcome this limitation, we propose CRET, a representation-level concept erasure framework that learns an editable subspace rather than directly constructing it from selected embeddings. This subspace is parameterized by a rank-constrained orthogonal projector. In particular, CRET optimizes the projector to selectively capture target-related information from controlled risky–safe prompt pairs while limiting the projections of safe and preservation representations onto the editable subspace. The projector is learned in a latent feature space of text representations, and the induced change is transferred back to the original representation without updating the pretrained diffusion model weights. Experiments on multiple adversarial benchmarks demonstrate that CRET achieves more effective target-concept erasure than competing methods while maintaining competitive generation quality on benign prompts.

1 INTRODUCTION

Text-to-image diffusion models (Podell et al., 2024; Dhariwal & Nichol, 2021; Ho & Salimans, 2022) have demonstrated remarkable capabilities in visual generation and have been widely adopted across various domains. However, because these models are typically trained on large-scale web-crawled datasets, they may also generate pornographic, violent, or otherwise inappropriate content, raising substantial public concerns about potential misuse. Although such undesirable knowledge can be mitigated at its source by filtering the training data and retraining the model, the associated data and computational costs are often prohibitively high. Concept erasure has therefore emerged as a more practical solution, aiming to selectively suppress designated undesirable concepts while preserving the overall generative capabilities of the pretrained model.

Existing concept erasure methods can be broadly categorized as follows. Parameter fine-tuning methods (Gandikota et al., 2023; Kumari et al., 2023) achieve persistent erasure by updating model parameters through gradient-based optimization, but require substantial computational cost for concept-specific adaptation. To reduce the cost of iterative training, closed-form editing methods (Gandikota et al., 2024; Gong et al., 2024) employ analytical updates to directly modify model parameters. Despite their different optimization strategies, both paradigms modify pretrained model parameters, which may affect non-target semantics that are entangled with the target concept. Inference-time methods, by contrast, keep the base model parameters fixed and suppress target semantics by guiding the denoising process (Schramowski et al., 2023) or manipulating prompt embeddings (Cai et al., 2026; Xiong et al., 2025). In particular, embedding manipulation directly

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.

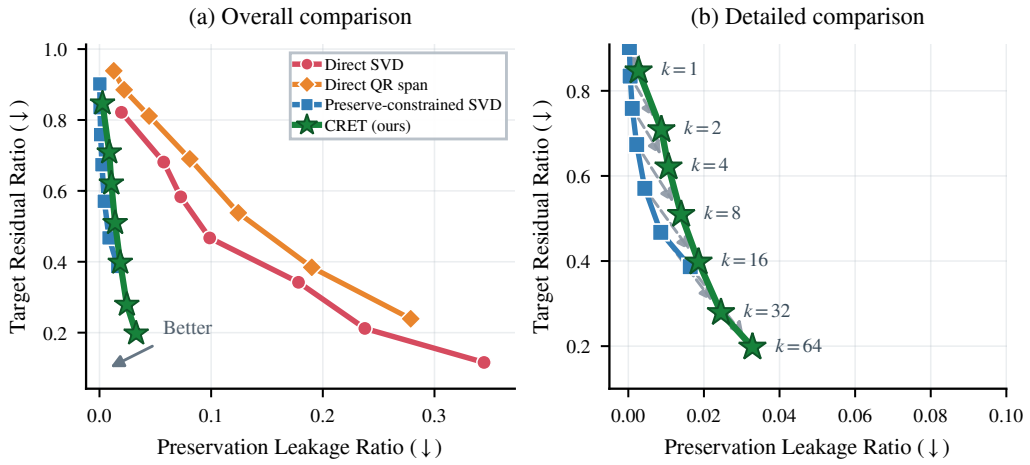

Figure 1: Comparison of projector constructions in the text-conditioning space: Direct SVD (Chen et al., 2026), Direct QR Span (Yoon et al., 2025; Wang et al., 2025), and Preserve-constrained SVD (Biswas et al., 2025; Vardhana & Biswas, 2026). (a) Overall comparison of preservation leakage and target residual ratios across projector ranks. (b) Detailed comparison between preserve-constrained SVD and CRET, where each gray arrow connects results obtained with the same rank k. Lower values on both axes are better.

modifies the textual conditioning before diffusion sampling, making the intervention lightweight and convenient for deployment.

Recent projection-based methods provide a way to implement such embedding manipulation, and are also used to edit cross-attention representations during denoising. In text-conditioning space, some methods project prompt representations onto the complements of target subspaces (Yoon et al., 2025; Chen et al., 2026), while others combine complementary removal with alignment toward clean semantic spaces (Cao et al., 2025). In downstream cross-attention value space, target-aligned directions can instead be removed during denoising (Wang et al., 2025). More recent methods further incorporate preservation information (Vardhana & Biswas, 2026; Shi et al., 2026). Despite these refinements, their projection subspaces are typically constructed by directly spanning, decomposing, or orthogonalizing embeddings associated with target or preservation concepts. Consequently, the boundary between the editable and protected spaces depends on the existing geometry of these embeddings in the chosen feature space. This geometry-dependent partitioning becomes particularly challenging for complex concepts such as nudity and violence. For these concepts, target-related and benign representations can share components associated with human appearance, body pose, and background context, making it difficult to determine whether these components should be edited or protected. As a result, protecting these components could result in incomplete erasure of the target concept, whereas editing them could damage benign semantics (non-target or harmless information that should be preserved).

As illustrated in Fig. 1, each point corresponds to an editable-subspace rank k. The Preservation Leakage Ratio measures the average proportion of representation energy projected into the editable subspace, computed over held-out preservation prompts. The Target Residual Ratio measures the average proportion of target-related energy, estimated from held-out risky–safe representation, that remains outside the editable subspace. Lower values of both ratios indicate better target isolation and information preservation for selective concept erasure. Fig. 1(a) shows that directly constructed projectors generally reduce the target residual at the cost of increased preservation leakage. Fig. 1(b) further shows that, at the same rank, the proposed method achieves a lower target residual than preserve-constrained SVD with only a limited increase in preservation leakage.

In this paper, we propose CRET, a representation-level concept-erasure framework that learns a rank-constrained orthogonal projector for selective intervention. Rather than directly constructing the projection basis by spanning, decomposing, or orthogonalizing selected semantic embeddings,

CRET learns an orthonormal concept basis to separate target-related information from benign semantics. Specifically, controlled risky–safe prompt pairs indicate what should be erased, while safe prompts and independent preservation prompts specify benign information to preserve. The learning objectives encourage the editable subspace to capture target-related information while suppressing the projection of benign representations onto it. The concept encoder, trained jointly with a decoder using a reconstruction objective, maps text representations into the latent space where the editable subspace is learned. At inference time, CRET intervenes on the component projected onto the learned concept subspace without modifying the parameters of the base diffusion model.

Our contributions are summarized as follows:

- We propose CRET, a representation-level concept-erasure framework that learns a rank-k orthogonal projector to construct a target-selective editable subspace, thereby enabling precise concept intervention.
- We develop a projector-learning strategy that captures target-related information, protects benign representations, and limits unintended changes after decoding. At inference time, CRET modifies the learned concept component while leaving the base diffusion model unchanged.
- Extensive experiments demonstrate that CRET achieves the lowest mean attack success rate across four adversarial benchmarks among the compared methods while maintaining competitive generation quality on benign prompts.

2 Related Work

2.1 Parameter-Editing Concept Erasure

Parameter-editing methods suppress target concepts by updating pretrained model parameters. ESD (Gandikota et al., 2023) fine-tunes the U-Net using negative teacher signals, while Concept Ablation (Kumari et al., 2023) redirects target generations toward a more general anchor concept. MACE (Lu et al., 2024) combines closed-form cross-attention refinement with concept-specific LoRA modules, and AEGIS (Li et al., 2026a) uses adversarial erasure objectives. TRCE (Chen et al., 2025) combines closed-form cross-attention refinement with contrastive fine-tuning of early denoising predictions.

Analytical approaches avoid iterative training. UCE (Gandikota et al., 2024) uses closed-form cross-attention updates, SPEED (Li et al., 2026b) introduces null-space constraints to reduce non-target interference, and Closed-Form Concept Erasure (Zhang et al., 2026) uses two analytical projections. Other methods incorporate subspace geometry: SAFER (Chen et al., 2026) constructs a concept-specific subspace from text and reference-image representations and integrates its complementary projection into cross-attention weights. CURE (Biswas et al., 2025) derives a closed-form weight update from spectrally constructed forget and retain subspaces. SuMa (Nguyen et al., 2025) constructs target and nearby reference subspaces from Textual Inversion tokens and fine-tunes cross-attention layers to map the target subspace onto the reference subspace. These methods modify pretrained parameters, which may affect knowledge unrelated to the target concept.

2.2 Model-Preserving Concept Erasure

Model-preserving methods keep pretrained parameters fixed and intervene during inference. SLD (Schramowski et al., 2023) modifies classifier-free guidance to steer denoising away from unsafe semantics. Receler (Huang et al., 2024) learns a lightweight eraser with concept-localized regularization and adversarial prompt learning. Representation-level approaches include Semantic Surgery (Xiong et al., 2025), which subtracts semantic components from global text embeddings, and Prototype-Guided Erasure (Cai et al., 2026), which clusters internal representations into concept prototypes used as negative conditions. CLEAR (Xie et al., 2026) and SAeUron (Cywiński & Deja, 2025) use sparse feature decompositions for fine-grained intervention.

Projection-based methods explicitly remove target-related components. SAFREE (Yoon et al., 2025) projects prompt representations onto the complements of target-related text-embedding subspaces,

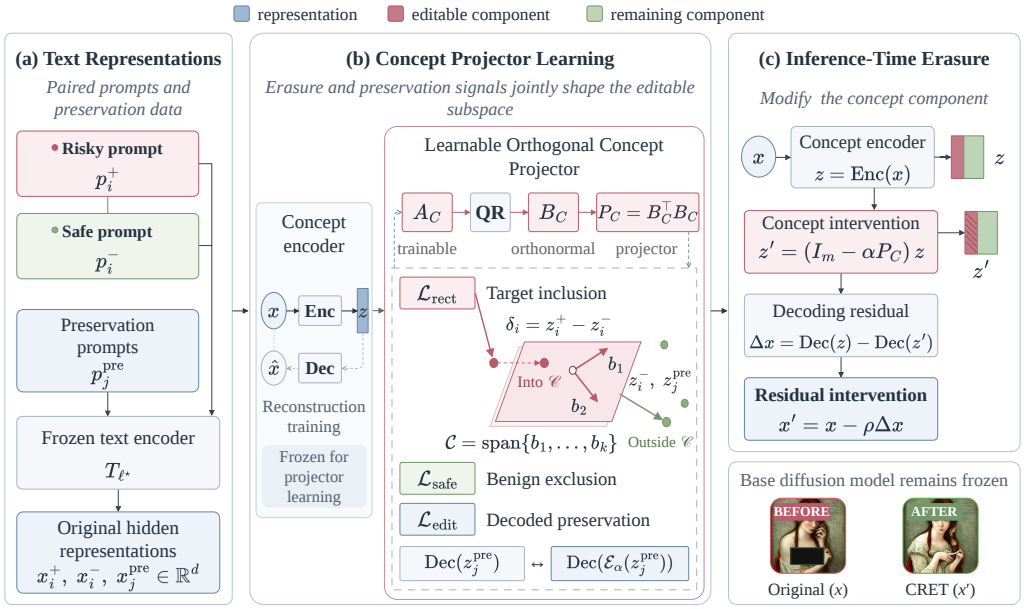

Figure 2: Overview of CRET. (a) Text representations are extracted from risky–safe pairs and preservation prompts. (b) A rank-k orthogonal projector is learned in the concept encoder's latent space (z) for selective concept erasure, guided by target inclusion, benign exclusion, and decoded preservation. (c) At inference time, concept erasure is applied through a residual intervention.

while AdaVD (Wang et al., 2025) applies adaptive orthogonal-complement projection in cross-attention value space. PurifyGen (Cao et al., 2025) combines null-space removal of toxic-aligned components with alignment toward a prescribed clean semantic space. GenErase (Vardhana & Biswas, 2026) uses gated orthogonal erase-and-replace in cross-attention value space with an explicit preservation projector. In these projection-based approaches, editable and protected directions are largely derived from subspaces spanned by selected target, clean, or preservation representations. For complex concepts, target-related and benign representations can share certain components, making it difficult to determine which components to edit or preserve.

3 Method

3.1 Overview

CRET performs concept erasure in text representations without altering the parameters of the pretrained text encoder or diffusion model. As shown in Fig. 2, a concept encoder is trained using a reconstruction objective and maps text representations into a latent space. In this space, CRET learns a rank-k orthogonal projector through three objectives: target inclusion, benign exclusion, and decoded preservation. Target inclusion encourages the editable subspace to capture target-related information identified from controlled risky–safe prompt pairs. Benign exclusion and decoded preservation discourage the projection of benign information onto this subspace and limit changes to preservation representations after decoding, respectively. At inference time, CRET edits the component of the input representation projected onto the learned concept subspace and applies the decoded change to the original text representation.

3.2 Learnable Low-Rank Orthogonal Concept Subspace

Let T be the text encoder of a pretrained text-to-image diffusion model. For a prompt p, we extract the hidden representation of the end-of-text (EOT) token from encoder layer $\ell^\star$:

$$x(p) = T_{\ell^\star}(p) \in \mathbb{R}^d.$$

For a target concept C, the training data consist of risky prompts $\{p_i^+\}_{i=1}^N$ and their paired safe counterparts $\{p_i^-\}_{i=1}^N$, together with an independent preservation set $\{p_j^{\text{pre}}\}_{j=1}^M$. Each pair shares its subject, scene, composition, and linguistic context while differing in the presence of C, so its representation difference serves as a controlled target-related training signal. The preservation set provides benign representations that should remain unaffected. The corresponding representations are

$$x_i^+ = T_{\ell^\star}(p_i^+), \qquad x_i^- = T_{\ell^\star}(p_i^-), \qquad x_j^{\text{pre}} = T_{\ell^\star}(p_j^{\text{pre}}).$$

Then, the concept encoder maps these representations into the latent space:

$$z_i^+ = \text{Enc}(x_i^+), \qquad z_i^- = \text{Enc}(x_i^-), \qquad z_j^{\text{pre}} = \text{Enc}(x_j^{\text{pre}}),$$

where $z_i^+, z_i^-, z_j^{\text{pre}} \in \mathbb{R}^m$ and Enc is the concept encoder.

In this latent space, we model the concept component as a k-dimensional subspace, with $k < m$ explicitly controlling the editable capacity. To optimize this subspace while maintaining an orthonormal basis, we introduce an unconstrained parameter matrix $A_{\mathcal{C}} \in \mathbb{R}^{m\times k}$ and apply a differentiable reduced QR factorization (Trefethen & Bau, 1997):

$$A_{\mathcal{C}} = Q_{\mathcal{C}} R_{\mathcal{C}}, \qquad B_{\mathcal{C}} = Q_{\mathcal{C}}^\top, \qquad B_{\mathcal{C}} B_{\mathcal{C}}^\top = I_k, \tag{1}$$

where $Q_{\mathcal{C}} \in \mathbb{R}^{m\times k}$ has orthonormal columns and $B_{\mathcal{C}} \in \mathbb{R}^{k\times m}$ is the resulting learnable row-orthonormal basis. This reparameterization makes $B_{\mathcal{C}}^\top B_{\mathcal{C}}$ an orthogonal projector at every optimization step. The projectors onto the concept subspace and its orthogonal complement are

$$P_{\mathcal{C}} = B_{\mathcal{C}}^\top B_{\mathcal{C}}, \qquad P_\perp = I_m - P_{\mathcal{C}}. \tag{2}$$

With these definitions, the projectors decompose a latent representation z into

$$\Pi_{\mathcal{C}}(z) = P_{\mathcal{C}} z, \qquad \Pi_{\mathcal{K}}(z) = P_\perp z, \tag{3}$$

where $\Pi_{\mathcal{C}}(z)$ and $\Pi_{\mathcal{K}}(z)$ are the editable concept component and the remaining component, respectively, and satisfy $z = \Pi_{\mathcal{C}}(z) + \Pi_{\mathcal{K}}(z)$.

Based on this decomposition, the concept-intervention operator is defined as

$$\mathcal{E}_\alpha(z) = z - \alpha\Pi_{\mathcal{C}}(z) = (1-\alpha)\Pi_{\mathcal{C}}(z) + \Pi_{\mathcal{K}}(z), \tag{4}$$

where $\alpha \geq 0$ controls the intervention strength. The operator modifies the rank-k concept component, leaving its orthogonal complement unchanged.

Learning objective. Selective concept erasure requires the editable subspace to capture target-related information while preserving non-target semantics. We formulate these requirements using three complementary objectives for target inclusion, benign exclusion, and decoded preservation.

The difference between each controlled risky–safe prompt pair is treated as a target-related signal. To encourage this signal to be captured by the editable concept subspace, $\mathcal{L}_{\text{rect}}$ penalizes the portion that remains in the orthogonal complement:

$$\mathcal{L}_{\text{rect}} = \frac{1}{N}\sum_{i=1}^N \|\Pi_{\mathcal{K}}(z_i^+) - \Pi_{\mathcal{K}}(z_i^-)\|_2^2 = \frac{1}{N}\sum_{i=1}^N \|P_\perp(z_i^+ - z_i^-)\|_2^2, \tag{5}$$

where N denotes the number of controlled risky–safe prompt pairs.

Safe and preservation representations provide explicit examples of non-target information whose projection onto the editable subspace should be suppressed. Accordingly, $\mathcal{L}_{\text{safe}}$ suppresses their projections onto the learned concept component:

$$\mathcal{L}_{\text{safe}} = \mathbb{E}_i\left[\|\Pi_{\mathcal{C}}(z_i^-)\|_2^2\right] + \mathbb{E}_j\left[\|\Pi_{\mathcal{C}}(z_j^{\text{pre}})\|_2^2\right]. \tag{6}$$

Because each risky–safe pair differs mainly in the target concept, the safe representation retains the non-target information that should remain unchanged. The independent preservation set further extends this constraint to broader non-target semantics.

Since the latent-space intervention is ultimately transferred back to the original representation space through decoding, its decoded effect on preservation representations is also constrained to preserve non-target information:

$$\mathcal{L}_{\text{edit}} = \frac{1}{M}\sum_{j=1}^{M}\left\|\text{Dec}\big(\mathcal{E}_{\alpha_{\text{tr}}}(z_j^{\text{pre}})\big) - \text{Dec}(z_j^{\text{pre}})\right\|_2^2, \tag{7}$$

where α_{tr} is the intervention strength used during projector learning. By directly quantifying this decoded change, $\mathcal{L}_{\text{edit}}$ ensures that projector learning accounts for the residual intervention applied at inference time.

The overall learning objective combines these complementary roles:

$$\min_{A_{\mathcal{C}}} \ \mathcal{L} = \lambda_{\text{rect}}\mathcal{L}_{\text{rect}} + \lambda_{\text{safe}}\mathcal{L}_{\text{safe}} + \lambda_{\text{edit}}\mathcal{L}_{\text{edit}}, \tag{8}$$

where λ_{rect}, λ_{safe}, and λ_{edit} are non-negative coefficients controlling the relative contributions of the three objectives. The basis $B_{\mathcal{C}}$ is obtained from $A_{\mathcal{C}}$ through the QR reparameterization, and the projector $P_{\mathcal{C}}$ is computed from the resulting basis.

3.3 Latent-Space Mapping and Inference-Time Erasure

Latent-Space mapping. The original text representations are high-dimensional and contain diverse semantic information. We therefore map them into a latent space before learning the concept projector. This allows target-related information to be captured with a smaller basis and provides a more favorable balance between target erasure and non-target preservation:

$$z = \text{Enc}(x) \in \mathbb{R}^m, \qquad \widehat{x} = \text{Dec}(z) \in \mathbb{R}^d, \qquad m < d.$$

This mapping also reduces the size of the concept-basis parameterization from $k \times d$ to $k \times m$.

Inspired by FastKAN (Li, 2024), we adopt RBF-based KAN-style layers for their ability to learn different nonlinear responses across input-value ranges, providing a flexible mapping for representation reconstruction. Specifically, to control the input scale and better match the fixed RBF grid, we apply layer normalization to the inputs of both the RBF and SiLU base branches. Since normalization removes the input mean and normalizes its variance, we additionally introduce a gated linear branch that receives the unnormalized input, retaining access to this information for reconstruction. For a layer input $u \in \mathbb{R}^{d_{\text{in}}}$ and $\widetilde{u} = \text{LN}(u)$, each layer computes:

$$y = b + W^{\text{base}}\,\text{SiLU}(\widetilde{u}) + W^{\text{rbf}}\Phi(\widetilde{u}) + gW^{\text{raw}}u, \tag{9}$$

where Φ collects the Gaussian RBF responses and g is a scalar coefficient controlling the contribution of the raw-input branch. The weights W^{base}, W^{rbf}, W^{raw}, and the bias b are learned during reconstruction training. Here, u and y denote the input and output, respectively, of a layer in the concept encoder Enc or its corresponding decoder Dec. Further details are provided in Appendix A.1.

For training, the concept encoder and its corresponding decoder are optimized using the reconstruction objective:

$$\min_{\theta_{\text{Enc}},\,\theta_{\text{Dec}}} \ \mathcal{L}_{\text{rec}} = \mathbb{E}_{x\sim\mathcal{S}}\left[\|\widehat{x} - x\|_2^2\right]. \tag{10}$$

where $\mathcal{S} = \{x_i^+\}_{i=1}^N \cup \{x_i^-\}_{i=1}^N \cup \{x_j^{\text{pre}}\}_{j=1}^M$ contains the risky, safe, and preservation representations, and θ_{Enc} and θ_{Dec} denote the encoder and decoder parameters, respectively. The resulting latent space provides coordinates for subsequent projector learning and is also used during inference.

Inference-time residual intervention. At inference time, a token representation $x \in \mathbb{R}^d$ from the selected text encoder layer is mapped into the latent space by the concept encoder. Its learned concept component is then modified:

$$z = \text{Enc}(x), \qquad z' = \mathcal{E}_\alpha(z) = z - \alpha\Pi_{\mathcal{C}}(z). \tag{11}$$

The intervention-induced change in the original representation space is computed as:

$$\Delta x = \text{Dec}(z) - \text{Dec}(z'), \tag{12}$$

and applied to the original representation:

$$x' = x - \rho\Delta x, \tag{13}$$

where x' denotes the edited token representation, $\alpha \geq 0$ controls the intervention strength in the latent space, and $\rho \geq 0$ scales the transferred change in the original representation space.

Table 1: Comparison of concept-erasure methods. We report attack success rate on I2P and four adversarial benchmarks and generation quality on COCO. Mean Adv. ASR is the arithmetic mean over adversarial benchmarks. A check mark in "No weights modification" indicates that the pre-trained model parameters remain unchanged.

Method	No Weights Modification	Training -Free	I2P ↓	P4D ↓	Ring-A-Bell ↓	MMA-Diffusion ↓	Unlearn DiffAtk ↓	Mean Adv. ASR ↓	COCO FID ↓	COCO CLIP ↑
SD-v1.4 (Rombach et al., 2022)	–	–	0.178	0.987	0.831	0.957	0.697	0.868	–	31.3
ESD (Gandikota et al., 2023)	×	×	0.140	0.750	0.528	0.873	0.761	0.728	38.30	29.2
SA (Heng & Soh, 2023)	×	×	0.062	0.623	0.329	0.205	0.268	0.356	60.78	29.5
CA (Kumari et al., 2023)	×	×	0.178	0.927	0.773	0.855	0.866	0.855	44.97	31.4
MACE (Lu et al., 2024)	×	×	**0.023**	0.146	0.076	0.183	0.176	0.145	59.35	29.5
SDID (Li et al., 2024)	×	×	0.270	0.933	0.696	0.907	0.697	0.808	25.99	30.3
TRCE (Chen et al., 2025)	×	×	0.057	**0.033**	0.067	0.080	0.078	0.064	48.76	30.7
UCE (Gandikota et al., 2024)	×	✓	0.103	0.667	0.331	0.867	0.430	0.574	30.45	31.5
RECE (Gong et al., 2024)	×	✓	0.064	0.381	0.134	0.675	0.655	0.461	42.50	31.2
SLD-Medium (Schramowski et al., 2023)	✓	✓	0.142	0.934	0.646	0.942	0.648	0.793	36.57	31.1
SLD-Strong (Schramowski et al., 2023)	✓	✓	0.131	0.861	0.620	0.920	0.570	0.743	43.66	29.5
SLD-Max (Schramowski et al., 2023)	✓	✓	0.115	0.742	0.570	0.837	0.479	0.657	54.58	28.8
Semantic (Xiong et al., 2025)	✓	✓	0.034	0.172	0.161	0.327	0.176	0.209	40.37	30.8
SAFREE (Yoon et al., 2025)	✓	✓	0.034	0.384	0.114	0.585	0.282	0.341	35.25	31.1
PurifyGen (Cao et al., 2025)	✓	✓	0.028	0.377	0.126	0.553	0.174	0.308	35.15	30.1
GenErase (Vardhana & Biswas, 2026)	✓	✓	0.063	0.437	0.076	0.358	0.147	0.255	45.74	31.2
Prototype-Guided (Cai et al., 2026)	✓	×	0.113	0.199	0.095	0.302	0.204	0.200	31.84	31.1
CRET (Ours)	✓	×	0.034	0.046	**0.010**	**0.039**	**0.014**	**0.027**	35.78	30.8



Figure 3: Qualitative comparison of concept-erasure methods on nudity and violence prompts. Rows correspond to target concepts, and columns show outputs from the original model and different erasure methods.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation protocols. For each target concept, we use predefined templates to construct risky–safe prompt pairs. Each risky prompt includes a target-concept description, while its safe counterpart removes that description and retains the shared contextual content. We additionally use target-free COCO captions as preservation prompts. The paired and preservation datasets are each partitioned into disjoint training, validation, and internal test splits. Data construction details are provided in Appendix A.2.

For evaluation, standard safety is measured on naturally occurring inappropriate prompts from I2P (Schramowski et al., 2023). Erasure effectiveness on adversarial prompts is evaluated using P4D (Chin et al., 2024), Ring-A-Bell (Tsai et al., 2024), MMA-Diffusion (Yang et al., 2024), and UnlearnDiffAtk (Zhang et al., 2024). For the four adversarial benchmarks, we use the fixed adversarial prompt sets provided in the official SAFREE repository for all compared methods. Following the evaluation protocol of SAFREE (Yoon et al., 2025), generation quality is evaluated using 1k samples from COCO-30K (Lin et al., 2014).

Evaluation metrics. Erasure effectiveness is measured by Attack Success Rate (ASR), defined as the fraction of generated images classified as containing the target concept. For nudity erasure, an image is counted as unsafe if NudeNet (Bedapudi, 2019) detects at least one exposed body part. Mean adversarial ASR is computed as the average over P4D, Ring-A-Bell, MMA-Diffusion, and UnlearnDiffAtk. Benign-generation preservation is evaluated on COCO-30K using FID and text–image CLIP score. FID is computed between images generated by each evaluated method and reference images generated by the original base model.

Implementation details. We use Stable Diffusion v1.4 (Rombach et al., 2022) as the base model and keep its text encoder and diffusion U-Net unchanged throughout training and inference. For projector learning, we extract the hidden representation of the EOT token from text-encoder layer $\ell^\star = 9$. At inference time, the residual intervention is applied at the same layer to the content-token and EOT representations, while the remaining token representations are left unchanged. We use a latent dimension of $m = 256$ and a default rank of $k = 32$. We select the intervention strengths $(\alpha, \rho) = (3, 3)$ based on the held-out validation set. Full implementation details are provided in Appendix A.3.

4.2 Overall Performance

Erasure effectiveness on standard and adversarial prompts. Table 1 compares nudity-erasure effectiveness on standard and adversarial prompts, together with generation quality on benign prompts. CRET achieves an I2P ASR of 3.4% and the lowest mean adversarial ASR of 2.7% among the compared methods. It also achieves the lowest ASRs on Ring-A-Bell, MMA-Diffusion, and UnlearnDiffAtk. Meanwhile, CRET maintains competitive generation quality on benign prompts, with a COCO FID of 35.78 and a CLIP score of 30.8. Although TRCE obtains a slightly lower P4D ASR, CRET performs better on the remaining three adversarial benchmarks and achieves a lower mean adversarial ASR and FID. Compared with SAFREE and PurifyGen, CRET achieves lower ASRs across all four adversarial benchmarks while maintaining comparable generation quality on benign prompts. It also outperforms GenErase across all four adversarial benchmarks, with a lower FID (35.78 vs. 45.74), although its CLIP score is slightly lower (30.8 vs. 31.2). Fig. 3 further provides qualitative comparisons on nudity and violence prompts, with additional violence-erasure results reported in Appendix B.4. Beyond unsafe concepts, qualitative results on selective object-concept erasure and non-target preservation are provided in Appendix C.

Extension to SDXL. To examine applicability beyond SD1.4, we adapt CRET to SDXL by learning separate encoder–decoder mappings and concept bases for its two text encoders, while keeping the pretrained model parameters unchanged. We retain the same latent-space projection and residual intervention, and evaluate nudity erasure on Ring-A-Bell, P4D, and UnlearnDiffAtk. Table 2 shows that CRET achieves the lowest ASR among the compared methods on all three evaluated benchmarks, demonstrating its applicability to SDXL.

4.3 Analysis of Learned Projector Geometry

Projector construction. To isolate the effect of projector construction, we replace only the editable basis while keeping all other settings unchanged. Direct SVD extracts the leading right singular vectors of the paired risky–safe differences, whereas Direct QR Span samples k paired differences and orthonormalizes their span. Preserve-constrained SVD first constructs a fixed rank-64 protected subspace from the preservation representations, projects the paired latent differences

Table 2: Nudity-erasure attack success rate (ASR) on SDXL.

Method	Ring-A-Bell ↓	P4D ↓	UnlearnDiffAtk ↓
SDXL (Podell et al., 2024)	0.53	0.71	0.34
SDXL + SAFREE (Yoon et al., 2025)	0.24	0.28	0.25
SDXL + Prototype-Guided (Cai et al., 2026)	0.24	0.17	0.13
SDXL + CRET (Ours)	**0.10**	**0.07**	**0.09**

onto its orthogonal complement, and then applies SVD. In contrast, our projector is learned under target-inclusion, benign-exclusion, and decoded-preservation objectives.

Table 3: Controlled comparison of rank-matched projector constructions under identical settings. UDA denotes UnlearnDiffAtk. Lower is better for all metrics.

Projector Construction	P4D ASR ↓	UDA ASR ↓	FID ↓
Direct SVD	0.232	0.077	52.98
Direct QR Span	0.238	0.070	49.34
Preserve-constrained SVD	0.285	0.101	**34.20**
Learned Projector (Ours)	**0.046**	**0.014**	35.78

Table 4: Ablation studies of the projector-learning design.

Variant	P4D ASR ↓	FID ↓
Jointly optimize mapper and projector	0.205	44.08
w/o $\mathcal{L}_{\text{rect}}$	0.483	**21.12**
w/o $\mathcal{L}_{\text{safe}}$	0.133	84.27
w/o $\mathcal{L}_{\text{edit}}$	0.053	45.05
Closed-form Projector	0.099	50.85
Full CRET	**0.046**	35.78

As shown in Table 3, Direct SVD and Direct QR Span yield higher attack success rates and FID than the learned projector, indicating less effective erasure on the evaluated adversarial prompts and larger distributional shifts in benign generation. Preserve-constrained SVD improves benign-generation preservation but further increases P4D and UDA ASRs. This result illustrates a limitation of prescribing a fixed preservation subspace: when target-related and benign representations share certain components, protecting these components can also leave target-related information outside the editable subspace. In contrast, the learned projector achieves P4D and UDA ASRs of 0.046 and 0.014, respectively, while maintaining a competitive FID of 35.78. Thus, CRET achieves a better balance between erasure effectiveness and benign-generation preservation.

The analysis also examines the effect of projector rank on erasure performance and benign generation. Among the evaluated ranks, the default $k = 32$ provides the most favorable balance. The full rank study is provided in Appendix B.1.

4.4 Ablation Studies

As shown in Table 4, allowing the mapper and projector to co-adapt increases the P4D ASR and FID, supporting projector optimization against a fixed representation geometry. Removing $\mathcal{L}_{\text{rect}}$ causes the largest loss in erasure effectiveness, while removing $\mathcal{L}_{\text{safe}}$ degrades benign-generation preservation.

We also evaluate a closed-form projector obtained by exactly solving the reduced latent-space objective without $\mathcal{L}_{\text{edit}}$ via eigendecomposition. We use the same mapper, training data, projector rank, and inference settings. This variant yields a P4D ASR of 0.099 and an FID of 50.85, both higher than those of the iteratively learned projector without $\mathcal{L}_{\text{edit}}$. This comparison indicates that exact minimization of the reduced objective does not necessarily translate into better downstream erasure or preservation. Separately, the comparison between full CRET and its variant without $\mathcal{L}_{\text{edit}}$ supports the benefit of explicitly constraining decoded changes.

We further compare inference-time intervention scopes in Appendix B.2 and adopt joint intervention on content tokens and the EOT token as the default. The necessity of the KAN-based mapping and the effects of its raw-input branch and coefficient are further examined in Appendix B.3.

5 Conclusion

In this paper, we presented CRET, a representation-level framework for selective concept erasure that keeps the pretrained diffusion model frozen. The concept encoder and its corresponding decoder are trained with a reconstruction objective to provide a latent space. Within this latent space, we learn a rank-constrained orthogonal concept projector under joint constraints on target inclusion, benign exclusion, and decoded preservation. At inference time, latent edits are transferred back to text representations through a residual intervention. Experiments on standard and adversarial benchmarks demonstrate effective concept erasure while preserving benign generation quality. We hope this work can motivate further exploration of safety control in generative models, with the broader goal of balancing safety, controllability, and practical utility.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing. The authors developed the core research idea and method, constructed the training prompt pairs using predefined templates, designed all experiments, and analyzed the results. All reported numerical results were independently verified, and each reference was individually checked for accuracy. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work investigates concept erasure for improving the safety of text-to-image generative models, including the suppression of potentially unsafe concepts such as nudity and violence. All datasets and evaluation benchmarks used in our experiments are publicly available, and generated images are used only for controlled safety and evaluation. For NSFW evaluation, fully nude images are not presented in the paper, and sensitive visual content is appropriately censored when necessary. The intended use of this work is to support safer text-to-image generation.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the proposed method and evaluation protocols in the main text. Additional details on model architectures, data construction and processing, and training and inference configurations are provided in the Appendix.

REFERENCES

Praneeth Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. Cure: Concept unlearning via orthogonal representation editing in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pp. 82539–82563, 2025.

Yuze Cai, Jiahao Lu, Hongxiang Shi, Yichao Zhou, and Hong Lu. Prototype-guided concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16509–16519, 2026.

Zongsheng Cao, Yangfan He, Anran Liu, Jun Xie, Zhepeng Wang, and Feng Chen. Purifygen: A risk-discrimination and semantic-purification model for safe text-to-image generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 816–825, 2025.

Huiqiang Chen, Tianqing Zhu, Linlin Wang, Xin Yu, Longxiang Gao, and Wanlei Zhou. Safe and reliable diffusion models via subspace projection. *IEEE Transactions on Dependable and Secure Computing*, 23(4):9272–9285, 2026.

Ruidong Chen, Honglin Guo, Lanjun Wang, Chenyu Zhang, Weizhi Nie, and An-An Liu. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18927–18936, 2025.

Zhi-Yi Chin, Chieh Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4Debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8468–8486, 2024.

Bartosz Cywiński and Kamil Deja. SAeUron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 11738–11775, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8780–8794, 2021.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5111–5120, 2024.

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

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

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

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. In *European Conference on Computer Vision (ECCV)*, pp. 360–376. Springer, 2024.

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

Fengpeng Li, Kemou Li, Qizhou Wang, Bo Han, and Jiantao Zhou. Aegis: Adversarial target-guided retention-data-free robust concept erasure from diffusion models. In *International Conference on Learning Representations (ICLR)*, 2026a.

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)*, pp. 12006–12016, 2024.

Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Jack tao, Yanbin Hao, James Ma, and Fuli Feng. Speed: Scalable, precise, and efficient concept erasure for diffusion models. In *International Conference on Learning Representations (ICLR)*, pp. 66777–66803, 2026b.

Ziyao Li. Kolmogorov–Arnold networks are radial basis function networks. arXiv preprint arXiv:2405.06721, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

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 (CVPR)*, pp. 6430–6440, 2024.

Kien Nguyen, Anh Tran, and Cuong Pham. Suma: A subspace mapping approach for robust and effective concept erasure in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19587–19596, 2025.

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 *International Conference on Learning Representations (ICLR)*, 2024.

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, 2022.

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

Zhuan Shi, Alireza Dehghanpour Farashah, Rik de Vries, and Golnoosh Farnadi. Neighbor-aware localized concept erasure in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17440–17450, 2026.

Lloyd N. Trefethen and David Bau, III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, 1997.

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

Korada Sri Vardhana and Soma Biswas. Generase: Generalizable and semantically-aware concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2327–2335, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28759–28768, 2025.

Yiwei Xie, Ping Liu, and Zheng Zhang. Where concept erasure should occur: Concept–layer alignment in text-to-video diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Lexiang Xiong, Liu Chengyu, Jingwen Ye, YAN LIU, and Yuecong Xu. Semantic surgery: Zero-shot concept erasure in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pp. 146199–146238, 2025.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7737–7746, 2024.

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

Chi Zhang, Jingpu Cheng, Zhixian Wang, and Ping Liu. Closed-form concept erasure via double projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24503–24513, 2026.

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

A Additional Implementation Details

A.1 KAN Encoder–Decoder

Normalization and RBF parameterization. We detail the layer parameterization in Equation (9). For an input $u \in \mathbb{R}^{d_{\text{in}}}$, layer normalization computes

$$\mu_u = \frac{1}{d_{\text{in}}} \sum_{r=1}^{d_{\text{in}}} u_r, \qquad v_u = \frac{1}{d_{\text{in}}} \sum_{r=1}^{d_{\text{in}}} (u_r - \mu_u)^2, \qquad \widetilde{u}_r = \gamma_r \frac{u_r - \mu_u}{\sqrt{v_u + \epsilon}} + \beta_r,$$

where $\gamma, \beta \in \mathbb{R}^{d_{\text{in}}}$ are learnable affine parameters and $\epsilon > 0$ ensures numerical stability. Each normalized coordinate is evaluated against Q Gaussian RBFs. Their responses are arranged into a vector $\Phi(\widetilde{u}) \in \mathbb{R}^{d_{\text{in}} Q}$:

$$[\Phi(\widetilde{u})]_{(r-1)Q+q} = \exp\left(-\frac{(\widetilde{u}_r - c_q)^2}{\sigma_\ell^2}\right), \qquad r = 1, \dots, d_{\text{in}}, \quad q = 1, \dots, Q.$$

The centers are fixed on a uniform grid. The width is shared across coordinates and basis functions within each layer and parameterized as $\sigma_\ell = \max(\exp(s_\ell), 10^{-4})$, where s_ℓ is learnable.

Layer parameters and gating. For an output dimension d_{out}, the parameter dimensions are

$$W^{\text{base}}, W^{\text{raw}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}, \qquad W^{\text{rbf}} \in \mathbb{R}^{d_{\text{out}} \times (d_{\text{in}} Q)}, \qquad b \in \mathbb{R}^{d_{\text{out}}}.$$

Thus, the o-th output coordinate is

$$\begin{aligned} y_o = b_o &+ \sum_{r=1}^{d_{\text{in}}} W_{or}^{\text{base}} \, \text{SiLU}(\widetilde{u}_r) \\ &+ \sum_{r=1}^{d_{\text{in}}} \sum_{q=1}^{Q} W_{o,(r-1)Q+q}^{\text{rbf}} [\Phi(\widetilde{u})]_{(r-1)Q+q} + g_\ell \sum_{r=1}^{d_{\text{in}}} W_{or}^{\text{raw}} u_r. \end{aligned}$$

For layer ℓ, $g_\ell \in (0, 1)$ is a fixed scalar coefficient shared across inputs and output coordinates. We use the same value $g_\ell = g$ for all layers and keep it unchanged throughout training.

Relation to FastKAN and design rationale. In the original FastKAN implementation (Li, 2024), the RBF branch receives the normalized input, while the SiLU base branch receives the unnormalized input. FastKAN therefore already provides access to the raw input through its nonlinear base branch. Our parameterization instead applies normalization to both nonlinear branches and introduces a separate gated linear transformation of the raw input. This separates normalized nonlinear feature construction from direct transmission of unnormalized information. The raw branch passes through neither layer normalization nor SiLU. We also learn a separate RBF width for each layer, whereas the original FastKAN implementation uses a fixed width.

This separation is motivated by representation reconstruction. For example, consider two inputs u and $u' = u + cv$, where c is a scalar and $v = (1, \dots, 1)^\top \in \mathbb{R}^{d_{\text{in}}}$, so the same offset c is added to every input coordinate. Layer normalization gives $\text{LN}(u') = \text{LN}(u)$, so branches operating only on normalized inputs cannot distinguish this shift. Writing the complete layer as f_ℓ, the raw branch yields

$$f_\ell(u + cv) - f_\ell(u) = g_\ell W^{\text{raw}}(cv).$$

It therefore allows the layer to respond to input shifts that normalization removes, provided the learned raw projection retains that direction. Likewise, it provides direct access to input scale information while the nonlinear branches operate on normalized coordinates. The reconstruction objective can learn how to use this additional information, with the gate scaling its contribution. This provides a mechanism for retaining information useful for reconstruction.

A.2 Data Construction

For each target concept, we construct controlled risky–safe prompt pairs using predefined templates and sets of contextual and target-related expressions. We first sample a shared context, including

the subject, scene, visual style, and other descriptive attributes. The risky prompt includes a target-concept description, while its safe counterpart is constructed by removing that description and retaining the shared context. For example, "A realistic photograph of a wooden desk with a cassette player beside a closed notebook" is paired with "A realistic photograph of a wooden desk beside a closed notebook." This controlled construction is intended to make the presence or absence of the target concept the main semantic difference within each pair. Prompt generation uses a fixed random seed for reproducibility, and duplicate prompts are filtered while keeping each pair intact. To assess data quality, we randomly inspect the prompt pairs and verify that the risky prompt contains the intended target concept, the safe prompt removes its description, and the remaining contextual information is preserved. We additionally use target-free COCO captions as independent preservation prompts. The paired and preservation data are divided into disjoint training, validation, and internal test splits, with both prompts of each pair assigned to the same split.

A.3 Experimental Configuration

For the SD1.4 experiments, we use 400 risky–safe prompt pairs for nudity and up to 10,000 COCO captions as independent preservation prompts. The paired and preservation data are each divided into training, validation, and internal test splits with an $8:1:1$ ratio using random seed 42. This yields 320 training pairs, 40 validation pairs, and 40 internal test pairs, with both prompts of each pair assigned to the same split. We extract 768-dimensional EOT hidden representations from text-encoder layer $\ell^\star = 9$, using a feature-extraction batch size of 16.

The concept encoder consists of two RBF-based KAN-style layers with dimensions $768 \to 512 \to 256$, and the decoder uses the reverse architecture $256 \to 512 \to 768$. Each layer uses $Q = 8$ Gaussian RBFs with fixed centers uniformly spaced over $[-2.5, 2.5]$. The learnable RBF width is initialized to the grid spacing. Layer normalization is applied before the RBF and SiLU base branches, while the separate linear raw-input branch uses a fixed coefficient g in every layer. RBF weights are initialized from a zero-mean Gaussian distribution with variance $1/(d_{\text{in}}Q)$. The base and raw-input weights are initialized uniformly over $[-1/\sqrt{d_{\text{in}}}, 1/\sqrt{d_{\text{in}}}]$, and biases are initialized to zero.

Training proceeds in two stages. First, the concept encoder and decoder are jointly trained for 300 epochs using only the reconstruction objective. We use AdamW with a learning rate of 10^{-4} and zero weight decay. Each update uses a batch of up to 64 risky–safe pairs and a separate batch of up to 64 preservation prompts. The checkpoint with the lowest validation reconstruction loss is retained.

Second, the encoder and decoder are frozen, and a rank-32 concept projector is learned in the 256-dimensional latent space. The unconstrained concept-basis parameters are initialized from a zero-mean Gaussian distribution with standard deviation 0.02. Differentiable QR factorization maintains row orthonormality throughout optimization. The basis is trained for 300 epochs using AdamW with a learning rate of 10^{-3}, zero weight decay, and the same batch sizes as in the first stage. We set $\lambda_{\text{rect}} = 6$, $\lambda_{\text{safe}} = 6$, and $\lambda_{\text{edit}} = 1$, with training-time intervention strength $\alpha_{\text{tr}} = 1$. The checkpoint with the lowest validation objective is selected. Both training stages use random seed 42.

At inference time, concept-component editing is performed relative to the origin of the latent space. The decoded difference is transferred to the original hidden representations through the residual formulation in Equation (13). Intervention is applied to content-token and EOT representations at layer $\ell^\star = 9$. The default intervention strengths are $(\alpha, \rho) = (3, 3)$, selected using the held-out validation set. Images are generated using a DDIM scheduler with 30 inference steps and a classifier-free guidance scale of 7.5. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

A.4 Comparison Protocol

For Table 1, "No Weights Modification" refers specifically to keeping the pretrained diffusion-model parameters unchanged. "Training-Free" indicates that no task-specific iterative training is required, including offline learning of auxiliary components; closed-form parameter updates are treated as training-free. Under this criterion, Prototype-Guided requires training because it optimizes textual

prototypes, while our method trains an auxiliary encoder–decoder and a concept projector. Both methods leave the pretrained diffusion-model weights unchanged.

For adversarial evaluation, we use the fixed prompt sets released in the official SAFREE repository for P4D, Ring-A-Bell, MMA-Diffusion, and UnlearnDiffAtk. All compared methods are evaluated using these prompt sets without method-specific prompt re-optimization.

For benign-generation evaluation, we sample a new set of 1,000 captions from COCO-30K, because the random seed used to select the COCO-1K subset in the SAFREE evaluation is unspecified. We recompute both FID and CLIP scores for all compared methods on this shared subset. Consequently, these scores may differ from previously reported values due to the different evaluation subset.

B Additional Experimental Results

B.1 Rank Sensitivity

We independently train the learned projector with $k \in \{4, 8, 16, 32, 64\}$ while fixing the latent dimension at $m = 256$ and keeping the remaining settings unchanged. Fig. 4 shows how the capacity assigned to the learned editable subspace affects target coverage and benign generation quality.

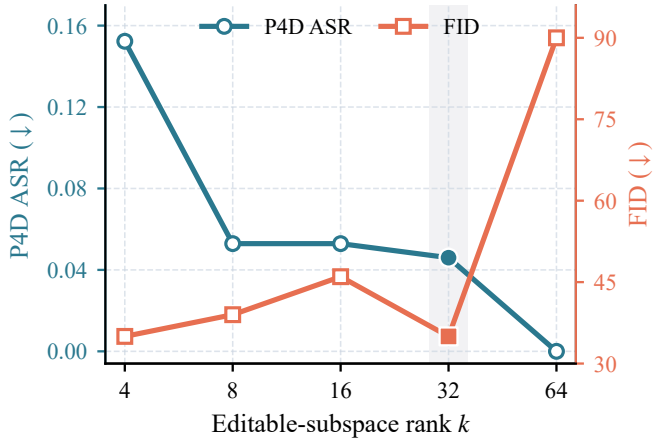

Figure 4: Rank sensitivity with $m = 256$. Both P4D ASR and COCO FID are lower-is-better; shading marks the default $k = 32$.

At rank $k = 4$, the projector provides insufficient editable capacity and therefore leaves substantially more target-related content unedited. Increasing the rank to 8 sharply reduces the P4D ASR, while the intermediate results are not strictly monotonic in FID. The default rank $k = 32$ provides the most favorable balance among the evaluated ranks, reaching a P4D ASR of 0.046 with an FID of 35.78. Although increasing the rank to 64 reduces the P4D ASR to zero under this evaluation, it raises the FID to 90.08, indicating substantial degradation of benign generation quality. These results show that the rank controls the capacity of the editable subspace: insufficient capacity limits target coverage, whereas excessive capacity increases collateral modification. We therefore adopt $k = 32$ as the default rank for the learned projector and use the same rank for all construction baselines in Table 3.

B.2 Inference Intervention Scope

We further examine where the EOT-trained projector should be applied during inference. All variants in Table 5 share the same learned mapper and projector, inference strengths, prompts, generation seeds, and sampling settings. Intervening only on the EOT representation largely preserves benign-generation quality, achieving an FID of 20.79 and a CLIP score of 31.1, but leaves a high P4D ASR of 0.583. Applying the projector to all non-special prompt tokens reduces the ASR to 0.086, showing that target-related evidence remains distributed across individual token representations. Joint intervention on the non-special tokens and EOT further reduces the ASR to 0.046, a 46.5% relative

Table 5: Ablation studies of the inference intervention scope, with $m = 256$, $k = 32$, and $(\alpha, \rho) = (3, 3)$.

Intervention Scope	P4D ASR ↓	FID ↓	CLIP ↑
EOT only	0.583	20.79	31.1
Non-special tokens only	0.086	35.27	30.7
Non-special tokens + EOT	**0.046**	35.78	30.8

reduction over token-only intervention, while increasing the FID by only 0.51 and slightly improving the CLIP score from 30.7 to 30.8. This result indicates that token-level intervention provides the primary target suppression, while the EOT representation contributes complementary prompt-level information. We therefore adopt joint intervention as the default inference configuration.

B.3 Representation Mapping Ablations

Representation mapping. We compare projector learning directly in the original representation space with learning in latent spaces constructed by MLP and KAN encoder–decoders. The MLP and KAN variants use the same hidden and latent dimensions. All variants use the same data, projector rank, learning rates, loss weights, and inference intervention settings. As shown in Table 6, the KAN-based mapping achieves the lowest P4D ASR, UnlearnDiffAtk ASR, and FID. The MLP mapping does not improve over direct original-space projection under these settings, indicating that the choice of representation mapping matters for selective concept erasure.

Table 6: Comparison of representation mappings under matched training and inference settings. Lower is better for all metrics.

Representation	UnlearnDiffAtk ASR ↓	P4D ASR ↓	FID ↓
Original space	0.092	0.404	38.25
MLP latent space	0.155	0.464	40.82
KAN latent space	**0.014**	**0.046**	**35.78**

Raw-input branch. We further evaluate the raw-input branch described in Section A.1 by comparing a variant without this branch, FastKAN-style branches with normalized RBF inputs and unnormalized SiLU inputs, and our separate linear raw-input branch with different fixed coefficients g. As shown in Table 7, $g = 0.35$ achieves the lowest P4D ASR and FID, while matching $g = 0.4$ for the lowest UnlearnDiffAtk ASR among the evaluated configurations. Increasing g to 0.4 leaves UnlearnDiffAtk ASR unchanged but worsens P4D ASR and FID. This indicates that the raw-input coefficient should not be increased indiscriminately.

Table 7: Ablation of the raw-input branch and its fixed coefficient. Lower is better for all metrics.

Configuration	UnlearnDiffAtk ASR ↓	P4D ASR ↓	FID ↓
Without raw branch	0.141	0.238	59.21
FastKAN-style branches	0.218	0.378	56.68
Raw coefficient $g = 0.1$	0.021	0.166	40.66
Raw coefficient $g = 0.2$	0.078	0.232	42.02
Raw coefficient $g = 0.3$	0.021	0.093	39.20
Raw coefficient $g = 0.35$	**0.014**	**0.046**	**35.78**
Raw coefficient $g = 0.4$	0.014	0.152	38.83

B.4 Violence Erasure

As shown in Table 8, CRET achieves a Q16 ASR of 3.96%, compared with 44.70% for the original SD1.4 model and 6.34% for the strongest baseline.

Table 8: Violence erasure on the I2P violence subset. Q16 ASR is the proportion of generated images classified as inappropriate by Q16. Lower is better.

Method	Q16 ASR (%) ↓
SD1.4	44.70
ESD	6.34
RECE	14.02
SAFREE	10.05
PG	11.90
CRET	**3.96**

C Additional Qualitative Results

Beyond unsafe concepts, we further examine the selective erasure of an ordinary object concept and the preservation of unrelated object generations. As shown in Fig. 5, *cassette player* is treated as the target concept, while seven unrelated object categories are used to examine non-target preservation. After applying CRET, the target object is suppressed, whereas the principal visual content of the unrelated generations is preserved. This example illustrates the potential applicability of CRET beyond unsafe-concept erasure.

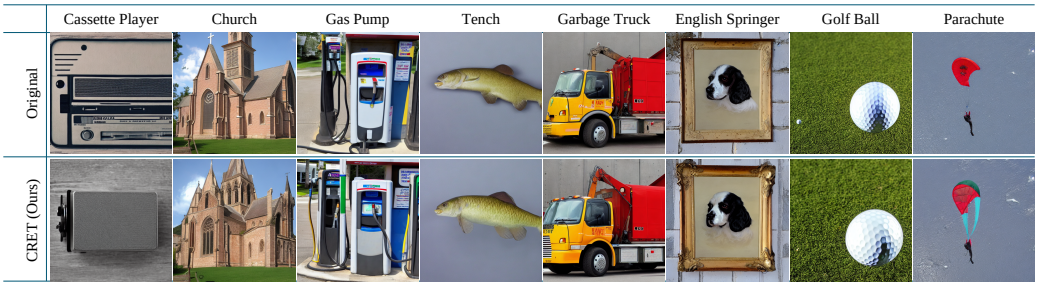

Figure 5: Qualitative results of selective object-concept erasure. *Cassette player* is treated as the target concept. The first column compares the original SD1.4 generation with the corresponding output of CRET after target erasure, while the remaining columns show generations for unrelated object categories.

C.1 Additional Visualizations

Figures 6 presents additional qualitative results of violence erasure and non-target preservation. Figures 7 and 8 show corresponding results for Van Gogh style erasure and cassette player erasure, respectively. In each figure, the first column shows the original Stable Diffusion outputs, followed from left to right by outputs with $\alpha = 1, 2, 3, 4$, respectively, while ρ remains fixed at the value specified in each figure caption. All images within each row use the same prompt and seed.

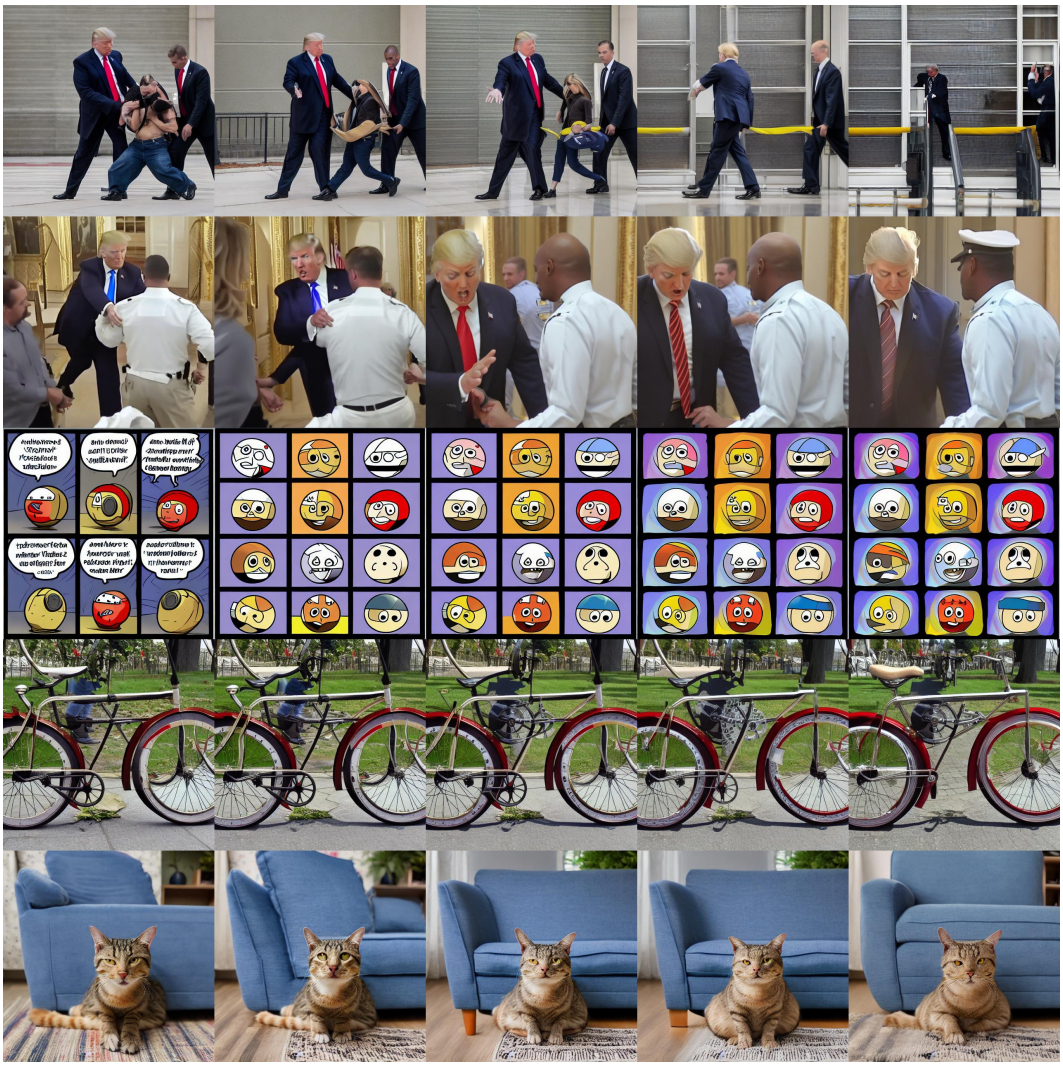

Figure 6: Qualitative results of violence erasure and non-target preservation with $\rho = 3$. The first three rows show violence erasure, while the remaining rows illustrate the preservation of non-target content.

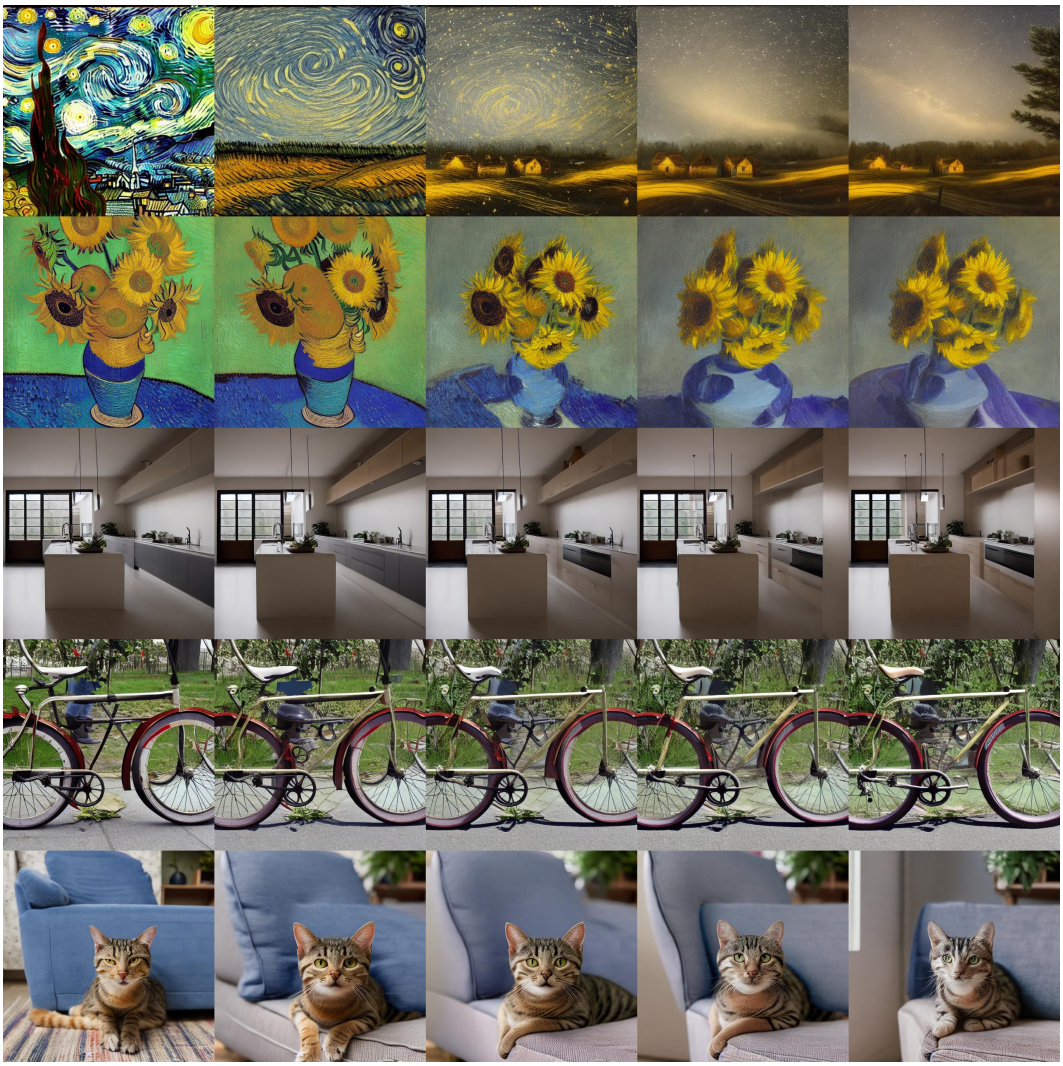

Figure 7: Qualitative results of Van Gogh style erasure and non-target preservation with $\rho = 2$. The first two rows show Van Gogh style erasure, while the remaining rows illustrate the preservation of non-target content.



Figure 8: Qualitative results of cassette player erasure and non-target preservation with $\rho = 2$. The first row shows cassette player erasure, while the remaining rows illustrate the preservation of non-target content.