Closed-Form Concept Erasure via Double Projections

Chi Zhang[1] Jingpu Cheng[1] Zhixian Wang[2] Ping Liu[3]✉
[1]National University of Singapore [2]Shanghai Jiao Tong University [3]University of Nevada, Reno
czhang24@nus.edu.sg chengjingpu@u.nus.edu jd.wzx@sjtu.edu.cn pino.pingliu@gmail.com

Abstract

While modern generative models such as diffusion-based architectures have enabled impressive creative capabilities, they also raise important safety and ethical risks. These concerns have led to growing interest in concept erasure, the process of removing unwanted concepts from model representations. Existing approaches often achieve strong erasure performance but rely on iterative optimization and may inadvertently distort unrelated concepts. In this work, we present a simple yet principled alternative: a linear transformation framework that achieves concept erasure analytically, without any training. Our method adapts a pre-trained model through two sequential, closed-form steps: first, computing a proxy projection of the target concept, and second, applying a constrained transformation within the left null space of known concept directions. This design yields a deterministic and geometrically interpretable procedure for safe, efficient, and theory-grounded concept removal. Across a wide range of experiments, including object and style erasure on multiple Stable Diffusion variants and the flow-matching model (FLUX), our approach matches or surpasses the performance of state-of-the-art methods while preserving non-target concepts more faithfully. Requiring only a few seconds to apply, it offers a lightweight and drop-in tool for controlled model editing, advancing the goal of safer and more responsible generative models. Code is available here.

1. Introduction

The remarkable capabilities of modern generative models, including diffusion-based [32, 59] and flow-based [19, 42, 48] methods, have revolutionized content creation. These systems can produce diverse, high-fidelity images and text from simple prompts, enabling a wide range of creative and practical applications [50, 52, 67]. Yet this power comes with risks: generative models may inadvertently reproduce copyrighted material, generate biased or harmful content, or reveal sensitive information [9, 10]. Such concerns have made concept erasure [23], the selective removal of undesired concepts from model representations, an increasingly important direction for safe and responsible AI.

Existing approaches pursue this goal through a range of mechanisms, including cross-attention layer modifications [23, 24, 45], model pruning strategies [13, 65], regularization-based editing [34], and adversarial-guided erasure [8, 72]. In general, these methods seek to remove target concepts by altering specific model components or parameters such as attention mechanisms, feature representations, or network weights. Collectively, these approaches have proven highly effective at suppressing targeted concepts in complex generative models, demonstrating that such information can indeed be localized and selectively removed from internal representations [23]. This progress has also enabled a variety of beneficial applications [24], including removing unwanted objects or artistic styles, enforcing copyright protection, mitigating harmful content, and promoting fairness in generative outputs.

Yet, in doing so, existing methods may also unintentionally affect other, non-target concepts, degrading the model's overall representational balance. For example, pruning neurons associated with a particular concept [13] can also remove neurons critical for other semantic attributes or generative behaviors, leading to noticeable drops in performance. This raises a central practical question: how can we effectively erase specific concepts while preserving a model's knowledge of non-target concepts?

To address this challenge, we introduce "Concept Erasure with Double Projections" (DP), a principled and efficient framework that explicitly minimizes interference with non-target representations. Instead of relying on iterative optimization or retraining, DP reformulates concept erasure as a pair of analytical projection steps with clear geometric interpretation. The first projection isolates the safe component of a target concept by aligning it with known non-target directions. The second applies a constrained transformation within the left nullspace of preserved representations, ensuring that removing the target concept minimally affects others. Importantly, both steps admit analytically

✉ Corresponding author.

closed-form solutions, yielding a *deterministic, training-free* method that operates in seconds.

We evaluate the proposed method across multiple concept-erasure settings, including object and style erasure, using several variants of Stable Diffusion [52] and the recent flow-matching model FLUX [3, 42]. Across all these architectures, our approach achieves erasure performance comparable to or better than existing state-of-the-art techniques in terms of removing the targeted concepts. More importantly, both qualitative and quantitative results consistently demonstrate that our approach better preserves the remaining non-target concepts, maintaining the overall generative quality and diversity of model outputs.

Overall, our study demonstrates that concept erasure can be formulated and solved efficiently within a *principled geometric framework*. By decoupling the optimization into two analytically solvable steps, the double projection solution achieves both interpretability and practicality, removing unwanted concepts in seconds without retraining or iterative fine-tuning. In practice, such a projection design offers several key advantages:

(1) **Closed-form formulation.** We reformulate concept erasure as an analytically solvable linear transformation problem, providing a one-shot solution with provable guarantees and eliminating any need for retraining.

(2) **Geometric interpretability.** The proposed double-projection design offers a principled geometric perspective that explicitly characterizes how erasure and preservation interact within the representation space.

(3) **Effective erasure and preservation.** Our method achieves state-of-the-art suppression of targeted concepts while minimizing interference with non-target semantics, preserving both visual quality and diversity.

(4) **Cross-model generality.** The framework operates consistently across multiple diffusion and flow-matching architectures, demonstrating robustness and scalability for diverse generative backbones.

2. Related Works

Deep Generative Models and Personalization. Deep generative models have become the foundation of modern image synthesis, with diffusion-based and flow-matching architectures leading recent advances [1, 32, 42, 44, 50, 52]. Diffusion models [32] generate images through iterative denoising from Gaussian noise, guided by learned score functions to produce highly realistic and semantically consistent outputs. Flow-matching methods [42] later introduced deterministic mappings between noise and data distributions, improving sample efficiency by aligning trajectories in a continuous latent space [42, 44]. These advances have enabled the synthesis of high-quality, semantically faithful imagery, driving widespread adoption across creative, industrial, and scientific applications. Building on this progress, personalization techniques have been developed to adapt generative models for user-specific concepts from only a few examples. In diffusion models, approaches such as DreamBooth [53], Textual Inversion [22], and parameter-efficient tuning [39, 57] enable subject-driven generation without retraining the full model. More recently, personalization has extended to flow-based architectures, with classifier-guided adaptation [60] and LoRA-based fine-tuning [18] supporting flexible concept encoding and efficient customization.

Risks and Safety Concerns in Deep Generative Models. Despite their remarkable versatility, generative models introduce serious ethical and safety challenges. One major concern is copyright infringement: large-scale models trained on web data can memorize and reproduce copyrighted works nearly verbatim [11, 58], leading to legal disputes with artists and creators [2]. Another is bias amplification—these models often internalize and reinforce stereotypes present in their training data [5, 16, 46], perpetuating harmful associations related to gender, race, or occupation. Generative models are also prone to producing unsafe or explicit content, including violent or pornographic imagery [35, 55], and safety filters designed to prevent such outputs can often be bypassed [51]. Furthermore, personalization techniques can exacerbate these issues by enabling malicious use cases such as nonconsensual deepfakes and imitation of artistic styles without consent [54, 56].

Concept Erasure. These multifaceted safety challenges have spurred extensive research on concept erasure techniques aimed at mitigating harmful generative behaviors [14, 23, 24, 37, 38, 61]. Concept erasure seeks to suppress a model's ability to reproduce undesired objects, styles, or identities while maintaining generation quality for non-targeted concepts. Representative approaches include fine-tuning methods [23, 30], cross-attention editing [24, 34, 45], and attention re-steering [70]. Other strategies involve regularization [74], pruning [13], adversarial training [8], Dumo [26], and trajectory-based techniques [12]. Overall, these methods seek to alter the behavior of pretrained models through post-training modifications, particularly efficient approaches [33, 68, 69]. Recent efforts also explore interpretability-driven erasure using sparse autoencoders [17, 36], training-free localized erasure via low-rank adaptation [40], and neuron-level precision removal [29]. There is also growing interest in robustness and evaluation [15, 43, 63, 73]. In addition to these diffusion-based models, extensions on concept erasure have also been proposed for flow-matching models [25], autoregressive transformers [27], and text-to-video generation [64, 66]. Related ideas are explored in large language models through nullspace-based editing [20], which focuses on MLP layers, whereas our work targets attention and embedding layers for visual generative models.

3. Concept Erasure: Problem Formulation and Geometric Insights

3.1. Problem Formulation

Modern generative models, such as diffusion [32] and transformer-based architectures [19], implicitly encode a rich set of semantic concepts within their latent representations. Let $f_{\theta_0} : \mathbb{R}^n \to \mathbb{R}^p$ denote the pretrained model parameterized by θ_0, which maps an internal latent code $z \in \mathbb{R}^n$ to an output feature $f_{\theta_0}(z) \in \mathbb{R}^p$. Then given a text prompt c, the model defines a conditional distribution

$$p_{\theta_0}(x \mid c), \quad x \in \mathcal{X},\ c \in \mathcal{C},$$

representing the likelihood of generating an image x conditioned on the prompt c.

Let $\mathcal{C}_{\text{target}} \subseteq \mathcal{C}$ denote the set of prompts corresponding to target concepts (e.g., specific objects, styles, or identities) to be erased, and let $\mathcal{X}_{\text{target}} \subseteq \mathcal{X}$ denote the associated undesired outputs. The goal of concept erasure is to transform the model parameters from θ_0 to θ such that the target concepts are effectively suppressed, while preserving the model's ability to generate and represent a set of non-target concepts. Formally, we seek a transformation $\{\theta_0 \to \theta\}$ satisfying two complementary objectives: *erasure* and *preservation*.

Erasure objective. The primary requirement is that the modified model should not produce undesired content when conditioned on any target prompt. This can be expressed as

$$\forall\, c \in \mathcal{C}_{\text{target}} : \quad \operatorname{supp}\big(p_\theta(x \mid c)\big) \cap \mathcal{X}_{\text{target}} = \emptyset, \tag{1}$$

where $\operatorname{supp}(p_\theta(x \mid c))$ denotes the *support* of the conditional distribution [21], i.e., the set of all possible samples x that the model can produce with nonzero probability under prompt c. In practice, this strict condition is relaxed to a probabilistic form:

$$\forall\, c \in \mathcal{C}_{\text{target}} : \quad \mathbb{P}_{x \sim p_\theta(\cdot|c)}[x \in \mathcal{X}_{\text{target}}] \le \delta, \tag{2}$$

where $\delta \ge 0$ sets a tolerance for the probability of undesired content generations.

Preservation objective. Equally important is preserving the model's capabilities on non-target prompts. Ideally, the modified model should exhibit identical behavior to the original model on all non-target concepts. Let $\mathcal{C}_{\text{pres}} \subseteq \mathcal{C}$ denote the set of prompts to be preserved, and let p_{θ_0} denote the pretrained model's distribution. A natural formulation is to require that the generated distributions remain close under some divergence measure $D(\cdot\|\cdot)$:

$$\forall\, c \in \mathcal{C}_{\text{pres}} : \quad D\left(p_\theta(\cdot \mid c) \,\|\, p_{\theta_0}(\cdot \mid c)\right) \le \varepsilon, \tag{3}$$

where $\varepsilon \ge 0$ controls the tolerance for deviation. Preservation can be expressed through alignment in feature space using a functional $\phi(\cdot)$ (e.g., CLIP embeddings [49], perceptual features, or aesthetic scores):

$$\forall c \in \mathcal{C}_{\text{pres}} : \ \|\mathbb{E}_{x \sim p_\theta(\cdot|c)}[\phi(x)] - \mathbb{E}_{x \sim p_{\theta_0}(\cdot|c)}[\phi(x)]\|_2 \le \varepsilon_\phi. \tag{4}$$

Empirical measures of this divergence include Maximum Mean Discrepancy (MMD), Fréchet Inception Distance (FID) [31], or performance-based metrics such as classification accuracy.

3.2. Empirical Concept Erasure with UCE

Directly optimizing the objectives in Eqs. (2), (3), and (4) within the full parameter space of a generative model is typically infeasible, owing to the distributed nature of concept representations, the high dimensionality of model parameters, and the nonlinear behavior of modern architectures. For practical deployment, empirical approaches operate within a restricted subspace, often targeting specific projection layers or attention matrices that encode concept-level information.

A representative example is Unified Concept Editing (UCE) [24], which applies modifications to selected model components (e.g., the Key and Value matrices in attention layers) and formulates the empirical objective as:

$$\min_W \left(\sum_{c_i \in \mathcal{C}_{\text{target}}} \|W c_i - W_0 c_i^*\|_2^2 + \sum_{c_j \in \mathcal{C}_{\text{pres}}} \|W c_j - W_0 c_j\|_2^2 \right), \tag{5}$$

where c_i^* denotes a proxy representation of the erased concept, typically chosen as a neutral anchor. The empirical objective in (5) consists of two complementary parts: the first term enforces the erasure of target concepts by aligning their transformed representations $W c_i$ with neutral proxy $W_0 c_i^*$, while the second term preserves non-target concepts by constraining the new mapping W to remain close to W_0 on preserved prompts.

A notable advantage of UCE is that it admits a *closed-form solution*, allowing direct computation of the optimal projection matrix without iterative training:

$$W = \Big(\sum_{c_i \in \mathcal{C}_{\text{target}}} v_i^* c_i^\top + \sum_{c_j \in \mathcal{C}_{\text{pres}}} W_0 c_j c_j^\top \Big) \Big(\sum_{c_i \in \mathcal{C}_{\text{target}}} c_i c_i^\top + \sum_{c_j \in \mathcal{C}_{\text{pres}}} c_j c_j^\top \Big)^{-1}, \tag{6}$$

where $v_i^* = W_0 c_i^*$ is the desired target vector.

This closed-form solution offers several practical advantages. First, it enables one-step computation of the updated projection matrix, avoiding iterative gradient-based optimization or retraining, which substantially reduces computational overhead. Moreover, by operating solely on the concept embeddings $\mathcal{C}_{\text{target}}$ and $\mathcal{C}_{\text{pres}}$, the method is entirely data-independent and does not require additional image sampling or backpropagation through the generative model.

3.3. Closed-Form Solution ≠ Good Solution

While closed-form approaches such as UCE offer clear advantages in efficiency, they do not inherently guarantee the preservation of non-target concepts. Despite the inclusion of a preservation term in Eq. (5), violations on preserved prompts can still occur—particularly when the target and preserved concepts are correlated or non-orthogonal in the latent space. The following geometric insights provide an intuitive understanding of why violations of preserved concepts may still occur.

Geometric Insights

In least-squares regression, the best-fit line minimizes total error but does not necessarily pass through every data point. Concept erasure in Eq. (5) behaves similarly: each concept embedding is a point in a high-dimensional space, and the optimization only finds a transformation W that minimizes the global loss. As such, target and preserved concepts may not lie on the fitted line, or even deviate significantly from the fitted solution, causing degradation and distortions in these concepts.

The following theorem provides a more concrete analysis of this phenomenon.

Theorem 3.1 (Perturbation of Preserved Concepts)**.** *Assume there is only one target vector c to be edited to v^* and let C_{pres} denote the concatenated preservation matrix. Let $N = cc^\top + C_{\text{pres}}C_{\text{pres}}^\top$, and assume that for some preserve vector p, $\langle N^{-1/2}c, N^{-1/2}p\rangle \geq \lambda\langle N^{-1/2}c, N^{-1/2}c\rangle$ for some $\lambda > 0$. Then we have*

$$\|\Delta W p\|_2 \geq \lambda \|\Delta W c\|_2. \tag{7}$$

That is, the perturbation on the non-target vector p is at least λ times the perturbation on the target vector c.

In practice, we also observe this phenomenon consistently across different models and concept sets. Non-target concepts experience noticeable degradation for both object and style erasure in Table 1 and 2. These observations underscore a crucial limitation: achieving a mathematically optimal solution under a least-squares objective does not imply controlled erasure and preservation. Maintaining their performance instead requires a more deliberate geometric design.

4. Concept Erasure with Double Projections

The geometric limitations of existing closed-form approaches motivate a more principled formulation of concept erasure. To this end, we propose "Concept Erasure with Double Projections" (DP), which explicitly decouples erasure and preservation through two sequential projections. By disentangling subspace interactions, DP provides analytical guarantees for training-free updates while retaining the efficiency and interpretability of a closed-form solution.

4.1. Formulation

We aim to identify an updated transformation $W \in \mathbb{R}^{p\times n}$ that effectively removes the representations of specific target concepts while preserving those of non-target concepts. Formally, we write $W = W_0 + \Delta W$ and optimize directly over ΔW:

$$\min_{W\in\mathbb{R}^{p\times n},\, c_i^*\in\mathcal{S}} \left(\|Wc_i - W_0 c_i^*\|_2^2 + \|WC_{\text{pres}} - W_0 C_{\text{pres}}\|_F^2\right), \tag{8}$$

where $W_0 \in \mathbb{R}^{p\times n}$ is the pretrained parameter matrix (e.g., an attention Key or Value matrix), $c_i \in \mathbb{R}^n$ is the embedding of a target concept to be erased, $C_{\text{pres}} = [c_1, c_2, \ldots, c_m] \in \mathbb{R}^{n\times m}$ collects the embeddings of preserved (non-target) concepts, and $\mathcal{S}$ defines the safe subspace within which the proxy vectors c_i^* are constrained to lie.

In essence, this optimization problem (8) involves two sets of variables: the weight matrix W and the proxy vector c_i^*. A common approach to solving such problems is through "alternating optimization" [4, 7, 41], which iteratively updates one variable while keeping the other fixed until convergence. Yet, these iterative procedures typically rely on gradient-based training and can be *computationally expensive*. Instead, we introduce a novel double projection method that yields a *closed-form, training-free* approximation to this optimization problem.

4.2. Projection 1: Proxy Construction in the Safe Subspace

We begin by computing a proxy vector c_i^* that captures the component of the target concept c_i lying within the safe subspace $\mathcal{S}$. Let $S \in \mathbb{R}^{n\times k}$ denote the matrix whose columns form a (possibly non-orthogonal) basis of $\mathcal{S}$. The proxy is then obtained through an orthogonal projection:

$$c_i^* = \text{proj}_{\mathcal{S}}(c_i) = S(S^\top S)^+ S^\top c_i. \tag{9}$$

where $(S^\top S)^+$ denotes the general Moore–Penrose pseudoinverse [47], ensuring the projection remains valid even if the basis vectors are linearly dependent. This step extracts the *safe component* of the target concept within the non-target subspace, effectively filtering out directions that could interfere with preserved concepts. In practice, one can construct a safe region by using multiple safe concepts, $\mathcal{S} = \text{span}\{s_1, s_2, \ldots, s_k\}$. Note that when $\mathcal{S}$ is defined using a single concept vector ($k = 1$), we require Eq. (9) to collapse to the UCE case [24].

4.3. Projection 2: Constrained Optimization for W

Given the proxy c_i^* from Projection 1, we now optimize the transformation ΔW while guaranteeing that updates are *or-*

thogonal to the space of preserved concepts. Let the preserved (non-target) concept embeddings be collected as

$$C_{\text{pres}} = [\, c_1^{\text{pres}},\, c_2^{\text{pres}},\, \ldots,\, c_m^{\text{pres}} \,] \in \mathbb{R}^{n\times m},$$

whose column space defines the subspace that must remain invariant during erasure. We parametrize the updated transformation as

$$W \;=\; W_0 + \Delta W, \qquad \text{s.t.} \quad \Delta W\, C_{\text{pres}} = 0,$$

so that any change lies in the left nullspace of C_{pres} and therefore leaves the preserved concepts untouched.

Substituting this into Eq. (8) reduces the problem to

$$\min_{\Delta W} \; \big\|(W_0+\Delta W)\, c_i - W_0 c_i^*\big\|_2^2 \quad \text{s.t. } \Delta W\, C_{\text{pres}} = 0, \;\; (10)$$

a linearly constrained least-squares problem in ΔW.

Let $U_2 \in \mathbb{R}^{n\times(n-r)}$ be an orthonormal basis for the left nullspace of C_{pres} ($r = \text{rank}(C_{\text{pres}})$), so that $U_2^\top C_{\text{pres}} = 0$. Any feasible update can be written as $\Delta W = Z U_2^\top$ with an unknown parameter $Z \in \mathbb{R}^{p\times(n-r)}$. Define

$$x \;=\; U_2^\top c_i \in \mathbb{R}^{n-r}, \quad b \;=\; W_0(c_i^* - c_i) \in \mathbb{R}^p.$$

Eq. (10) becomes a standard least-squares problem,

$$\min_{Z} \; \big\|Zx - b\big\|_2^2,$$

whose minimum-norm solution (when $x \neq 0$) is

$$Z^\star \;=\; b\, x^\top \big(xx^\top\big)^+ \;=\; \frac{b\, x^\top}{\|x\|_2^2}. \qquad (11)$$

The update is therefore admitting a *closed-form solution*:

$$\Delta W^\star \;=\; Z^\star U_2^\top \;=\; \frac{W_0(c_i^* - c_i)\, x^\top}{\|x\|_2^2}\, U_2^\top. \qquad (12)$$

Note for multiple-concept erasure $C_{\text{tgt}} = [c_1, \ldots, c_T]$, we can solve (11) for concept matrix X in a similar way.

4.4. Discussions and Geometric Insights

The first projection is *optional*, and one could directly specify a proxy c_i^* as in UCE [24] for simplicity, effectively bypassing this projection. However, constructing a richer safe subspace $\mathcal{S}$ generally reduces the magnitude of the update, leading to smaller $\|\Delta W\|_F^2$ and thus less perturbations to the original model. In contrast, the second projection is *essential*: constraining ΔW to the nullspace of C_{pres} guarantees orthogonality, ensuring that model modifications minimally affect the preserved representations. This geometric intuition is made precise in the following theorem.

Theorem 4.1 (Preservation of Non-Target Concepts)**.** *Let $C_{\text{pres}} \in \mathbb{R}^{n\times m}$ denote the matrix of non-target concept embeddings, and let $W_0 \in \mathbb{R}^{p\times n}$ be the pretrained transformation. If the update $\Delta W \in \mathbb{R}^{p\times n}$ satisfies $\Delta W\, C_{\text{pres}} = 0$, then for $W^\star := W_0 + \Delta W$ it holds that $W^\star v = W_0 v$ for all $v \in \text{col}(C_{\text{pres}})$; that is, all non-target concept representations are exactly preserved.*

Most importantly, both projections in Eqs. (9) and (12) admit exact *closed-form solutions*. Each step, from computing the proxy vector c_i^* to updating the transformation W, can be derived analytically without any iterative optimization or gradient-based training. This makes the entire procedure fully *deterministic and training-free*, combining computational efficiency with clear geometric interpretability. In practice, c_i^* and C_{pres} are shared by all layers. Moreover, these closed-form updates enable DP to be performed within seconds, in contrast to optimization-based approaches that often require minutes or hours.

Meanwhile, it is not necessary to include all available concepts as preservation targets. Studies from AGE [8] indicate that concept erasure exhibits a largely *localized* effect: removing one concept mainly affects a small neighborhood of semantically related concepts, which can be identified through a concept graph. Hence, C_{pres} can be constructed from a compact, semantically relevant subset. Moreover, when C_{pres} includes many concepts, a low-rank truncation can be applied via its singular value decomposition, $C_{\text{pres}} = U_1 \Sigma V^\top$, where $U_1 = [u_1, \ldots, u_r] \in \mathbb{R}^{n\times r}$ contains left singular vectors ordered by singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Retaining only the top-k components, $U_{1,k} = [u_1, \ldots, u_k]$, captures the dominant subspace while discarding low-energy, redundant directions. The update rule in Eq. (10) can then be parameterized as $\Delta W = Z\, U_{2,k}^\top$, admitting a similar closed-form solution. The following theorems provide the lower bound for the erasing targets and the upper bound for non-target concepts.

Theorem 4.2 (Preservation Bound for Truncated Cases)**.** *For any preserve vector $p_i \in \mathcal{C}_{\text{pres}}$, we have*

$$\|(W' - W_0)p_i\|_2 \leq \|Z^*\|_2 \sigma_{k+1}(C_{\text{pres}}), \qquad (13)$$

where σ_{k+1} denotes the $k+1$-th singular value of C_{pres}.

Note that when $k = r$ (i.e., without truncation), the right-hand side of (13) vanishes, degenerating to Theorem 4.1. Moreover, let C_{tgt}^* be the anchor concepts and $C_\perp^{(k)} := U_{2,k}^\top C_{\text{tgt}}$. Define $B := W_0\big(C_{\text{tgt}}^\star - C_{\text{tgt}}\big)$ (See Appendix A.1 for detailed explanations on notations), and we have the following erasure bound.

Theorem 4.3 (Erasure Bound for Truncated Cases)**.** *Let the thin SVD of $C_\perp^{(k)}$ be $C_\perp^{(k)} = U_{q_k}^{(k)} \Sigma_{q_k}^{(k)} V_{q_k}^{(k)\top}$ with rank $q_k \in \{0, \ldots, \min(n-k, T)\}$. We have*

$$(W - W_0)C_{\text{tgt}} = B\, V_{q_k}^{(k)} V_{q_k}^{(k)\top}. \qquad (14)$$

Moreover, assume that $\ker B$ and the row space of $C_\perp^{(k)}$ intersect trivially, i.e., $\ker B \cap \text{row}(C_\perp^{(k)}) = \{0\}$. Equivalently, $BV_{q_k}^{(k)}$ has full column rank and $\sigma_{\min}(BV_{q_k}^{(k)}) > 0$. Then, for each target column c_i, letting $y_i := V_{q_k}^{(k)\top} e_i$,

$$\|(W - W_0)c_i\|_2 \;\geq\; \sigma_{\min}\big(BV_{q_k}^{(k)}\big)\, \|y_i\|_2. \qquad (15)$$

Table 1. Results on SD 1.4 for all algorithms. Each block includes both original and post-update accuracies. Left: **Target Class** shows erasure performance (**Erased Accuracy** ↓). Right: **Other Classes** reports the accuracy of preserved concepts (**Preservation Drop** ↓). Lower values indicate stronger erasure and better preservation. *Note:* UCE and DP are **deterministic methods** and thus no standard deviation is reported. †Detailed analysis of why UCE performs worse in these two cases are provided in the Appendix D.

Object	Target Class	Erased Accuracy (%) ↓					Other Classes	Preservation Drop (%) ↓				
	Original	ESD	CP	AGE	UCE	DP	Original	ESD	CP	AGE	UCE	DP
Cassette Player	78.0	$20.5_{\pm 1.5}$	$4.3_{\pm 0.3}$	$24.0_{\pm 1.0}$	$12.0^{\dagger}$	**2.0**	86.8	$22.6_{\pm 3.1}$	$31.7_{\pm 3.5}$	$5.8_{\pm 0.4}$	20.8	**3.3**
Chain Saw	80.0	**0.0**$_{\pm 0.0}$	$1.0_{\pm 0.0}$	$1.0_{\pm 0.0}$	**0.0**	**0.0**	86.6	$22.4_{\pm 1.7}$	$34.1_{\pm 3.2}$	$6.7_{\pm 0.5}$	0.6	**0.3**
Church	84.0	**4.0**$_{\pm 1.0}$	$24.6_{\pm 0.6}$	$16.3_{\pm 1.7}$	**4.0**	**4.0**	86.1	$25.3_{\pm 2.1}$	$17.4_{\pm 1.0}$	$7.2_{\pm 0.4}$	15.1	**6.1**
Gas Pump	77.0	$8.3_{\pm 1.0}$	$4.0_{\pm 0.7}$	$4.0_{\pm 1.0}$	6.0	**2.0**	86.9	$14.1_{\pm 1.7}$	$31.8_{\pm 2.1}$	$3.9_{\pm 0.3}$	5.6	**2.6**
Tench	73.0	$3.0_{\pm 1.0}$	**0.0**$_{\pm 0.0}$	$12.0_{\pm 2.0}$	**0.0**	**0.0**	87.3	$12.1_{\pm 1.1}$	$34.3_{\pm 3.1}$	$6.4_{\pm 0.3}$	9.9	**4.6**
Garbage Truck	78.0	$20.5_{\pm 3.0}$	**0.0**$_{\pm 0.0}$	$21.0_{\pm 2.7}$	**0.0**	**0.0**	86.8	$11.3_{\pm 0.9}$	$37.2_{\pm 4.2}$	$4.6_{\pm 0.3}$	0.1	**-1.2**
English Springer	95.0	$4.7_{\pm 0.7}$	**0.0**$_{\pm 0.0}$	**0.0**$_{\pm 0.0}$	**0.0**	**0.0**	84.9	$20.7_{\pm 1.2}$	$30.9_{\pm 3.3}$	$5.3_{\pm 0.5}$	0.0	**-0.8**
Golf Ball	99.0	$6.3_{\pm 1.0}$	$21.3_{\pm 1.3}$	$6.0_{\pm 1.0}$	$56.0^{\dagger}$	**0.0**	84.4	$28.2_{\pm 0.0}$	$32.5_{\pm 0.0}$	$5.6_{\pm 0.0}$	6.2	**5.4**
Parachute	95.0	$4.0_{\pm 0.3}$	$1.0_{\pm 0.0}$	$12.0_{\pm 1.3}$	**0.0**	**0.0**	84.9	$17.4_{\pm 2.1}$	$39.9_{\pm 4.1}$	$2.6_{\pm 0.3}$	4.0	**-0.8**
French Horn	100.0	$1.0_{\pm 0.0}$	$1.0_{\pm 0.0}$	**0.0**$_{\pm 0.0}$	**0.0**	**0.0**	84.3	$20.4_{\pm 2.1}$	$34.7_{\pm 3.1}$	$7.9_{\pm 0.6}$	4.7	**4.2**
Mean	85.9	7.2	5.7	9.6	7.8	**0.8**	85.9	19.5	32.5	5.6	6.7	**2.4**

5. Experiments

We now turn to the empirical evaluation of concept erasure and preservation, examining the proposed approach under various scenarios such as object and style removal across different Stable Diffusion variants and modern flow-matching models.

5.1. Experimental Setup

Backbones and Tasks Our experiments are first conducted on Stable Diffusion v1.4 (SD1.4), the most widely used backbone in prior concept-erasure studies. To assess generality across architectures, we further evaluate our method on Stable Diffusion v1.5 (SD1.5) and the recent flow-matching generative model FLUX [3, 42]. Following prior work [23, 24], we consider two standard evaluation tracks: (i) object-level erasure on ten ImageNet categories including *cassette player*, *chain saw*, *church*, *gas pump*, *tench*, *garbage truck*, *English springer*, *golf ball*, *parachute*, and *French horn* and (ii) style-level erasure targeting five artistic concepts including *Pablo Picasso*, *Vincent van Gogh*, *Rembrandt*, *Andy Warhol*, and *Caravaggio*. For object-level erasure, we report the Top-1 classification accuracy of a pretrained ResNet-50 [28] on generated images. For style-level erasure, we measure the CLIP [49] text–image similarity between generated samples and the corresponding style prompts as [8, 23, 24].

Erasure Methods We benchmark representative concept-erasure methods that collectively span projection-, fine-tuning-, adversarial-, and pruning-based paradigms. Specifically, we compare against Unified Concept Editing (UCE) [24], Erased Stable Diffusion (ESD) [23], Concept-Prune (CP) [13] and AGE [8]. These baselines cover a diverse methodological spectrum, enabling a comprehensive assessment of DP's effectiveness and efficiency relative to existing approaches. For each method, ten image variants are generated per prompt.

5.2. Object Erasure with Stable Diffusion

We first focus on object-level concept erasure using Stable Diffusion v1.4 (SD 1.4), a canonical benchmark backbone for prior erasure studies [8, 23, 24]. We follow these prior works in selecting the same ten ImageNet object categories to ensure comparability with established erasure benchmarks. However, our evaluation protocol adopts a **stricter and more realistic criterion** than previous studies. Specifically, we unify visually and semantically similar concepts (e.g., treating "cassette player" and "tape player" as equivalent categories) to mitigate the category ambiguity in diffusion outputs. Furthermore, unlike prior evaluations [8] that relied on Top-5 accuracy, we report Top-1 accuracy throughout.

Table 1 reports the performance of all algorithms. In terms of **concept erasure**, several existing methods achieve strong suppression of the target object, confirming that diffusion backbones are generally amenable to concept-level editing. Methods like UCE, CP and DP, for instance, demonstrate effective removal on easily separable categories such as "Chain Saw" and "English Springer", where the erased accuracy drops close to zero. These results indicate that when the concept subspace is well localized, single-projection or pruning-based updates can adequately diminish target activations.

In terms of **preservation**, the proposed DP algorithm consistently achieves the smallest degradation across all objects, demonstrating a clear advantage in maintaining non-target representations. While competing approaches often introduce secondary distortions, such as performance drops exceeding 20% for ESD and CP, DP preserves nearly unchanged accuracy on the remaining nine categories, typically within only a few percentage points. This stability stems from its double-projection mechanism, particularly the nullspace projection, which explicitly constrains updates to the left nullspace of preserved representations.

Table 2. Results on SD 1.4 for **artistic style erasure**. Each block includes both original and post-update accuracies. *Note:* each artist is given a set of its own labels to compute the CLIP score. UCE and DP are **deterministic methods** and no standard deviation is reported.

Style	**Target Class**	**Erased Accuracy (%) ↓**					**Other Classes**	**Preservation Drop (%) ↓**				
	Original	ESD	CP	AGE	UCE	DP	Original	ESD	CP	AGE	UCE	DP
Andy Warhol	86.0	$14.7_{\pm 1.3}$	$21.5_{\pm 2.0}$	$31.5_{\pm 3.0}$	15.0	**12.5**	92.1	$4.9_{\pm 2.1}$	$10.4_{\pm 1.4}$	$5.2_{\pm 0.7}$	2.1	**2.0**
Caravaggio	82.0	$23.3_{\pm 1.3}$	$20.3_{\pm 1.3}$	$\mathbf{7.5}_{\pm 1.0}$	16.0	11.5	93.6	$8.8_{\pm 2.1}$	$27.5_{\pm 3.2}$	$11.4_{\pm 1.1}$	0.7	**0.4**
Pablo Picasso	81.5	$40.5_{\pm 2.0}$	$\mathbf{24.5}_{\pm 2.5}$	$26.0_{\pm 1.0}$	30.0	26.0	91.4	$14.7_{\pm 1.2}$	$13.4_{\pm 2.1}$	$4.3_{\pm 0.3}$	2.9	**2.6**
Rembrandt	85.0	$18.0_{\pm 2.0}$	$15.0_{\pm 1.0}$	$3.5_{\pm 0.5}$	4.5	**3.0**	84.4	$8.8_{\pm 0.9}$	$23.3_{\pm 2.1}$	$6.9_{\pm 0.7}$	-1.0	**-1.4**
Van Gogh	63.0	$8.5_{\pm 0.5}$	$14.7_{\pm 1.7}$	$17.7_{\pm 0.3}$	7.0	**5.5**	89.9	$6.7_{\pm 0.6}$	$8.9_{\pm 1.2}$	$5.8_{\pm 0.4}$	0.8	**-1.3**
Mean	79.5	21.0	19.2	17.2	14.5	**11.7**	90.3	8.8	16.7	6.7	1.1	**0.5**

Table 3. Results on the FLUX model for object erasure. Rows 1–3 report target erasure (**Erased Accuracy** ↓). Rows 4–6 report preservation fidelity (**Preservation Drop** ↓). Visualization of sample generated images is available in Appendix H.

Metric	Cassette Player	Chain Saw	Church	Gas Pump	Tench	Garbage Truck	English Springer	Golf Ball	Parachute	French Horn	**Mean**
					Target Class (Erased Accuracy, % ↓)						
Original	39.0	100.0	99.0	100.0	89.0	98.0	82.0	100.0	99.0	100.0	90.6
UCE	**0.0**	**0.0**	63.0	73.0	6.0	21.0	**0.0**	76.0	**0.0**	**0.0**	23.9
DP	**0.0**	**0.0**	**12.0**	**0.0**	**0.0**	**0.0**	**0.0**	**1.0**	**0.0**	**0.0**	**1.0**
					Other Classes (Preservation Drop, % ↓)						
Original	96.3	89.6	89.7	89.6	90.8	89.8	91.6	89.6	89.7	89.6	90.6
UCE	2.0	2.7	2.1	2.0	1.5	2.2	2.3	2.3	1.9	3.0	2.2
DP	**0.5**	**0.9**	**0.3**	**1.4**	**-0.1**	**0.4**	**1.2**	**-0.1**	**1.0**	**0.3**	**0.6**

Consequently, the erasure operation remains geometrically orthogonal to the non-target embeddings, ensuring that both the visual quality and semantic fidelity of unaffected generations are largely retained.

5.3. Why is perfect preservation not observed?

Beyond the observed performance improvements, it is also important to rethink why perfect preservation is not achieved in this experiment. Ideally, non-target concepts should remain entirely unaffected under DP, since the update ΔW is explicitly designed to be orthogonal to their embeddings, as established in Theorem 4.1. In practice, however, perfect preservation is not always observed. This minor deviation arises from the presence of positional embeddings in diffusion models: although DP enforces $\Delta W c_j = 0$ to preserve non-target content embeddings, the model operates on representations of the form $z_j = c_j + q_j$, where q_j denotes the positional embedding. This additive coupling, which is also present in other closed-form methods such as UCE, introduces small but consistent deviations from perfect preservation, as confirmed empirically (see Appendix C for details). Moreover, this issue is further compounded by the self-attention mechanism in the encoder, which introduces additional interactions across token representations.

In the following FLUX example, we demonstrate that this fluctuation can be mitigated by performing concept erasure directly on the embedding layers of our encoders.

5.4. Artistic Styles Erasure with Stable Diffusion

We next evaluate the proposed DP algorithm on the task of artistic style erasure, using Stable Diffusion 1.4 as the base model. Following prior studies [24], we focus on five representative artistic styles that exhibit diverse visual characteristics and degrees of abstraction. Performance is evaluated using the CLIP text–image similarity between generated images and their corresponding style prompts.

As shown in Table 2, existing methods already demonstrate competitive performance across several artistic styles. For instance, CP achieves particularly strong removal on "Picasso", while AGE performs slightly better on "Caravaggio", indicating that localized or style-specific optimization can yield strong suppression. Nevertheless, across all styles, DP achieves comparable or better erasure quality on all five artistic styles. The main advantage of DP lies in its strong ability to preserve non-target concepts. While other methods often degrade unrelated styles due to overlapping feature directions, DP ensures that style-independent components remain relatively intact. As a result, the model retains its ability to accurately reproduce unaffected artistic styles with minimal performance drop, typically within only a few percentage points.

5.5. Switching to Flow Matching

Recent advances [42] demonstrate that flow matching offers an equally powerful and more theoretically grounded alternative. To test the generality of our erasure approach, we further evaluate DP on a recent flow-matching model, FLUX [3]. Note that ESD operates on predicted noise, whereas flow-matching models predict vector fields, making ESD incompatible with these experiments.There is also no direct support to utilize pruning or adversarial methods in flow matching. Consequently, we exclude these approaches from our comparison. Closed-form erasure methods like UCE and DP, by contrast, exhibit *broader applicability* because they directly operate on linear mappings

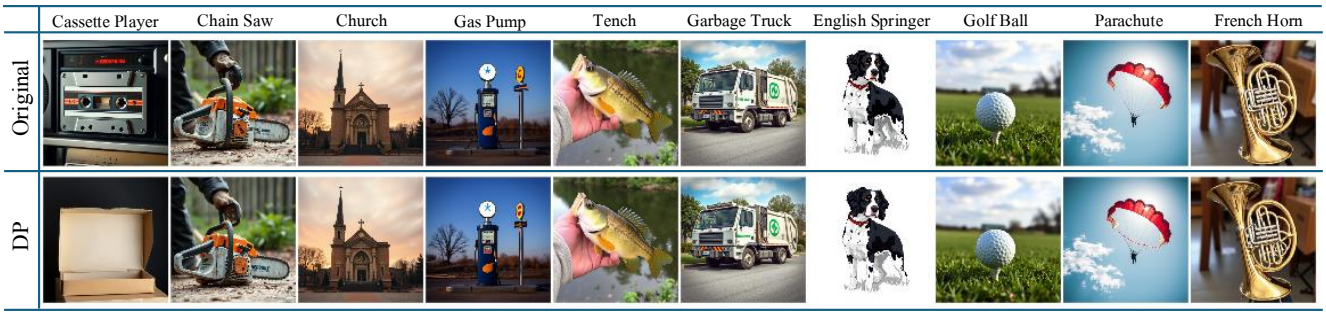

Figure 1. Visualization of concept erasure on "Cassette player". Results indicated that the target concept (column 1) is effectively suppressed, while the remaining nine categories (columns 2–10) show minimal impact.

rather than model-specific generative dynamics. Note that in the FLUX model, we apply these closed-form updates to the *embedding layers* rather than attention blocks. This formulation also naturally eliminates interference from positional embeddings, enabling a cleaner concept erasure process.

As shown in Table 3, both closed-form methods, UCE and DP, achieve effective object erasure within the flow-matching framework. UCE successfully suppresses most target concepts, demonstrating its adaptability beyond diffusion models; however, its residual accuracies on complex categories such as "Church" and "Gas Pump" suggest that direct linear projections may not fully capture the flow field's geometric structure. In contrast, DP consistently attains near-zero residual accuracies across all ten objects, confirming its ability to generalize across generative paradigms. Notably, DP also yields markedly smaller preservation drops, averaging only 0.6% compared to UCE's 23.9%, indicating that the nullspace constraint effectively isolates target directions even in a flow-based representation space. Specifically, Figure 1 presents the generated images from the original FLUX model and our proposed DP approach for the concept "Cassette Player". Results indicate that the erased concept is effectively suppressed, while the remaining nine categories are largely preserved. These results highlight that DP maintains its theoretical advantages, while also providing robust, architecture-agnostic concept erasure with minimal interference to non-target concepts on the FLUX model.

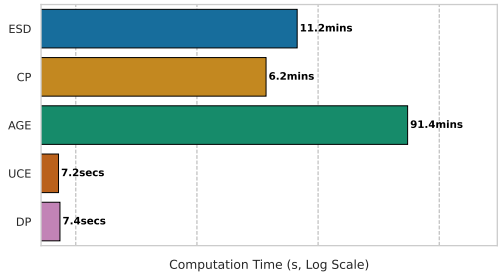

Figure 2. Computation time comparison across erasure methods (log scale). Results indicate that closed-form approaches such as UCE and DP perform concept erasure within only a few seconds. Experiments are conducted for SD 1.4 on Nvidia 3090.

5.6. Time Consumption

One key advantage of closed-form erasure methods lies in their exceptional *computational efficiency*. Iterative approaches, such as ESD, CP and AGE, require repeated optimization steps to update noise parameters or perform pruning, leading to substantial time costs on the order of several minutes or even hours, as shown in Figure 2. In contrast, closed-form formulations like UCE and DP complete the erasure process almost instantaneously, requiring only a few seconds. This is due to their underlying training-free mechanism.

5.7. Additional Experiments

Due to space constraints, we present additional experiments in the Appendix. (1) We provide quantitative measures including LPIPS [71], PSNR [6], SSIM [62], and FID [31], in Appendix E. (2) Moreover, we conduct experiments on alternative model variants, such as Stable Diffusion v1.5 in Appendix F. (3) We also present complete visualizations of generated images for both Stable Diffusion and FLUX in Appendix G and Appendix H, respectively. (4) Additionally, ablation studies are provided in Appendix I. (5) Generalization beyond C_{pres} is reported in Appendix J.

6. Conclusion

In this work, we introduced DP, a closed-form, training-free framework for principled concept erasure in generative models. By formulating the task as a pair of sequential projections, first extracting the safe component of a concept, and then constraining updates within the left nullspace of preserved representations, DP offers a deterministic solution with clear geometric interpretability and analytical guarantees. Comprehensive experiments on both Stable Diffusion and the flow-matching model demonstrate that DP achieves erasure quality comparable to or exceeding existing baselines, while consistently minimizing preservation loss. This broad applicability, coupled with a runtime of only a few seconds, highlights the practicality of DP as a drop-in tool for safe and controllable concept erasure.

Acknowledgements

This research is partially supported by the National Research Foundation, Singapore, under the NRF fellowship (project No.NRF-NRFF13-2021-0005).

References

[1] Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023. 2

[2] Sarah Andersen, Kelly McKernan, and Karla Ortiz. Sarah Andersen, Kelly McKernan, and Karla Ortiz et al. v. Stability AI Ltd. et al., 2023. Case No. 3:2023cv00201, US District Court for the Northern District of California. 2

[3] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 2, 6, 7

[4] James C. Bezdek and Richard J. Hathaway. Convergence of alternating optimization. *Neural, Parallel & Scientific Computations*, 11(4):351–368, 2003. 4

[5] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1493–1504, 2023. 2

[6] Alan C. Bovik. *Handbook of Image and Video Processing*. Academic Press, 2000. 8, 19

[7] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3 (1):1–122, 2011. 4

[8] Anh Bui, Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. 2025. 1, 2, 5, 6, 16

[9] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Jacob Steinhardt. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, 2021. 1

[10] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *ICLR*, 2022. 1

[11] Nicholas 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)*, 2023. 2

[12] Finn Carter. Trace: Trajectory-constrained concept erasure in diffusion models. *arXiv preprint arXiv:2505.23312*, 2025. 2

[13] Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. In *ICLR*, 2025. 1, 2, 6

[14] Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. Machine unlearning under retain–forget entanglement. In *ICLR*, 2026. 2

[15] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023. 2

[16] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023. 2

[17] Bartosz Cywiński and Kamil Deja. SAeUron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *arXiv preprint arXiv:2501.18052*, 2025. 2

[18] Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. LoRAShop: Training-free multi-concept image generation and editing with rectified flow transformers. *arXiv preprint arXiv:2505.23758*, 2025. 2

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[20] Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained model editing for language models. In *ICLR*, 2025. 2

[21] Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999. 3

[22] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2022. 2

[23] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, 2023. 1, 2, 6

[24] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *WACV*, 2024. 1, 2, 3, 4, 5, 6, 7

[25] Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *ICML*, 2025. 2

[26] Feng Han, Kai Chen, Chao Gong, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Dumo: Dual encoder modulation network for precise concept erasure. In *AAAI*, 2025. 2

[27] Feng Han, Chao Gong, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Vce: Safe autoregressive image generation via visual contrast exploitation. *arXiv preprint arXiv:2509.16986*, 2025. 2

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[29] Qinqin He, Jiaqi Weng, Jialing Tao, and Hui Xue. A single neuron works: Precise concept erasure in text-to-image diffusion models. *arXiv preprint arXiv:2509.21008*, 2025. 2

[30] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *NeurIPS*, 2023. 2

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 3, 8, 19

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3

[33] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[34] 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 *ECCV*, 2024. 1, 2

[35] Tatum Hunter. AI porn is easy to make now. For women, that's a nightmare. The Washington Post, 2023. 2

[36] Dahye Kim and Deepti Ghadiyaram. Concept steerers: Leveraging k-sparse autoencoders for controllable generations. *arXiv preprint arXiv:2501.19066*, 2025. 2

[37] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*, 2023. 2

[38] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, 2023. 2

[39] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2

[40] Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *CVPR*, 2025. 2

[41] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *NeurIPS*, 2000. 4

[42] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1, 2, 6, 7

[43] Ping Liu and Chi Zhang. Erased or dormant? rethinking concept erasure through reversibility. *arXiv preprint arXiv:2505.16174*, 2025. 2

[44] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 2

[45] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *CVPR*, 2024. 1, 2

[46] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 2

[47] Eliakim H Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the american mathematical society*, 26: 294–295, 1920. 4

[48] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 6

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[51] Javier Rando, Daniel Paleka, David Lindner, Lennard Heim, and Florian Tramèr. Red-teaming the Stable Diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 2

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2

[53] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[54] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023. 2

[55] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, 2023. 2

[56] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*, 2023. 2

[57] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 2

[58] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *CVPR*, 2023. 2

[59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[60] Zhicheng Sun, Zhenhao Yang, Yang Jin, Haozhe Chi, Kun Xu, Liwei Chen, Hao Jiang, Yang Song, Kun Gai, and Yadong Mu. Rectifid: Personalizing rectified flow with anchored classifier guidance. In *NeurIPS*, 2024. 2

[61] Jiacheng Wang, Ping Liu, and Wei Xu. Unified diffusion-based rigid and non-rigid editing with text and image guidance. In *ICME*, 2024. 2

[62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8, 19

[63] Yiwei Xie, Ping Liu, and Zheng Zhang. Erasing concepts, steering generations: A comprehensive survey of concept suppression. *arXiv preprint arXiv:2505.19398*, 2025. 2

[64] Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. Videoeraser: Concept erasure in text-to-video diffusion models. In *EMNLP*, 2025. 2

[65] Tianyun Yang, Juan Cao, and Chang Xu. Pruning for robust concept erasing in diffusion models. In *NeurIPS Workshops*, 2024. 1

[66] Xiaoyu Ye, Songjie Cheng, Yongtao Wang, Yajiao Xiong, and Yishen Li. T2vunlearning: A concept erasing method for text-to-video diffusion models. *arXiv preprint arXiv:2505.17550*, 2025. 2

[67] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 1

[68] Chi Zhang, Cheng Jingpu, Yanyu Xu, and Qianxiao Li. Parameter-efficient fine-tuning with controls. In *Forty-first International Conference on Machine Learning*, 2024. 2

[69] Chi Zhang, REN Lianhai, Jingpu Cheng, and Qianxiao Li. From weight-based to state-based fine-tuning: Further memory reduction on lora with parallel control. In *Forty-second International Conference on Machine Learning*, 2025. 2

[70] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *CVPR*, pages 1755–1764, 2024. 2

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR (CVPR)*, 2018. 8, 19

[72] 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. In *NeurIPS*, 2024. 1

[73] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *ECCV*, 2024. 2

[74] Mengnan Zhao, Lihe Zhang, Tianhang Zheng, Yuqiu Kong, and Baocai Yin. Separable multi-concept erasure from diffusion models. *arXiv preprint arXiv:2402.05947*, 2024. 2