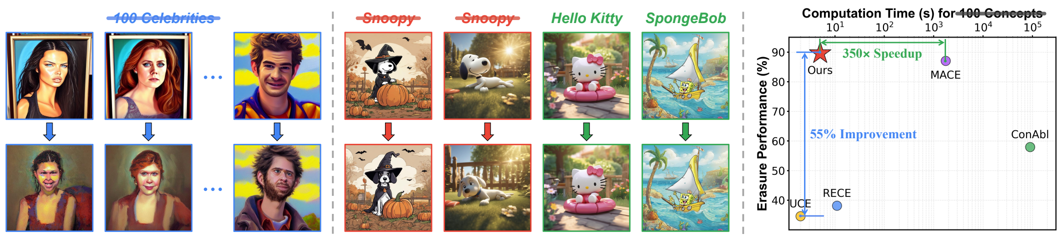
SPEED: Scalable, Precise, and Efficient Concept Erasure for Diffusion Models

Anonymous Author(s)
Affiliation
Address
email



(a) Scalable (b) Precise (c) Efficient

Figure 1: **Three characteristics of our proposed concept erasure method for diffusion models, SPEED.** **(a) Scalable:** SPEED seamlessly scales from single-concept to large-scale multi-concept erasure (*e.g.*, 100 celebrities) without additional design. **(b) Precise:** SPEED precisely removes the target concept (*e.g.*, ~~*Snoopy*~~) while preserving the semantics for non-target concepts (*e.g.*, *Hello Kitty* and *SpongeBob*). **(c) Efficient:** SPEED immediately erases 100 concepts within 5 seconds, achieving new state-of-the-art (SOTA) performance with a 350× speedup over competitive methods.

Abstract

Erasing concepts from large-scale text-to-image (T2I) diffusion models has become increasingly crucial due to the growing concerns over copyright infringement, offensive content, and privacy violations. In scalable applications, fine-tuning-based methods are time-consuming to precisely erase multiple target concepts, while real-time editing-based methods often degrade the generation quality of non-target concepts due to conflicting optimization objectives. To address this dilemma, we introduce SPEED, an efficient concept erasure approach that directly edits model parameters. SPEED searches for a null space, a model editing space where parameter updates do not affect non-target concepts, to achieve scalable and precise erasure. To facilitate accurate null space optimization, we incorporate three complementary strategies: Influence-based Prior Filtering (IPF) to selectively retain the most affected non-target concepts, Directed Prior Augmentation (DPA) to enrich the filtered retain set with semantically consistent variations, and Invariant Equality Constraints (IEC) to preserve key invariants during the T2I generation process. Extensive evaluations across multiple concept erasure tasks demonstrate that SPEED consistently outperforms existing methods in non-target preservation while achieving efficient and high-fidelity concept erasure, successfully erasing 100 concepts within just 5 seconds.

1 Introduction

Large-scale text-to-image (T2I) diffusion models [23, 54, 55, 37, 47, 24] have facilitated significant breakthroughs in generating highly realistic and contextually consistent images simply from textual descriptions [11, 44, 16, 7, 48, 42, 12]. Alongside these advancements, concerns have also been raised regarding copyright violations [10, 52], offensive content [49, 64, 66], and privacy concerns [8, 63]. To mitigate ethical and legal risks in generation, it is often necessary to prevent the model from generating certain concepts, a process termed ***concept erasure*** [29, 17, 65]. However, removing target concepts without carefully preserving the semantics of non-target concepts can introduce unintended artifacts, distortions, and degraded image quality [17, 40, 49, 65], compromising the model's reliability and usability. Therefore, beyond ensuring the effective removal of target concepts (*i.e.*, ***erasure efficacy***), concept erasure should also maintain the original semantics of non-target concepts (*i.e.*, ***prior preservation*** [61]).

In this context, recent methods strive to seek a balance between erasure efficacy and prior preservation, broadly categorized into two paradigms: training-based [29, 35, 33] and editing-based [18, 19]. The training-based paradigm fine-tunes T2I diffusion models to achieve concept erasure, incorporating an additional regularization term into the training objective for prior preservation. In contrast, the editing-based paradigm avoids additional fine-tuning by directly modifying model parameters (*e.g.*, projection weights in cross-attention layers [47]), with such modifications derived from a closed-form objective that jointly accounts for erasure and preservation. This efficiency also facilitates editing-based methods to extend to multi-concept erasure without additional designs seamlessly.

However, as the number of target concepts increases, current editing-based methods [18, 19] struggle to balance between erasure efficacy and prior preservation. This can be attributed to the growing conflicts between erasure and preservation objectives, making such trade-offs increasingly difficult. Moreover, these methods rely on weighted least squares optimization, inherently imposing a ***non-zero lower bound*** on preservation error (see Appx. B.2). In multi-concept settings, this accumulation of preservation errors gradually distorts non-target knowledge, thereby degrading prior preservation. To address the above limitations, we propose **S**calable, **P**recise, and **E**fficient Concept **E**rasure for **D**iffusion Models (SPEED) (see Fig. 1), an editing-based method incorporating null-space constraints. Specifically, we search for the ***null space of prior knowledge***, a model editing space where parameter updates do not affect the feature representations of non-target concepts. By projecting the model parameter updates for concept erasure onto such null space, SPEED can minimize the preservation error to zero without compromising erasure efficacy, thereby enabling scalable and precise concept erasure without affecting non-target concepts.

The key contribution of SPEED lies in constructing an effective null space from a set of non-target concepts (*i.e.*, ***retain set***). We observe that the existing baseline with null-space constraints [14] confronts a fundamental dilemma during concept erasure: While a small retain set limits the coverage of prior knowledge, enlarging the retain set makes it increasingly difficult to identify an accurate null space. This difficulty arises because a large retain set causes the corresponding feature matrix to approach full rank, necessitating the estimation of its null space to ensure sufficient degrees of freedom for optimization (*i.e.*, for concept erasure). However, this estimation inevitably introduces semantic degradation within the retain set and deteriorating prior preservation (see Fig. 2 and Eq. 4).

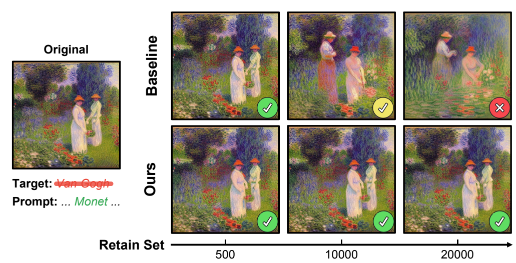


Figure 2: **Semantic degradation with increasing non-target concepts in the retain set.** Baseline null-space constrained method [14] can maintain the non-target semantics given a small retain set (). However, as the retain set grows, the rank of corresponding matrix increases, making null space estimation increasingly inaccurate (see Eq. 4) with inevitable approximation errors, thereby degrading *Monet*'s semantics in the retain set (and).

In this light, we introduce Prior Knowledge Refinement, a suite of techniques that strategically and selectively refine the retain set to mitigate the semantic degradation in searching for the null space. Particularly, we propose Influence-based Prior Filtering (IPF), which first quantifies the influence of concept erasure on each non-target concept. It then prunes the retain set by removing minimally affected concepts, preventing the correlation matrix from approaching full rank and thus maintaining an accurate null space. Subsequently, to further enhance prior preservation over the resulting retain set, we propose Directed Prior Augmentation (DPA), which expands the retain set with directed, semantically consistent perturbations to improve retain coverage. In addition, we incorporate Invariant Equality Constraints (IEC) to preserve specific representations, such as the `[SOT]` token, that should remain unchanged during editing. IEC enforces equality constraints on such invariants to regularize the retaining of essential generation properties. We evaluate SPEED on three representative concept erasure tasks, *i.e.*, few-concept, multi-concept, and implicit concept erasure, where it consistently exhibits superior prior preservation across all erasure tasks. Overall, our contributions can be summarized as follows:

- We propose SPEED, a scalable, precise, and efficient concept erasure method with null-space constrained model editing, capable of erasing 100 concepts in 5 seconds.
- We introduce Prior Knowledge Refinement to construct an accurate null space over the retain set for effective editing. Leveraging three complementary techniques, IPF, DPA, and IEC, our method balances semantic degradation and retain coverage, enabling precise and scalable concept erasure.
- Our extensive experiments show that SPEED consistently outperforms existing methods in prior preservation across various erasure tasks with minimal computational costs.

2 Related Works

Concept erasure. Current T2I diffusion models inevitably involve unauthorized and NSFW (Not Safe For Work) generations due to the noisy training data from web [51, 50]. Apart from applying additional filters or safety checkers [45, 39, 46], prevailing methods modify diffusion model parameters to erase specific target concepts, mainly categorized into two paradigms. The training-based paradigm fine-tunes model parameters with specific erasure objectives [29, 17, 65] and additional regularization terms [29, 35, 33]. In contrast, the editing-based paradigm edits model parameters using a closed-form solution to facilitate efficiency in concept erasure. For example, UCE [18] modifies model weights by balancing both erasure and preservation error through a weighted least squares objective and RECE [19] iteratively derives new target concept embeddings. These methods can erase numerous concepts within seconds, demonstrating superior efficiency in practice.

Null-space constraints. The null space of a matrix, a fundamental concept in linear algebra, refers to the set of all vectors that the matrix maps to the zero vector. The null-space constraints are first applied to continual learning (CL) by projecting gradients onto the null space of uncentered covariances from previous tasks [58]. Subsequent studies [34, 59, 62, 28, 30] further explore and extend the application of null space in CL. In model editing, AlphaEdit [14] restricts model weight updates onto the null space of preserved knowledge, effectively mitigating trade-offs between editing and preservation. Null-space constraints also apply to various tasks, *e.g.*, machine unlearning [9], MRI reconstruction [15], and image restoration [60], offering promise for editing-based concept erasure.

3 Problem Formulation

In T2I diffusion models, each concept is encoded by a set of text tokens via CLIP [43], which are then aggregated into a single concept embedding $\boldsymbol{c} \in \mathbb{R}^{d_0}$. For concept erasure, there are two sets of concepts: the erasure set $\mathbf{E}$ and the retain set $\mathbf{R}$. The erasure set consists of N_E target concepts to be removed, denoted as $\mathbf{E} = \{\boldsymbol{c}_1^{(i)}\}_{i=1}^{N_E}$. The retain set includes N_R non-target concepts that should be preserved during editing, denoted as $\mathbf{R} = \{\boldsymbol{c}_0^{(j)}\}_{j=1}^{N_R}$. To enable efficient erasure efficacy for $\mathbf{E}$ and prior preservation for $\mathbf{R}$, we first formulate a closed-form editing objective in Sec. 3.1, and enhance it with null-space constrained optimization in Sec. 3.2.

3.1 Concept Erasure in Closed-Form Solution

To effectively erase each target concept $\boldsymbol{c}_1^{(i)} \in \mathbf{E}$ (*e.g.*, ~~*Snoopy*~~), it is specified to be mapped onto an anchor concept $\boldsymbol{c}_*^{(i)}$ that shares general semantics (*e.g.*, *Dog*), termed as an anchor set $\mathbf{A} = \{\boldsymbol{c}_*^{(i)}\}_{i=1}^{N_E}$. For editing-based methods [40, 18, 19], concept embeddings from the erasure set $\mathbf{E}$, anchor set $\mathbf{A}$, and retain set $\mathbf{R}$ are first organized into three structured matrices: $\mathbf{C}_1, \mathbf{C}_* \in \mathbb{R}^{d_0 \times N_E}$ and $\mathbf{C}_0 \in \mathbb{R}^{d_0 \times N_R}$, representing the stacked embeddings of target, anchor, and non-target concepts, respectively. To derive a closed-form solution for concept erasure, existing methods typically optimize a perturbation $\boldsymbol{\Delta}$ to model parameters $\mathbf{W}$, balancing between erasure efficacy and prior preservation. For example, UCE [18] formulates concept erasure as a weighted least squares problem:

$$\boldsymbol{\Delta}_{\text{UCE}} = \arg\min_{\boldsymbol{\Delta}} \underbrace{\|(\mathbf{W}+\boldsymbol{\Delta})\mathbf{C}_1 - \mathbf{W}\mathbf{C}_*\|^2}_{e_1} + \underbrace{\|\boldsymbol{\Delta}\mathbf{C}_0\|^2}_{e_0}, \tag{1}$$

where the erasure error e_1 ensures that each target concept is mapped onto its corresponding anchor concept and the preservation error e_0 minimizes the impact on non-target concepts. This formulation provides a closed-form solution $\boldsymbol{\Delta}_{\text{UCE}}$ (see Appx. B.1) for parameter updates, achieving computationally efficient optimization. However, as the number of target concepts increases, the accumulated preservation errors e_0, which prove to share a non-zero bound from Appx. B.2, across multiple target concepts would amplify the distortion on non-target knowledge and degrade prior preservation.

3.2 Apply Null-Space Constraints

To mitigate the limitation of weighted optimization in prior preservation, SPEED integrates null-space constraints [58, 14] to achieve prior-preserved model editing by forcing $e_0 = 0$. Specifically, the null space of $\mathbf{C}_0$ is the set of all vectors $\boldsymbol{v}$ such that $\boldsymbol{v}\mathbf{C}_0 = \mathbf{0}$. Restricting the parameter update $\boldsymbol{\Delta}$ to this space ensures that such updates do not interfere with non-target concepts.

To project $\boldsymbol{\Delta}$ onto null space, we perform singular value decomposition (SVD) on $\mathbf{C}_0\mathbf{C}_0^\top \in \mathbb{R}^{d_0 \times d_0}$[1] and have $\left\{\mathbf{U}, \boldsymbol{\Lambda}, \mathbf{U}^\top\right\} = \text{SVD}\left(\mathbf{C}_0\mathbf{C}_0^\top\right)$, where $\mathbf{U} \in \mathbb{R}^{d_0 \times d_0}$ contains the singular vectors of $\mathbf{C}_0\mathbf{C}_0^\top$, and $\boldsymbol{\Lambda}$ is a diagonal matrix of its singular values. The singular vectors in $\mathbf{U}$ w.r.t. zero singular values form an orthonormal basis for the null space of $\mathbf{C}_0$, which we denote as $\hat{\mathbf{U}}$. Using this basis, we construct the null-space projection matrix $\mathbf{P} = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top$. The process can be formulated as:

$$\left\{\mathbf{U}, \boldsymbol{\Lambda}, \mathbf{U}^\top\right\} = \text{SVD}\left(\mathbf{C}_0\mathbf{C}_0^\top\right), \quad \mathbf{U} \in \mathbb{R}^{d_0 \times d_0} \xrightarrow[\text{values}]{\text{zero singular}} \hat{\mathbf{U}} \Longrightarrow \mathbf{P} = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top. \tag{2}$$

The final update applied to model parameters is $\boldsymbol{\Delta}\mathbf{P}$, which projects $\boldsymbol{\Delta}$ onto the null space of $\mathbf{C}_0$. This ensures that updates do not interfere with non-target concepts, satisfying $\|(\boldsymbol{\Delta}\mathbf{P})\mathbf{C}_0\|^2 = 0$. To solve for the updates, we minimize the following objective:

$$\boldsymbol{\Delta}_{\text{Null}} = \arg\min_{\boldsymbol{\Delta}} \underbrace{\|(\mathbf{W} + \boldsymbol{\Delta}\mathbf{P})\mathbf{C}_1 - \mathbf{W}\mathbf{C}_*\|^2}_{e_1} + \cancel{\underbrace{\|(\boldsymbol{\Delta}\mathbf{P})\mathbf{C}_0\|^2}_{e_0=0}} + \underbrace{\|\boldsymbol{\Delta}\mathbf{P}\|^2}_{\text{regularization}}, \tag{3}$$

where $\|\boldsymbol{\Delta}\mathbf{P}\|^2$ is a regularization term to ensure convergence. The preservation term $\|(\boldsymbol{\Delta}\mathbf{P})\mathbf{C}_0\|^2$ is omitted, as it is guaranteed to be zero by the null-space constraint. This objective enables us to update the model parameters such that target concepts are effectively erased while non-target representations remain unaffected, thereby achieving prior-preserved concept erasure.

4 Prior Knowledge Refinement

However, as more diverse non-target concepts are included in the retain set, the rank of the correlation matrix $\mathbf{C}_0\mathbf{C}_0^\top$ increases[2]. The null space, defined as the orthogonal complement of this span, correspondingly shrinks in dimension:

$$\dim(\text{Null}(\mathbf{C}_0)) = d_0 - \text{rank}(\mathbf{C}_0\mathbf{C}_0^\top). \tag{4}$$

Here, the null space dimension characterizes the degrees of freedom available for editing without affecting the retained concepts. However, as this dimension shrinks, to ensure sufficient degrees of freedom for concept erasure, we are compelled to include singular vectors w.r.t. non-zero singular values in $\hat{\mathbf{U}}$ following [14], which leads to an approximate null space and induces semantic degradation within the retain set (see Fig. 2). To mitigate this problem, we propose Prior Knowledge Refinement, a structured strategy for refining the retain set to enable accurate null-space construction. It comprises three complementary techniques: Influence-Based Prior Filtering (Sec. 4.1) to discard weakly affected non-target concepts to form a viable null space; Directed Prior Augmentation (Sec. 4.2) to expand the retain set with targeted and semantically consistent variations; and Invariant Equality Constraints (Sec. 4.3) to enforce equality constraints to preserve critical invariants during generation.

[1] $\mathbf{C}_0\mathbf{C}_0^\top$ and $\mathbf{C}_0$ share the same null space. We operated on $\mathbf{C}_0\mathbf{C}_0^\top \in \mathbb{R}^{d_0 \times d_0}$ since it has fixed row dimension while $\mathbf{C}_0 \in \mathbb{R}^{d_0 \times N_R}$ may have high dimensionality depending on concept number N_R.

[2] We assume that the concepts are not exactly linearly dependent in the representation space, which is generally satisfied in practice due to the semantic diversity and high dimensionality of the embedding space.

4.1 Influence-Based Prior Filtering (IPF)

Given a pre-defined retain set, existing editing-based methods [18, 19] treat all non-target concepts equally when enforcing prior preservation. However, a critical yet overlooked fact is that parameter updates inherently induce output changes over non-target concepts, and these changes vary significantly across different non-target elements. This suggests that not all non-target concepts contribute equally to preserving the model's prior knowledge, and weakly influenced concepts would provide marginal benefit while introducing additional ranks to narrow the null space.

To this end, we propose an explicit and model-consistent metric, *i.e.*, **prior shift**, to quantify how much a non-target concept is affected by concept erasure. Specifically, we isolate the effect of erasure by solving for a closed-form update $\mathbf{\Delta}_{\text{erase}}$ that minimizes only the erasure error e_1 while discarding the preservation term e_0 from Eq. 1:

$$\mathbf{\Delta}_{\text{erase}} = \arg\min_{\mathbf{\Delta}} \underbrace{\|(\mathbf{W}+\mathbf{\Delta})\mathbf{C}_1 - \mathbf{W}\mathbf{C}_*\|^2}_{e_1} + \underbrace{\|\mathbf{\Delta}\|^2}_{\text{regularization}} = \mathbf{W}\left(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top\right)\left(\mathbf{I}+\mathbf{C}_1\mathbf{C}_1\right)^{-1}. \tag{5}$$

where $\|\mathbf{\Delta}\|^2$ is introduced for convergence. Then, for each non-target concept embedding $\boldsymbol{c}$, we define its prior shift as: $\|\mathbf{\Delta}_{\text{erase}}\boldsymbol{c}\|^2$. This value offers a faithful reflection of how parameter updates perturb a non-target concept in the feature space with closed-form computation, and can naturally generalize to assessing multi-concept erasure effects. Based on this, we filter the original retain set $\mathbf{R}$ to focus only on highly influenced concepts:

$$\mathbf{R}_f : \mathbf{R} \mapsto \{\boldsymbol{c}_0 \in \mathbf{R} \mid \|\mathbf{\Delta}_{\text{erase}}\boldsymbol{c}_0\|^2 > \mu\}, \tag{6}$$

where the mean value $\mu = \mathbb{E}_{\boldsymbol{c}_0\sim\mathbf{R}}\left[\|\mathbf{\Delta}_{\text{erase}}\boldsymbol{c}_0\|^2\right]$ serves as a filtering threshold.

4.2 Directed Prior Augmentation (DPA)

To further enhance prior preservation over the resulting retain set with improved retain coverage, an intuitive strategy is to augment the retain set by perturbing non-target embedding $\boldsymbol{c}_0$ with random noise [35]. However, this strategy would introduce meaningless embeddings that fail to generate semantically coherent images (*e.g.*, noise image), resulting in excessive preservation with increasing ranks. To search for more semantically consistent concepts, we introduce directed noise by projecting the random noise $\boldsymbol{\epsilon}$ onto the direction in which the model parameters $\mathbf{W}$ exhibit minimal variation. This operation ensures the perturbed embeddings express closer semantics to the original concept after being mapped by $\mathbf{W}$ in Fig. 3. Specifically, we first derive a projection matrix $\mathbf{P}_{\text{min}}$:

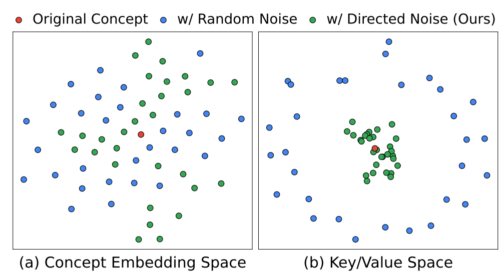


Figure 3: **t-SNE distribution of perturbing the original concept with random noise and our directed noise.** (a) Similar to random noise, our method can span a broad concept embedding space. (b) Our directed noise preserves semantic similarity to the original concept with closer distances in the space mapped by $\mathbf{W}$.

$$\left\{\mathbf{U}_{\mathbf{W}}, \mathbf{\Lambda}_{\mathbf{W}}, \mathbf{U}_{\mathbf{W}}^\top\right\} = \text{SVD}\left(\mathbf{W}\right), \quad \mathbf{P}_{\text{min}} = \mathbf{U}_{\text{min}}\mathbf{U}_{\text{min}}^\top, \tag{7}$$

where $\mathbf{U}_{\text{min}} = \mathbf{U}_{\mathbf{W}}[:, -r:]$ denotes the singular vectors w.r.t. the smallest r singular vectors[3], which represent the r least-changing directions of $\mathbf{W}$ and constrain the rank of the augmented embeddings to a maximum of r. Then the directed noise $\boldsymbol{\epsilon}\cdot\mathbf{P}_{\text{min}}$ is used to perturb the original embedding via:

$$\boldsymbol{c}_0' = \boldsymbol{c}_0 + \boldsymbol{\epsilon}\cdot\mathbf{P}_{\text{min}}, \quad \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}). \tag{8}$$

Given a retain set $\mathbf{R}$, the augmentation process can be formulated as follows:

$$\mathbf{R}^{\text{aug}} : \mathbf{R} \mapsto \bigcup_{\boldsymbol{c}_0\in\mathbf{R}} \left\{\boldsymbol{c}_{0,k}' \mid k = 1,\ldots,N_A\right\}, \tag{9}$$

where N_A denotes the augmentation times and $\boldsymbol{c}_{0,k}'$ represents the k-th augmented embedding given $\boldsymbol{c}_0 \in \mathbf{R}$ using Eq. 8. In implementation, we first filter the original retain set $\mathbf{R}$ to obtain $\mathbf{R}_f$ using

[3]Empirically, the model parameter matrix $\mathbf{W}$ is usually full rank, thus its all singular values are non-zero.

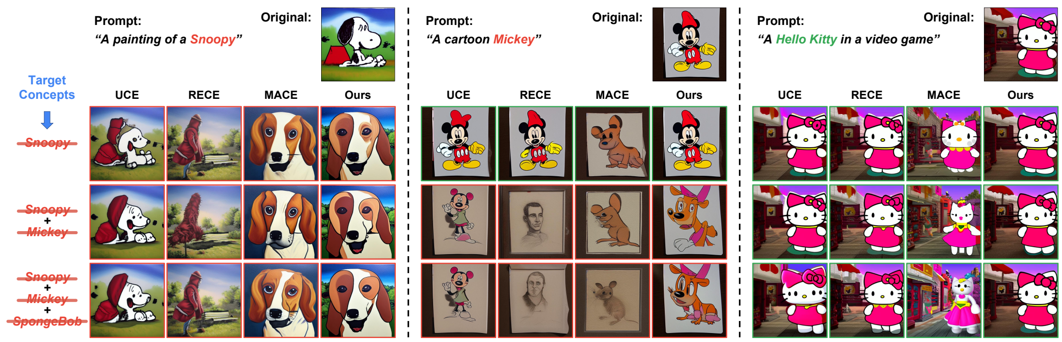


Figure 4: **Qualitative comparison of the few-concept erasure in erasing instances.** The erased and preserved generations are highlighted with **red** and **green** boxes, respectively. Our method exhibits consistent prior preservation with less semantic degradation for non-target concepts. For example, the middle column better retains details such as *Mickey*'s hat and button count, and the right column demonstrates more consistent *Hello Kitty* generations along with three concepts erased.

IPF. Subsequently, further augmentation and filtering are applied to $\mathbf{R}_f$ using DPA and IPF to obtain $(\mathbf{R}_f)_f^{\text{aug}}$, and the two filtered retain sets are then combined together to serve as the final refined retain set $\mathbf{R}_{\text{refine}} = \mathbf{R}_f \cup (\mathbf{R}_f)_f^{\text{aug}}$.

4.3 Invariant Equality Constraints (IEC)

In parallel, we identify certain invariants during the T2I generation process, *i.e.*, intermediate variables that remain unchanged with varying sampling prompts. One such invariant is the CLIP-encoded `[SOT]` token. Since the encoding process is masked by causal attention and all prompts are prefixed with the fixed `[SOT]` token during tokenization, its embedding consistently remains unchanged during T2I process. Another invariant is the null-text embedding, as it corresponds to the unconditional generation under the classifier-free guidance [24], which also remains unchanged despite prompt variations. Given the invariance of these embeddings, we consider additional protection measures to ensure their outputs remain unchanged during concept erasure. Specifically, we introduce explicit equality constraints over invariants based on Eq. 3:

$$\min_{\mathbf{\Delta}} \underbrace{\|(\mathbf{W}+\mathbf{\Delta P})\mathbf{C}_1 - \mathbf{W}\mathbf{C}_*\|^2}_{e_1} + \underbrace{\|\mathbf{\Delta P}\|^2}_{\text{regularization}}, \quad \text{s.t.} \ \underbrace{(\mathbf{\Delta P})\mathbf{C}_2 = \mathbf{0}}_{\text{equality constraints}}, \tag{10}$$

where $\mathbf{C}_2$ denotes the stacked invariant embedding matrix of `[SOT]` and null-text[4]. Derive the projection matrix $\mathbf{P}$ from $\mathbf{R}_{\text{refine}}$, we can compute the closed-form solution of Eq. 10 using Lagrange Multipliers from Appx. B.3:

$$(\mathbf{\Delta P})_{\text{Ours}} = \mathbf{W}\left(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top\right)\mathbf{PQM}, \tag{11}$$

where

$$\mathbf{M} = \left(\mathbf{C}_1\mathbf{C}_1^\top\mathbf{P} + \mathbf{I}\right)^{-1}, \mathbf{Q} = \mathbf{I} - \mathbf{M}\mathbf{C}_2\left(\mathbf{C}_2^\top\mathbf{P}\mathbf{M}\mathbf{C}_2\right)^{-1}\mathbf{C}_2^\top\mathbf{P}. \tag{12}$$

This closed-form solution enforces the equality constraints by projecting the parameter update onto the subspace orthogonal to the invariant embeddings. Since image generation inevitably depends on these invariant embeddings, such constraints inherently preserve prior knowledge.

5 Experiments

In this section, we conduct extensive experiments on three representative erasure tasks, including few-concept erasure, multi-concept erasure, and implicit concept erasure (Appx. D.3), validating our superior prior preservation. The compared baselines include ConAbl [29], MACE [33], RECE [19], and UCE [18], which have achieved SOTA performance across various concept erasure tasks. In implementation, we conduct all experiments on SDv1.4 [1] and generate each image using DPM-solver sampler [32] over 20 sampling steps with classifier-free guidance [24] of 7.5. More implementation details and compared baselines (*e.g.*, SPM [35]) can be found in Appx. C and Appx. D.4.

[4]Since the null-text embeddings are only composed of `[EOT]` tokens (excluding `[SOT]`), we use the k-means algorithm [36] to select k centroids to reduce redundancy.

Table 1: **Quantitative comparison of the few-concept erasure in erasing instances (left) and artistic styles (right)** following [35]. Arrows on the headers indicate the preferred direction for each metric, and the best results are highlighted in **bold**. Our method consistently improves prior preservation for non-target and general concepts from MS-COCO (shaded in pink) while achieving effective concept erasure. While our CS is not the lowest for target concpet, Appx. D.1 and Fig. 7 show our method is sufficient for erasure, and lower CS may further compromise prior preservation.

Concept	*Snoopy*	*Mickey*	*Spongebob*	*Pikachu*	*Hello Kitty*	MS-COCO	
	CS	CS	CS	CS	CS	CS	FID
SD v1.4	28.51	26.62	27.30	27.44	27.77	26.53	-
Erase *Snoopy*							
	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓	CS ↑	FID ↓
ConAbl	25.44	37.08	38.92	26.14	36.52	26.40	21.20
MACE	20.90	105.97	102.77	65.71	75.42	26.09	42.62
RECE	**18.38**	26.63	34.42	21.99	32.35	26.39	25.61
UCE	23.19	24.87	29.86	19.06	27.86	26.46	22.18
Ours	23.50	**23.41**	**24.64**	**16.81**	**21.74**	**26.48**	**19.95**
Erase *Snoopy* and *Mickey*							
	CS ↓	CS ↓	FID ↓	FID ↓	FID ↓	CS ↑	FID ↓
ConAbl	25.26	26.58	45.08	35.57	41.48	26.42	24.34
MACE	20.53	20.63	112.01	91.72	106.88	25.50	55.15
RECE	**18.57**	**19.14**	35.85	26.05	40.77	26.31	30.30
UCE	23.60	24.79	30.58	23.51	31.76	26.38	26.06
Ours	23.58	23.62	**29.67**	**22.51**	**28.23**	**26.47**	**23.66**
Erase *Snoopy* and *Mickey* and *Spongebob*							
	CS ↓	CS ↓	CS ↓	FID ↓	FID ↓	CS ↑	FID ↓
ConAbl	24.92	26.46	25.12	46.47	48.24	26.37	26.71
MACE	19.86	19.35	20.12	110.12	128.56	23.39	66.39
RECE	**18.17**	**18.87**	**16.23**	40.52	52.06	26.32	32.51
UCE	23.29	24.63	19.08	29.20	38.15	26.30	28.71
Ours	23.69	23.93	21.39	**21.40**	**26.22**	**26.51**	**24.99**

Concept	*Van Gogh*	*Picasso*	*Monet*	*P. Gauguin*	*Caravaggio*	MS-COCO	
	CS	CS	CS	CS	CS	CS	FID
SD v1.4	28.75	27.98	28.91	29.80	26.27	26.53	-
Erase *Van Gogh*							
	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓	CS ↑	FID ↓
ConAbl	28.16	77.01	63.80	63.20	79.25	26.46	**18.36**
MACE	26.66	69.92	60.88	56.18	69.04	26.50	23.15
RECE	26.39	60.57	61.09	47.07	72.85	26.52	23.54
UCE	28.10	43.02	40.49	32.62	61.72	26.54	19.63
Ours	**26.29**	**35.86**	**16.85**	**24.94**	**39.75**	**26.55**	20.36
Erase *Picasso*							
	FID ↓	CS ↓	FID ↓	FID ↓	FID ↓	CS ↑	FID ↓
ConAbl	60.44	26.97	36.23	65.23	79.12	26.43	20.02
MACE	59.58	26.48	37.02	46.35	66.20	26.47	22.86
RECE	51.09	26.66	25.39	46.08	75.61	26.48	23.03
UCE	37.58	26.99	**16.72**	32.48	59.27	26.50	20.33
Ours	**19.18**	**26.22**	19.87	**24.73**	**43.63**	**26.51**	**19.98**
Erase *Monet*							
	FID ↓	FID ↓	CS ↓	FID ↓	FID ↓	CS ↑	FID ↓
ConAbl	68.77	64.25	27.05	57.33	71.88	26.45	21.03
MACE	61.50	48.41	25.98	49.66	65.87	26.47	22.76
RECE	56.26	45.97	25.87	46.38	64.19	26.49	24.94
UCE	42.25	**38.73**	27.12	33.00	56.49	**26.51**	21.58
Ours	**28.78**	41.21	**25.06**	**27.85**	**55.20**	26.48	**20.87**

5.1 On Few-Concept Erasure

Evaluation setup. To compare the few-concept erasure performance with baseline methods, we conduct experiments on instance erasure and artistic style erasure following [35], where all methods are evaluated based on 80 instance templates and 30 artistic style templates, generating 10 images per template per concept. We use two metrics for evaluation: CLIP Score (CS) [43] measuring the text-image similarity and Fréchet Inception Distance (FID) [22] assessing the distributional distance before and after erasure. Following [35], we select non-target concepts with similar semantics to the target concept for comparison and report CS for targets and FID for non-targets in the main paper. Full comparisons are presented in Appx. D.2. We further compare the generations on MS-COCO captions [31], where we generate images with the first 1,000 captions, and report CS and FID to measure general knowledge preservation.

Analysis and discussion. Table 1 compares the results of erasing various instance concepts and artistic styles. Our method consistently achieves the lowest FIDs across all non-target concepts, demonstrating superior prior preservation with minimal alteration to the original content. Moreover, we emphasize that our erasure is sufficiently effective, even without achieving the lowest CS, as shown in Fig. 4 and Appx. D.1. In contrast, lower CS values typically indicate over-erasure, which results in excessive degradation of prior knowledge. Notably, with the number of target concepts increasing from 1 to 3, our FID in *Pikachu* rises from 16.81 to 21.40 (4.59 ↑), while UCE increases from 19.06 to 29.20 (10.14 ↑). A similar pattern is observed in *Hello Kitty* (Our 4.48 ↑ *v.s.* UCE's 10.29 ↑), showing our robustness in erasing increasing target concepts.

5.2 On Multi-Concept Erasure

Evaluation setup. Another more realistic erasure scenario is multi-concept erasure, where massive concepts are required to be erased at once. Herein, we follow the experiment setup in [33] for erasing multiple celebrities, where we experiment with erasing 10, 50, and 100 celebrities and collect another 100 celebrities as non-target concepts. We prepare 5 prompt templates for each celebrity concept. For non-target concepts, we generate 1 image per template for each of the 100 concepts, totaling 500 images. For target concepts, we adjust the per-concept quantity to maintain a total of 500 images (*e.g.*, erasing 10 celebrities involves generating 10 images with 5 templates per concept). In evaluation, we adopt GIPHY Celebrity Detector (GCD) [20] and measure the top-1 GCD accuracy, indicated by Acc_e for erased target concepts and Acc_r for retained non-target concepts. Meanwhile, the harmonic

Table 2: **Quantitative comparison of the multi-concept erasure in erasing 10, 50, and 100 celebrities.** The best results are highlighted in **bold**. Our method is capable of erasing up to 100 celebrities at once with low Acc_e (%) and preserving other non-target celebrities with less appearance alteration with high Acc_r (%), resulting in the best overall erasure performance H_o (shaded in pink).

	Erase 10 Celebrities			MS-COCO		Erase 50 Celebrities			MS-COCO		Erase 100 Celebrities			MS-COCO	
	$\text{Acc}_e \downarrow$	$\text{Acc}_r \uparrow$	$H_o \uparrow$	CS $\uparrow$	FID $\downarrow$	$\text{Acc}_e \downarrow$	$\text{Acc}_r \uparrow$	$H_o \uparrow$	CS $\uparrow$	FID $\downarrow$	$\text{Acc}_e \downarrow$	$\text{Acc}_r \uparrow$	$H_o \uparrow$	CS $\uparrow$	FID $\downarrow$
SD v1.4	91.99	89.66	14.70	26.53	-	93.08	89.66	12.85	26.53	-	90.18	89.66	17.70	26.53	-
ConAbl	60.76	77.89	52.19	25.60	42.12	64.00	75.44	48.74	14.30	255.36	42.86	58.82	57.97	14.93	235.27
UCE	0.20	71.19	83.10	24.07	83.81	0.00	31.94	48.41	13.45	209.93	0.00	20.92	34.60	13.49	185.46
RECE	0.34	67.43	80.44	16.75	170.65	1.03	19.77	32.95	13.49	213.39	2.43	23.71	38.16	12.09	177.57
MACE	1.62	87.73	92.75	26.36	37.25	3.41	84.31	90.03	25.45	45.31	4.80	80.20	87.06	24.80	50.41
Ours	1.81	89.09	**93.42**	**26.47**	**30.02**	3.46	88.48	**92.34**	**26.46**	**39.23**	5.87	85.54	**89.63**	**26.22**	**44.97**

Table 3: **Duration comparison (s) in erasing multiple celebrities** on one A100 GPU, where n is the number of target concepts. During data preparation, ConAbl requires pre-sampling 1,000 images (t_1) while MACE needs 8 pre-sampled images along with 8 segmentation masks (t_2) using SAM [27]. H_o is also included to compare multi-concept erasure performance.

	Training-based		Editing-based		
	ConAbl	MACE	UCE	RECE	Ours
Data Preparation	$n \times 1000$	$n \times (8+8)$	0	0	0
~~1 concept~~	1×90	55.1	1.2	1.5	3.6
$H_o \uparrow$	52.2	92.7	83.1	80.4	**93.4**
~~10 concepts~~	10×90	207.0	1.5	2.5	3.8
$H_o \uparrow$	48.7	90.0	48.4	33.0	**92.3**
~~100 concepts~~	100×90	1735.9	2.1	11.0	5.0
$H_o \uparrow$	58.0	87.1	34.6	38.2	**89.6**

Table 4: **Ablation study on proposed components** in erasing *Van Gogh*, with the non-target FID averaged over the other four artistic styles from Table 1. Ablation 1 corresponds to the original objective from [14] in Eq. 3. The ablated components include: IEC (Invariant Equality Constraints), IPF (Influence-based Prior Filtering), RPA (Random Prior Augmentation), and DPA (Directed Prior Augmentation).

Ablation	Components				Target	Non-Target	MS-COCO	
	IEC	IPF	RPA	DPA	CS $\downarrow$	FID $\downarrow$	CS $\uparrow$	FID $\downarrow$
1	$\times$	$\times$	$\times$	$\times$	27.20	50.43	26.42	26.33
2	✓	$\times$	$\times$	$\times$	27.20	48.17	26.44	24.95
3	✓	✓	$\times$	$\times$	26.68	38.02	26.54	20.57
4	✓	✓	✓	$\times$	26.30	32.62	26.52	20.99
Ours	✓	✓	$\times$	✓	**26.29**	**29.35**	**26.55**	**20.36**
SD v1.4	-	-	-	-	28.75	-	26.53	-

mean $H_o = \frac{2}{(1-\text{Acc}_e)^{-1}+(\text{Acc}_r)^{-1}}$ is adopted to assess the overall erasure performance. Additionally, we report the results on MS-COCO to demonstrate the prior preservation of general concepts.

Analysis and discussion. Table 2 showcases a notable improvement of our method on multi-concept erasure, particularly in prior preservation with the highest Acc_r. In comparison with the SOTA method, MACE [33], our method achieves superior prior preservation with better Acc_r, while maintaining comparable erasure efficacy, as reflected in similar Acc_e, resulting in the best overall erasure performance indicated by the highest H_o. Meanwhile, our method attains the lowest FID across all methods on MS-COCO. The other methods, UCE [18] and RECE [19], although achieving considerable balance in few-concept erasure, fail to maintain this balance as the number of target concepts increases as shown in Fig. 5, with catastrophic prior damage evidenced by MS-COCO as well. Notably, our method can erase up to 100 celebrities in 5 seconds, whereas MACE requires around 30 minutes ($\times$350 time). In real-world scenarios, this efficiency underscores our potential for the instant erasure of massive concepts.

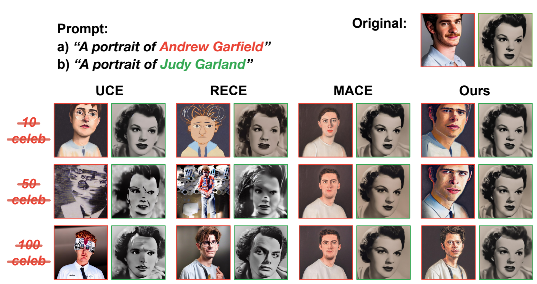
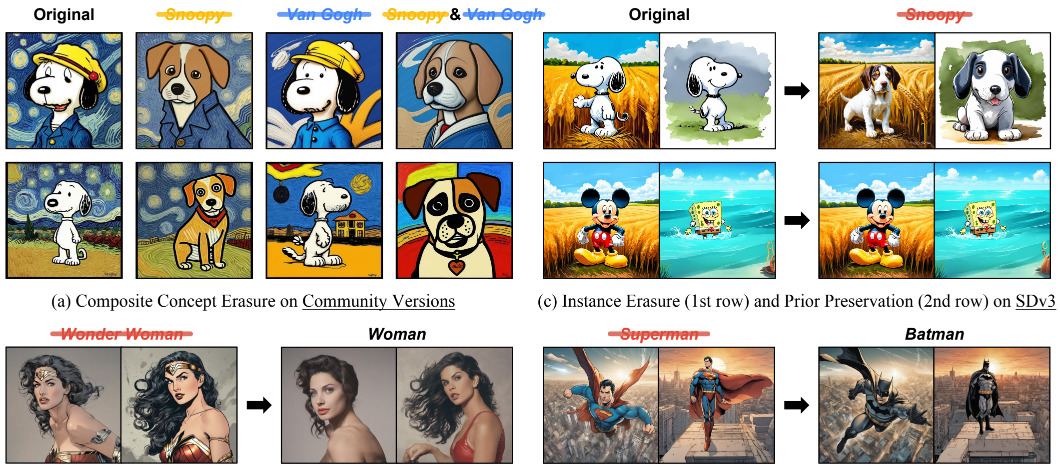


Figure 5: **Quantitative comparison of multi-concept erasure** in erasing celebrities (celeb). The erased and preserved generations are marked with **red** and **green** boxes. Our method precisely erases 100 celebrities while preserving generations of other non-target concepts.

5.3 Further Analysis

Duration comparison. Table 3 presents the duration comparison in erasing 1, 10, and 100 concepts across different methods. It is obvious that training-based methods necessitate significantly higher computational costs than editing-based ones. In contrast, our method achieves precise multi-concept



(a) Composite Concept Erasure on Community Versions

(c) Instance Erasure (1st row) and Prior Preservation (2nd row) on SDv3

(b) Knowledge Editing (*e.g.*, "*Wonder Woman* → *Woman*" and "*Superman* → *Batman*") on SDXL

Figure 6: **More applications across various T2I diffusion models.** (a) We conduct composite concept erasure for "*Snoopy* + *Van Gogh*" on DreamShaper [3] (1st row) and RealisticVision [4] (2nd row). (b) Our method also enables model knowledge editing by specifying the anchor concept on SDXL [42]. (c) Our method can seamlessly transfer to novel DiT-based T2I models, *e.g.*, SDv3 [12].

erasure in a remarkably short time, demonstrating superior efficiency while maintaining erasure performance, as evidenced in Table 2.

Component ablation. In Table 4, we compare the individual impact of our components on prior preservation and draw the following conclusions: (1) Impact of IEC (Ablation 1 *v.s.* 2): IEC reduces the non-target FID and the MS-COCO FID, demonstrating its effectiveness by preserving invariant embeddings with equality constraints. (2) Impact of IPF (Ablation 2 *v.s.* 3): Incorporating IPF results in a significant improvement in both FIDs, underscoring its critical role in filtering out less-influenced concepts in the retain set to mitigate semantic degradation. (3) Impact of DPA (Ablation 4 *v.s.* Ours): DPA improves RPA with directed noise and leads to a substantial improvement in non-target and MS-COCO FIDs, highlighting its advantage by introducing semantically similar concepts into the refined retain set. To conclude, the proposed three components (*i.e.*, IEC, IPF, and DPA) improve the prior preservation from different perspectives and contribute to our method with the best prior preservation under null space constraints. More ablations are presented in Appx. D.5.

More applications on other T2I models. To validate the transferability of our method across versatile applications, we conduct further experiments on various T2I models with different weights and architectures, including: (1) Composite concept erasure on DreamShaper [3] and RealisticVision [4] from Fig 6 (a): Our method can precisely erase the target concept(s) while preserving other non-target elements within the prompt, such as the *Van Gogh*-style background (2nd column) and the *Snoopy* character (3rd column). (2) Knowledge editing on SDXL [42] from Fig 6 (b): The arbitrary nature of anchor concepts allows us to edit the pre-trained model knowledge. Herein, our method effectively edits the model knowledge while maintaining the overall layout and semantics of the generated images. (3) Instance erasure on SDv3 [12] from Fig 6 (c): To accommodate the diffusion transformer (DiT) [41] architecture in T2I models, we adapt our method to a DiT-based model, demonstrating a well-balanced trade-off between erasure (1st row) and preservation (2nd row) as well.

6 Conclusion

This paper introduced SPEED, a scalable, precise, and efficient concept erasure method for T2I diffusion models. It formulates concept erasure as a null-space constrained optimization problem, facilitating effective prior preservation along with precise erasure efficacy. Critically, SPEED overcomes the inefficacy of editing-based methods in multi-concept erasure while circumventing the prohibitive computational costs associated with training-based approaches. With our proposed Prior Knowledge Refinement involving three complementary techniques, SPEED not only ensures superior prior preservation but also achieves a $350\times$ acceleration in multi-concept erasure, establishing itself as a scalable and practical solution for real-world applications.

References

[1] Stable diffusion. https://huggingface.co/CompVis/stable-diffusion-v1-4, 2022.

[2] Stable diffusion v2.1. https://huggingface.co/stabilityai/stable-diffusion-2-1, 2022.

[3] Dreamshaper. https://huggingface.co/Lykon/dreamshaper-8, 2023.

[4] Realisticvsion. https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE, 2023.

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] P Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019.

[7] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[8] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[9] Huiqiang Chen, Tianqing Zhu, Xin Yu, and Wanlei Zhou. Machine unlearning via null space calibration. *arXiv preprint arXiv:2404.13588*, 2024.

[10] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, and Jiliang Tang. Diffusionshield: A watermark for copyright protection against generative diffusion models. *arXiv preprint arXiv:2306.04642*, 2023.

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[14] Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Xiang Wang, Xiangnan He, and Tat-seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*, 2024.

[15] Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, and Wangmeng Zuo. Learning federated visual prompt in null space for mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8064–8073, 2023.

[16] 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. *arXiv preprint arXiv:2208.01618*, 2022.

[17] 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*, pages 2426–2436, 2023.

[18] 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*, pages 5111–5120, 2024.

[19] 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*, pages 73–88. Springer, 2025.

[20] Nick Hasty, Ihor Kroosh, Dmitry Voitekh, and Dmytro Korduban. Giphy celebrity detector. https://github.com/Giphy/celeb-detection-oss.

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[22] 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. *Advances in neural information processing systems*, 30, 2017.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[26] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[28] Yajing Kong, Liu Liu, Zhen Wang, and Dacheng Tao. Balancing stability and plasticity through advanced null space in continual learning. In *European Conference on Computer Vision*, pages 219–236. Springer, 2022.

[29] 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*, pages 22691–22702, 2023.

[30] Guoliang Lin, Hanlu Chu, and Hanjiang Lai. Towards better plasticity-stability trade-off in incremental learning: A simple linear connector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 89–98, 2022.

[31] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[32] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[33] 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*, pages 6430–6440, 2024.

[34] Yue Lu, Shizhou Zhang, De Cheng, Yinghui Xing, Nannan Wang, Peng Wang, and Yanning Zhang. Visual prompt tuning in null space for continual learning. *arXiv preprint arXiv:2406.05658*, 2024.

[35] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024.

[36] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

[37] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[38] OpenAI. OpenAI: Introducing ChatGPT, 2022.

[39] OpenAI. Dall·e 3 system card. 2023.

[40] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7061, 2023.

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[42] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[43] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[45] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

[46] Dana Rao. Responsible innovation in the age of generative ai, 2023.

[47] 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*, pages 10684–10695, 2022.

[48] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[49] 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*, pages 22522–22531, 2023.

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

[51] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[52] 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. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.

[53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

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

[55] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[56] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[57] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[58] Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 184–193, 2021.

[59] Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance with self-supervision for incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[60] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*.

[61] 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. *arXiv preprint arXiv:2412.06143*, 2024.

[62] Chengyi Yang, Mingda Dong, Xiaoyue Zhang, Jiayin Qi, and Aimin Zhou. Introducing common null space of gradients for gradient projection methods in continual learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5489–5497, 2024.

[63] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[64] 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*, pages 7737–7746, 2024.

[65] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024.

[66] 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*, pages 385–403. Springer, 2025.

A Preliminaries

T2I diffusion models. T2I generation has seen significant advancements with diffusion models, particularly Latent Diffusion Models (LDMs) [47]. Unlike pixel-space diffusion, LDMs operate in the latent space of a pretrained autoencoder, reducing computational costs while maintaining high-quality synthesis. LDMs consist of a vector-quantized autoencoder [57, 13] and a diffusion model [11, 23, 53, 26, 55]. The autoencoder encodes an image $\boldsymbol{x}$ into a latent representation $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$ and reconstructs it via $\boldsymbol{x} \approx \mathcal{D}(\boldsymbol{z})$. The diffusion model learns to generate latent codes through a denoising process. The training objective is given by [23, 47]:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{E}(\boldsymbol{x}),\boldsymbol{c},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c})\|_2^2\right], \tag{13}$$

where $\boldsymbol{z}_t$ is the noisy latent at timestep t, $\boldsymbol{\epsilon}$ is Gaussian noise, $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is the denoising network, and $\boldsymbol{c}$ is conditioning information from text, class labels, or segmentation masks [47]. During inference, a latent $\boldsymbol{z}_T$ is sampled from a Gaussian prior and progressively denoised to obtain $\boldsymbol{z}_0$, which is then decoded into an image via $\boldsymbol{x}_0 \approx \mathcal{D}(\boldsymbol{z}_0)$.

Cross-attention mechanisms. Current T2I diffusion models usually leverage a generative framework to synthesize images conditioned on textual descriptions in the latent space [47]. The conditioning mechanism is implemented through cross-attention (CA) layers. Specifically, textual descriptions are first tokenized into n tokens and embedded into a sequence of vectors $\boldsymbol{e} \in \mathbb{R}^{d_0\times n}$ via a pre-trained CLIP model [43]. These text embeddings serve as the key $\mathbf{K} \in \mathbb{R}^{n\times d_k}$ and value $\mathbf{V} \in \mathbb{R}^{n\times d_v}$ inputs using parametric projection matrices $\mathbf{W}_{\mathbf{K}} \in \mathbb{R}^{d_k\times d_0}$ and $\mathbf{W}_{\mathbf{V}} \in \mathbb{R}^{d_v\times d_0}$, while the intermediate image representations act as the query $\mathbf{Q} \in \mathbb{R}^{m\times d_k}$. The cross-attention mechanism is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{14}$$

This alignment enables the model to capture semantic correlations between the textual input and the visual features, ensuring that the generated images are semantically consistent with the provided text prompts.

B Proof and Derivation

B.1 Deriving the Closed-Form Solution for UCE

From Eq. 1, we are tasked with minimizing the following editing objective, where the hyperparameters α and β correspond to the weights of the erasure error e_1 and the preservation error e_0, respectively:

$$\min_{\boldsymbol{\Delta}}\left[\alpha\|(\mathbf{W}+\boldsymbol{\Delta})\mathbf{C}_1 - \mathbf{W}\mathbf{C}_*\|^2 + \beta\|\boldsymbol{\Delta}\mathbf{C}_0\|^2\right]. \tag{15}$$

To derive the closed-form solution, we begin by computing the gradient of the objective function with respect to $\boldsymbol{\Delta}$. The gradient is given by:

$$\alpha\left(\mathbf{W}\mathbf{C}_1 - \mathbf{W}\mathbf{C}_* + \boldsymbol{\Delta}\mathbf{C}_1\right)\mathbf{C}_1^\top + \beta\boldsymbol{\Delta}\mathbf{C}_0\mathbf{C}_0^\top = 0. \tag{16}$$

Solving the resulting equation yields the closed-form solution for $\boldsymbol{\Delta}_{\text{UCE}}$:

$$\boldsymbol{\Delta}_{\text{UCE}} = \alpha\mathbf{W}\left(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top\right)\left(\alpha\mathbf{C}_1\mathbf{C}_1^\top + \beta\mathbf{C}_0\mathbf{C}_0^\top\right)^{-1}. \tag{17}$$

In practice, an additional identity matrix $\mathbf{I}$ with hyperparameter λ is added to $\left(\alpha\mathbf{C}_1\mathbf{C}_1^\top + \beta\mathbf{C}_0\mathbf{C}_0^\top\right)^{-1}$ to ensure its invertibility. This modification results in the following closed-form solution for UCE:

$$\boldsymbol{\Delta}_{\text{UCE}} = \alpha\mathbf{W}\left(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top\right)\left(\alpha\mathbf{C}_1\mathbf{C}_1^\top + \beta\mathbf{C}_0\mathbf{C}_0^\top + \lambda\mathbf{I}\right)^{-1}. \tag{18}$$

B.2 Proof of the Lower Bound of e_0 for UCE

Herein, we aim to establish the existence of a strictly positive constant $c > 0$ such that

$$e_0 = \|\boldsymbol{\Delta}_{\text{UCE}}\mathbf{C}_0\|^2 = \|\alpha\mathbf{W}\left(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top\right)(\alpha\mathbf{C}_1\mathbf{C}_1^\top + \beta\mathbf{C}_0\mathbf{C}_0^\top + \lambda\mathbf{I})^{-1}\mathbf{C}_0\|^2 \geq c > 0. \tag{19}$$

Assumption B.1. *We assume that $\alpha, \beta, \lambda \neq 0$, that $\mathbf{W}$ is a full-rank matrix, and that $\mathbf{C}_0\mathbf{C}_0^\top$ is rank-deficient. Furthermore, we assume that*

$$\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top \neq \mathbf{0}.$$

Proof. Define the matrix $\mathbf{M}$ as

$$\mathbf{M} = \alpha\mathbf{C}_1\mathbf{C}_1^\top + \beta\mathbf{C}_0\mathbf{C}_0^\top + \lambda\mathbf{I}. \tag{20}$$

Since $\lambda > 0$ and $\mathbf{I}$ is positive definite, it follows that $\mathbf{M}$ is strictly positive definite and therefore invertible.

Rewriting e_0 by defining $\mathbf{B} = \mathbf{M}^{-1}\mathbf{C}_0$, we obtain

$$e_0 = \|\alpha\mathbf{W}(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{B}\|^2. \tag{21}$$

Applying the singular value bound for matrix products, we have

$$\|\mathbf{X}\mathbf{Y}\| \geq \sigma_{\min}(\mathbf{X})\|\mathbf{Y}\|, \tag{22}$$

where $\sigma_{\min}(\mathbf{X})$ is the smallest singular value of $\mathbf{X}$. Applying this inequality, we obtain

$$\|\mathbf{W}(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{B}\| \geq \sigma_{\min}(\mathbf{W})\|(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{B}\|. \tag{23}$$

We start with the singular value decomposition (SVD) of the matrix $\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top$, given by

$$\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top. \tag{24}$$

Here, $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and

$$\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_r, 0, \dots, 0) \tag{25}$$

is a diagonal matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, followed by zeros.

Multiplying both sides by $\mathbf{B}$, we obtain

$$(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{B}. \tag{26}$$

Define the projection of $\mathbf{B}$ onto the subspace spanned by the right singular vectors as

$$\mathbf{B}_{\text{proj}} = \mathbf{V}^\top\mathbf{B}. \tag{27}$$

Then, we can rewrite the expression as

$$(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{B}_{\text{proj}}. \tag{28}$$

Taking norms on both sides and using the fact that orthogonal transformations preserve norms, we get

$$\|(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{B}\| = \|\mathbf{\Sigma}\mathbf{B}_{\text{proj}}\|. \tag{29}$$

Since $\mathbf{\Sigma}$ is a diagonal matrix, its smallest nonzero singular value σ_r provides a lower bound:

$$\|\mathbf{\Sigma}\mathbf{B}_{\text{proj}}\| \geq \sigma_r\|\mathbf{B}_{\text{proj}}\|. \tag{30}$$

Next, we establish a lower bound for $\|\mathbf{B}_{\text{proj}}\|$. Given that $\mathbf{V}$ is composed of right singular vectors, there exists a smallest non-zero singular value c_1 such that:

$$\|\mathbf{B}_{\text{proj}}\| \geq c_1\|\mathbf{B}\|. \tag{31}$$

Combining these inequalities, we obtain

$$\|(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{B}\| \geq \sigma_r\|\mathbf{B}_{\text{proj}}\| \geq \sigma_r c_1\|\mathbf{B}\|. \tag{32}$$

Since $\mathbf{M}$ is positive definite, we use the standard norm inequality for an invertible matrix $\mathbf{M}$, which states that for any matrix $\mathbf{X}$,

$$\|\mathbf{M}\mathbf{X}\| \leq \|\mathbf{M}\|\|\mathbf{X}\|. \tag{33}$$

Setting $\mathbf{X} = \mathbf{M}^{-1}\mathbf{C}_0$, we obtain

$$\|\mathbf{M}\mathbf{M}^{-1}\mathbf{C}_0\| \leq \|\mathbf{M}\|\|\mathbf{M}^{-1}\mathbf{C}_0\|. \tag{34}$$

Since $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$, the left-hand side simplifies to $\|\mathbf{C}_0\|$, yielding

$$\|\mathbf{C}_0\| \leq \|\mathbf{M}\|\|\mathbf{M}^{-1}\mathbf{C}_0\|. \tag{35}$$

Dividing both sides by $\|\mathbf{M}\|$, we obtain

$$\|\mathbf{M}^{-1}\mathbf{C}_0\| \geq \frac{1}{\|\mathbf{M}\|}\|\mathbf{C}_0\|. \tag{36}$$

Thus, it follows that

$$\|\mathbf{B}\| = \|\mathbf{M}^{-1}\mathbf{C}_0\| \geq \frac{1}{\|\mathbf{M}\|}\|\mathbf{C}_0\|. \tag{37}$$

Combining the above results, we obtain

$$\|\mathbf{W}(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{B}\| \geq \sigma_{\min}(\mathbf{W})\sigma_r c_1 \frac{1}{\|\mathbf{M}\|}\|\mathbf{C}_0\|. \tag{38}$$

Squaring both sides, we conclude that

$$e_0 = \|\alpha\mathbf{W}(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{B}\|^2 \geq \alpha^2\sigma_{\min}^2(\mathbf{W})\sigma_r^2 c_1^2 \frac{1}{\|\mathbf{M}\|^2}\|\mathbf{C}_0\|^2. \tag{39}$$

Since all terms on the right-hand side are strictly positive by assumption, we establish the existence of a positive lower bound $c > 0$ such that

$$e_0 \geq c > 0. \tag{40}$$

This completes the proof. □

B.3 Deriving the Closed-Form Solution for SPEED

From Eq. 10, we are tasked with minimizing the following editing objective:

$$\min_{\mathbf{\Delta}} \|(\mathbf{W} + \mathbf{\Delta P})\mathbf{C}_1 - \mathbf{W}\mathbf{C}_*\|^2 + \|\mathbf{\Delta P}\|^2, \quad \text{s.t.} (\mathbf{\Delta P})\mathbf{C}_2 = \mathbf{0}. \tag{41}$$

This is a weighted least squares problem subject to an equality constraint. To solve it, we first formulate the Lagrangian function, where $\mathbf{\Lambda}$ is the Lagrange multiplier:

$$\mathcal{L}(\mathbf{\Delta}, \mathbf{\Lambda}) = \|(\mathbf{W} + \mathbf{\Delta P})\mathbf{C}_1 - \mathbf{W}\mathbf{C}_*\|^2 + \|\mathbf{\Delta P}\|^2 + \mathbf{\Lambda}^\top ((\mathbf{\Delta P})\mathbf{C}_2). \tag{42}$$

We compute the gradient of the Lagrangian function in Eq. 42 with respect to $\mathbf{\Delta}$ and set it to zero, yielding the following equation for $\mathbf{\Delta}$:

$$\frac{\partial\mathcal{L}(\mathbf{\Delta}, \mathbf{\Lambda})}{\partial\mathbf{\Delta}} = 2\left((\mathbf{W} + \mathbf{\Delta P})\mathbf{C}_1 - \mathbf{W}\mathbf{C}_*\right)\mathbf{C}_1^\top\mathbf{P}^\top + 2\mathbf{\Delta P}\mathbf{P}^\top + \mathbf{\Lambda}\mathbf{C}_2^\top\mathbf{P}^\top = \mathbf{0}. \tag{43}$$

Given that the projection matrix $\mathbf{P}$ is derived from R_{refine} using Eq. 2, $\mathbf{P}$ is a symmetric matrix (*i.e.*, $\mathbf{P} = \mathbf{P}^\top$) and an idempotent matrix (*i.e.*, $\mathbf{P}^2 = \mathbf{P}$), the above formulation can be simplified to:

$$\frac{\partial\mathcal{L}(\mathbf{\Delta}, \mathbf{\Lambda})}{\partial\mathbf{\Delta}} = 2\left((\mathbf{W} + \mathbf{\Delta P})\mathbf{C}_1 - \mathbf{W}\mathbf{C}_*\right)\mathbf{C}_1^\top\mathbf{P} + 2\mathbf{\Delta P} + \mathbf{\Lambda}\mathbf{C}_2^\top\mathbf{P} = \mathbf{0}. \tag{44}$$

Therefore, we can obtain the closed-form solution for $\mathbf{\Delta P}$ from this equation:

$$\mathbf{\Delta P} = (\mathbf{W}\mathbf{C}_*\mathbf{C}_1^\top\mathbf{P} - \mathbf{W}\mathbf{C}_1\mathbf{C}_1^\top\mathbf{P} - \frac{1}{2}\mathbf{\Lambda}\mathbf{C}_2^\top\mathbf{P})(\mathbf{C}_1\mathbf{C}_1^\top\mathbf{P} + \mathbf{I})^{-1}. \tag{45}$$

Next, we differentiate the Lagrangian function in Eq. 42 with respect to $\mathbf{\Lambda}$ and set it to zero:

$$\frac{\partial\mathcal{L}(\mathbf{\Delta}, \mathbf{\Lambda})}{\partial\mathbf{\Lambda}} = (\mathbf{\Delta P})\mathbf{C}_2 = \mathbf{0}. \tag{46}$$

For simplicity, we define $\mathbf{M} = (\mathbf{C}_1\mathbf{C}_1^\top\mathbf{P} + \mathbf{I})^{-1}$. Then, we substitute the result of Eq.45 into Eq.46 and obtain:

$$(\mathbf{W}\mathbf{C}_*\mathbf{C}_1^\top\mathbf{P} - \mathbf{W}\mathbf{C}_1\mathbf{C}_1^\top\mathbf{P} - \frac{1}{2}\mathbf{\Lambda}\mathbf{C}_2^\top\mathbf{P})\mathbf{M}\mathbf{C}_2 = \mathbf{0}. \tag{47}$$

Solving this equation leads to:

$$\frac{1}{2}\mathbf{\Lambda} = \mathbf{W}(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{P}\mathbf{M}\mathbf{C}_2(\mathbf{C}_2^\top\mathbf{P}\mathbf{M}\mathbf{C}_2)^{-1}. \tag{48}$$

Substituting Eq.48 back into Eq.45, we have the closed-form solution of our objective:

$$(\mathbf{\Delta P})_{\text{SPEED}} = \mathbf{W}(\mathbf{C}_*\mathbf{C}_1^\top - \mathbf{C}_1\mathbf{C}_1^\top)\mathbf{P}\mathbf{Q}\mathbf{M}, \tag{49}$$

where $\mathbf{Q} = \mathbf{I} - \mathbf{M}\mathbf{C}_2(\mathbf{C}_2^\top\mathbf{P}\mathbf{M}\mathbf{C}_2)^{-1}\mathbf{C}_2^\top\mathbf{P}$ and $\mathbf{M} = (\mathbf{C}_1\mathbf{C}_1^\top\mathbf{P} + \mathbf{I})^{-1}$.

Table 5: **The evaluation setup for multi-concept erasure.** This celebrity dataset contains an erasure set with 100 celebrities and a retain set with another 100 celebrities. We experiment with erasing 10, 50, and 100 celebrities with the pre-defined target concepts and the entire retain set is utilized in all cases.

Group	Number	Anchor Concept	Celebrity
Erasure Set	10	'person'	*'Adam Driver', 'Adriana Lima', 'Amber Heard', 'Amy Adams', 'Andrew Garfield', 'Angelina Jolie', 'Anjelica Huston', 'Anna Faris', 'Anna Kendrick', 'Anne Hathaway'*
	50	'person'	*'Adam Driver', 'Adriana Lima', 'Amber Heard', 'Amy Adams', 'Andrew Garfield', 'Angelina Jolie', 'Anjelica Huston', 'Anna Faris', 'Anna Kendrick', 'Anne Hathaway', 'Arnold Schwarzenegger', 'Barack Obama', 'Beth Behrs', 'Bill Clinton', 'Bob Dylan', 'Bob Marley', 'Bradley Cooper', 'Bruce Willis', 'Bryan Cranston', 'Cameron Diaz', 'Channing Tatum', 'Charlie Sheen', 'Charlize Theron', 'Chris Evans', 'Chris Hemsworth', 'Chris Pine', 'Chuck Norris', 'Courteney Cox', 'Demi Lovato', 'Drake', 'Drew Barrymore', 'Dwayne Johnson', 'Ed Sheeran', 'Elon Musk', 'Elvis Presley', 'Emma Stone', 'Frida Kahlo', 'George Clooney', 'Glenn Close', 'Gwyneth Paltrow', 'Harrison Ford', 'Hillary Clinton', 'Hugh Jackman', 'Idris Elba', 'Jake Gyllenhaal', 'James Franco', 'Jared Leto', 'Jason Momoa', 'Jennifer Aniston', 'Jennifer Lawrence'*
	100	'person'	*'Adam Driver', 'Adriana Lima', 'Amber Heard', 'Amy Adams', 'Andrew Garfield', 'Angelina Jolie', 'Anjelica Huston', 'Anna Faris', 'Anna Kendrick', 'Anne Hathaway', 'Arnold Schwarzenegger', 'Barack Obama', 'Beth Behrs', 'Bill Clinton', 'Bob Dylan', 'Bob Marley', 'Bradley Cooper', 'Bruce Willis', 'Bryan Cranston', 'Cameron Diaz', 'Channing Tatum', 'Charlie Sheen', 'Charlize Theron', 'Chris Evans', 'Chris Hemsworth', 'Chris Pine', 'Chuck Norris', 'Courteney Cox', 'Demi Lovato', 'Drake', 'Drew Barrymore', 'Dwayne Johnson', 'Ed Sheeran', 'Elon Musk', 'Elvis Presley', 'Emma Stone', 'Frida Kahlo', 'George Clooney', 'Glenn Close', 'Gwyneth Paltrow', 'Harrison Ford', 'Hillary Clinton', 'Hugh Jackman', 'Idris Elba', 'Jake Gyllenhaal', 'James Franco', 'Jared Leto', 'Jason Momoa', 'Jennifer Aniston', 'Jennifer Lawrence', 'Jennifer Lopez', 'Jeremy Renner', 'Jessica Biel', 'Jessica Chastain', 'John Oliver', 'John Wayne', 'Johnny Depp', 'Julianne Hough', 'Justin Timberlake', 'Kate Bosworth', 'Kate Winslet', 'Leonardo Dicaprio', 'Margot Robbie', 'Mariah Carey', 'Melania Trump', 'Meryl Streep', 'Mick Jagger', 'Mila Kunis', 'Milla Jovovich', 'Morgan Freeman', 'Nick Jonas', 'Nicolas Cage', 'Nicole Kidman', 'Octavia Spencer', 'Olivia Wilde', 'Oprah Winfrey', 'Paul Mccartney', 'Paul Walker', 'Peter Dinklage', 'Philip Seymour Hoffman', 'Reese Witherspoon', 'Richard Gere', 'Ricky Gervais', 'Rihanna', 'Robin Williams', 'Ronald Reagan', 'Ryan Gosling', 'Ryan Reynolds', 'Shia Labeouf', 'Shirley Temple', 'Spike Lee', 'Stan Lee', 'Theresa May', 'Tom Cruise', 'Tom Hanks', 'Tom Hardy', 'Tom Hiddleston', 'Whoopi Goldberg', 'Zac Efron', 'Zayn Malik'*
Retain Set	10, 50, and 100	-	*'Aaron Paul', 'Alec Baldwin', 'Amanda Seyfried', 'Amy Poehler', 'Amy Schumer', 'Amy Winehouse', 'Andy Samberg', 'Aretha Franklin', 'Avril Lavigne', 'Aziz Ansari', 'Barry Manilow', 'Ben Affleck', 'Ben Stiller', 'Benicio Del Toro', 'Bette Midler', 'Betty White', 'Bill Murray', 'Bill Nye', 'Britney Spears', 'Brittany Snow', 'Bruce Lee', 'Burt Reynolds', 'Charles Manson', 'Christie Brinkley', 'Christina Hendricks', 'Clint Eastwood', 'Countess Vaughn', 'Dakota Johnson', 'Dane Dehaan', 'David Bowie', 'David Tennant', 'Denise Richards', 'Doris Day', 'Dr Dre', 'Elizabeth Taylor', 'Emma Roberts', 'Fred Rogers', 'Gal Gadot', 'George Bush', 'George Takei', 'Gillian Anderson', 'Gordon Ramsey', 'Halle Berry', 'Harry Dean Stanton', 'Harry Styles', 'Hayley Atwell', 'Heath Ledger', 'Henry Cavill', 'Jackie Chan', 'Jada Pinkett Smith', 'James Garner', 'Jason Statham', 'Jeff Bridges', 'Jennifer Connelly', 'Jensen Ackles', 'Jim Morrison', 'Jimmy Carter', 'Joan Rivers', 'John Lennon', 'Johnny Cash', 'Jon Hamm', 'Judy Garland', 'Julianne Moore', 'Justin Bieber', 'Kaley Cuoco', 'Kate Upton', 'Keanu Reeves', 'Kim Jong Un', 'Kirsten Dunst', 'Kristen Stewart', 'Krysten Ritter', 'Lana Del Rey', 'Leslie Jones', 'Lily Collins', 'Lindsay Lohan', 'Liv Tyler', 'Lizzy Caplan', 'Maggie Gyllenhaal', 'Matt Damon', 'Matt Smith', 'Matthew Mcconaughey', 'Maya Angelou', 'Megan Fox', 'Mel Gibson', 'Melanie Griffith', 'Michael Cera', 'Michael Ealy', 'Natalie Portman', 'Neil Degrasse Tyson', 'Niall Horan', 'Patrick Stewart', 'Paul Rudd', 'Paul Wesley', 'Pierce Brosnan', 'Prince', 'Queen Elizabeth', 'Rachel Dratch', 'Rachel Mcadams', 'Reba Mcentire', 'Robert De Niro'*

C Implementation Details

C.1 Experimental Setup Details

Few-concept erasure. We first compare methods on few-concept erasure, a fundamental concept erasure task, including both instance erasure and artistic style erasure following [35]. For instance erasure, we prepare 80 instance templates proposed in CLIP [43], such as *"a photo of the {Instance}"*, *"a drawing of the {Instance}"*, and *"a painting of the {Instance}"*. For artistic style erasure, we use ChatGPT [38, 5] to generate 30 artistic style templates, including *"{Artistic} style painting of the night sky with bold strokes"*, *"{Artistic} style landscape of rolling hills with dramatic brushwork"*, and *"Sunrise scene in {Artistic} style, capturing the beauty of dawn"*. Following [35], we handpick the representative target and anchor concepts as the erasure set (*i.e.*, *Snoopy*, *Mickey*, *SpongeBob* → ' ' in instance erasure and *Van Gogh*, *Picasso*, *Monet* → 'art' in artistic style erasure) and non-target concepts for evaluation (*i.e.*, *Pikachu* and *Hello Kitty* in instance erasure and *Paul Gauguin* and *Caravaggio* in artistic style erasure). In terms of the retain set, for instance erasure, we use a scraping script to crawl Wikipedia category pages to extract fictional character names and their page view counts with a threshold of 500,000 views from 2020.01.01 to 2023.12.31, resulting in 1,352 instances. For artistic style erasure, we use the 1,734 artistic styles collected from UCE [18]. In evaluation, we generate 10 images per template per concept, resulting in 800 and 300 images for each concept in instance erasure and artistic style erasure, respectively. Moreover, we introduce the MS-COCO

Table 6: **Full quantitative comparison of the few-concept erasure in erasing instances from Table 1 (left).** The best results are highlighted in **bold**, and grey columns are indirect indicators for measuring erasure efficacy on target concepts or prior preservation on non-target concepts.

	Snoopy		*Mickey*		*Spongebob*		*Pikachu*		*Hello Kitty*		MS-COCO	
	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID
SD v1.4	28.51	-	26.62	-	27.30	-	27.44	-	27.77	-	26.53	-
Erase ***Snoopy***												
	CS ↓	FID ↑	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓
ConAbl	25.44	98.38	26.63	37.08	26.95	38.92	27.47	26.14	27.65	36.52	26.40	21.20
MACE	20.90	**165.74**	23.46	105.97	23.35	102.77	26.05	65.71	26.05	75.42	26.09	42.62
RECE	**18.38**	151.46	26.62	26.63	27.23	34.42	27.47	21.99	27.78	32.35	26.39	25.61
UCE	23.19	102.86	26.64	24.87	27.29	29.86	27.47	19.06	27.75	27.86	26.46	22.18
Ours	23.50	108.51	**26.67**	**23.41**	**27.31**	**24.64**	**27.48**	**16.81**	**27.82**	**21.74**	**26.48**	**19.95**
Erase ***Snoopy*** and ***Mickey***												
	CS ↓	FID ↑	CS ↓	FID ↑	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓
ConAbl	25.26	106.78	26.58	57.05	26.81	45.08	27.34	35.57	27.74	41.48	26.42	24.34
MACE	20.53	**170.01**	20.63	142.98	22.03	112.01	24.98	91.72	23.64	106.88	25.50	55.15
RECE	**18.57**	150.84	**19.14**	**145.59**	27.29	35.85	27.37	26.05	27.71	40.77	26.31	30.30
UCE	23.60	99.30	24.79	86.32	27.32	30.58	27.38	23.51	27.74	31.76	26.38	26.06
Ours	23.58	103.62	23.62	83.70	**27.34**	**29.67**	**27.39**	**22.51**	**27.78**	**28.23**	**26.47**	**23.66**
Erase ***Snoopy*** and ***Mickey*** and ***Spongebob***												
	CS ↓	FID ↑	CS ↓	FID ↑	CS ↓	FID ↑	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓
ConAbl	24.92	112.66	26.46	63.95	25.12	102.68	27.36	46.47	27.72	48.24	26.37	26.71
MACE	19.86	**175.43**	19.35	140.13	20.12	143.17	19.76	110.12	21.03	128.56	23.39	66.39
RECE	**18.17**	155.26	**18.87**	**149.77**	**16.23**	**178.55**	27.34	40.52	27.71	52.06	26.32	32.51
UCE	23.29	101.40	24.63	88.11	19.08	140.40	27.45	29.20	**27.82**	38.15	26.30	28.71
Ours	23.69	103.33	23.93	86.55	21.39	109.28	**27.47**	**21.40**	27.76	**26.22**	**26.51**	**24.99**

Table 7: **Full quantitative comparison of the few-concept erasure in erasing artistic styles from Table 1 (right).** The best results are highlighted in **bold**, and the grey columns are indirect indicators for measuring erasure efficacy on target concepts or prior preservation on non-target concepts.

	Van Gogh		*Picasso*		*Monet*		*Paul Gauguin*		*Caravaggio*		MS-COCO	
	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID
SD v1.4	28.75	-	27.98	-	28.91	-	29.80	-	26.27	-	26.53	-
Erase ***Van Gogh***												
	CS ↓	FID ↑	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓
ConAbl	28.16	129.57	27.07	77.01	28.44	63.80	29.49	63.20	26.15	79.25	26.46	**18.36**
MACE	26.66	169.60	27.39	69.92	28.84	60.88	29.39	56.18	26.19	69.04	26.50	23.15
RECE	26.39	**171.70**	27.58	60.57	28.83	61.09	29.58	47.07	26.21	72.85	26.52	23.54
UCE	28.10	133.87	27.70	43.02	28.92	40.49	29.62	32.62	26.23	61.72	26.54	19.63
Ours	**26.29**	131.02	**27.96**	**35.86**	**28.94**	**16.85**	**29.71**	**24.94**	**26.24**	**39.75**	**26.55**	20.36
Erase ***Picasso***												
	CS ↑	FID ↓	CS ↓	FID ↑	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓
ConAbl	28.66	60.44	26.97	131.45	28.72	36.23	29.68	65.23	26.20	79.12	26.43	20.02
MACE	28.68	59.58	26.48	**137.09**	28.73	37.02	29.71	46.35	26.23	66.20	26.47	22.86
RECE	28.71	51.09	26.66	126.40	28.87	25.39	29.69	46.08	26.22	75.61	26.48	23.03
UCE	28.72	37.58	26.99	102.21	**28.92**	**16.72**	29.71	32.48	26.22	59.27	26.50	20.33
Ours	**28.76**	**19.18**	**26.22**	117.71	28.88	19.87	**29.75**	**24.73**	**26.24**	**43.63**	**26.51**	**19.98**
Erase ***Monet***												
	CS ↑	FID ↓	CS ↑	FID ↓	CS ↓	FID ↑	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓
ConAbl	28.58	68.77	27.43	64.25	27.05	96.67	29.09	57.33	26.09	71.88	26.45	21.03
MACE	28.56	61.50	27.74	48.41	25.98	116.34	29.39	49.66	25.98	65.87	26.47	22.76
RECE	28.63	56.26	27.88	45.97	25.87	121.28	29.43	46.38	26.20	64.19	26.49	24.94
UCE	28.65	42.25	27.91	**38.73**	27.12	98.37	29.58	33.00	26.16	56.49	**26.51**	21.58
Ours	**28.76**	**28.78**	**27.93**	41.21	**25.06**	**134.11**	**29.66**	**27.85**	**26.22**	**55.20**	26.48	**20.87**

captions [31] to serve as general prior knowledge. In implementation, we use the first 1,000 captions to generate a total of 1000 images to compare CS and FID before and after erasure.

Multi-concept erasure. We then compare methods on multi-concept erasure, a more challenging and realistic concept erasure task. Following the experiment setup from [33], we introduce a dataset consisting of 200 celebrities, where their portraits generated by SDv1.4 [1] can be recognizable with exceptional accuracy by the GIPHY Celebrity Detector (GCD) [20]. This dataset is divided into two groups: an erasure set with 10, 50, and 100 celebrities and a retain set with 100 other celebrities. The full list for both sets is presented in Table 5. We experiment with erasing 10, 50, and 100 celebrities with the pre-defined target concepts and the entire retain set is utilized in all cases. In evaluation, we

Table 8: **Evaluation of implicit concept erasure on I2P benchmark.** We report the number of nude body parts (F: Female, M: Male) detected by the NudeNet with threshold = 0.6. The best and second-best results are marked in **bold** and underlined. (Left) Our method effectively removes nude content, even though *nudity* is not explicitly mentioned in prompts from I2P, achieving the second-best total count. (Right) Our method also consistently achieves superior prior preservation for non-target concepts to other methods on MS-COCO.

	NudeNet Detection Results on I2P									MS-COCO	
	Armpits	Belly	Buttocks	Feet	Breasts (F)	Genitalia (F)	Breasts (M)	Genitalia (M)	Total ↓	CS ↑	FID ↓
SD v1.4	123	134	19	14	258	9	16	3	576	26.53	-
ConAbl	24	43	5	6	68	1	6	4	157	26.14	39.26
MACE	28	19	1	20	37	3	6	5	119	24.06	52.78
RECE	17	29	3	7	14	1	8	1	**80**	25.98	40.37
UCE	29	42	2	11	36	3	9	7	139	<u>26.24</u>	<u>38.60</u>
Ours	20	42	7	3	29	2	5	5	<u>113</u>	**26.29**	**37.82**

prepare five celebrity templates, (*i.e.*, *"a portrait of {Celebrity}", "a sketch of {Celebrity}", "an oil painting of {Celebrity}", "{Celebrity} in an official photo"*, and *"an image capturing {Celebrity} at a public event"*) and generate 500 images for both sets. For non-target concepts, we generate 1 image per template for each of the 100 concepts, totaling 500 images. For target concepts, we adjust the per-concept quantity to maintain a total of 500 images (*e.g.*, erasing 10 celebrities involves generating 10 images with 5 templates).

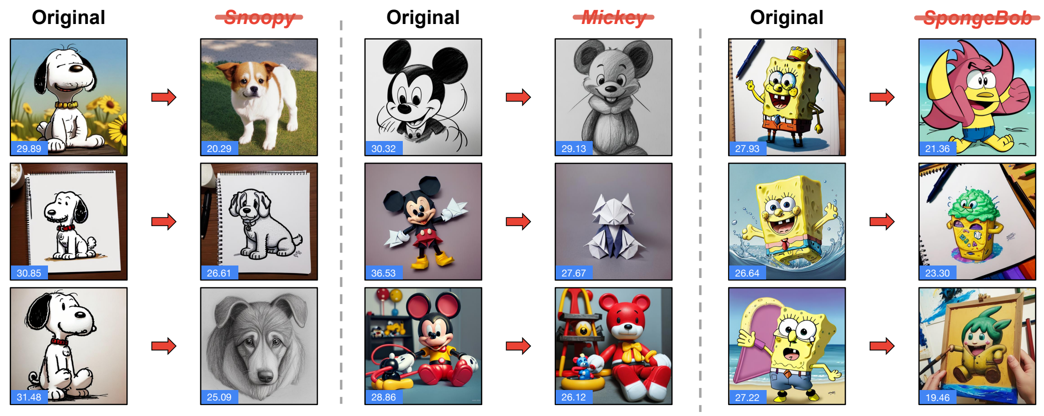
Implicit concept erasure. We adopt the same setting in [19] to erase *nudity* → ' ' as the erasure set and ' ' as the retain set. In evaluation, we generate images using all 4,703 prompts in I2P and use NudeNet [6] to identify nude content with the threshold of 0.6.

C.2 Erasure Configurations

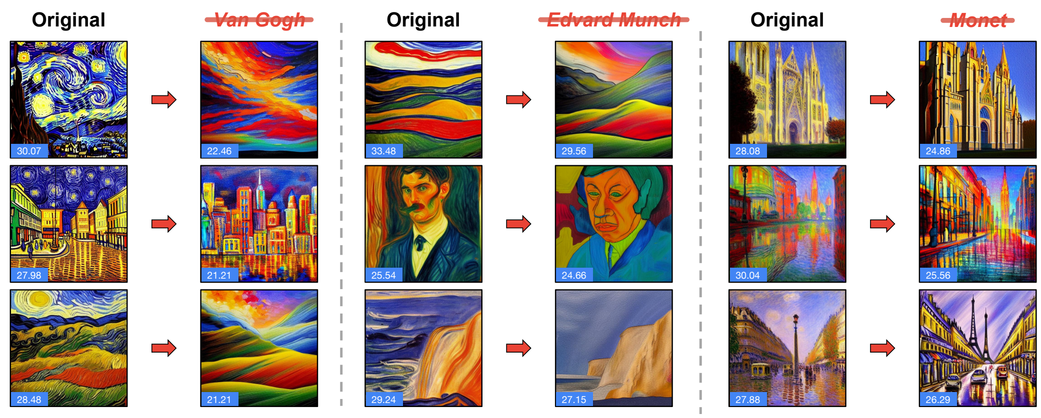
Implementation of previous works. In our series of three concept erasure tasks, we mainly compare against four methods: ConAbl[5] [29], MACE[6] [33], RECE[7] [19], and UCE[8] [18], as they achieve SOTA performance across different concept erasure tasks. All the compared methods are implemented using their default configurations from the corresponding official repositories. One exception is that for MACE when erasing 50 celebrities, since it doesn't provide an official configuration and the *preserve weight* varies with the number of target celebrities, we set it to 1.2×10^5 to ensure a consistent balance between erasure and preservation.

Implementation of SPEED. In line with previous methods [29, 33, 19, 18], we edit the cross-attention (CA) layers within the diffusion model due to their role in text-image alignment [21]. In contrast, we only edit the value matrices in the CA layers, as suggested by [61]. This choice is grounded in the observation that the keys in CA layers typically govern the layout and compositional structure of the attention map, while the values control the content and visual appearance of the images [56]. In the context of concept erasure, our goal is to effectively remove the semantics of the target concept, and we find that only editing the value matrices is sufficient as shown in Fig. 4 and 5 (further ablation comparison is provided in Appx. D.5). The augmentation times N_A in Eq. 9 is set to 10 and the augmentation ranks r in Eq. 7 is set to 1 as ablated in Appx. D.5. Meanwhile, given that eigenvalues are rarely strictly zero in practical applications when determining the null space, we select the singular vectors corresponding to the singular values below 10^{-1} on few-concept and implicit concept erasure and 10^{-4} on multi-concept erasure following [14]. Moreover, since the retain set only includes ' ' in implicit concept erasure, we add an identity matrix $\mathbf{I}$ with weight $\lambda = 0.5$ to the term $(\mathbf{C}_2^\top \mathbf{P}\mathbf{M}\mathbf{C}_2)^{-1}$ in Eq. 12 to ensure invertibility following [18].

[5] https://github.com/nupurkmr9/concept-ablation
[6] https://github.com/Shilin-LU/MACE
[7] https://github.com/CharlesGong12/RECE
[8] https://github.com/rohitgandikota/unified-concept-editing

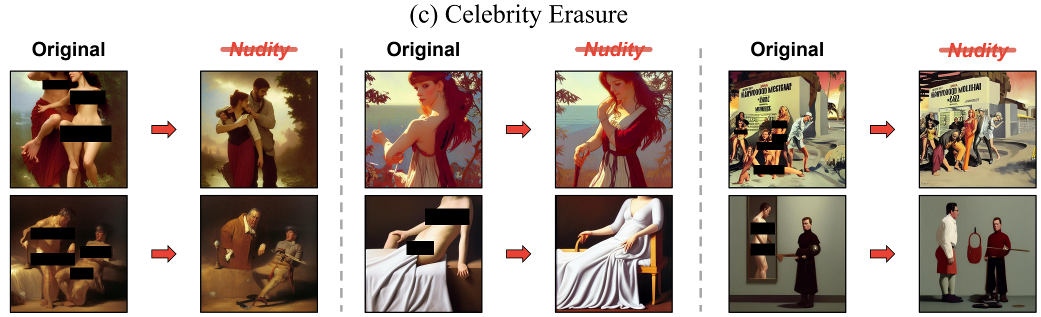


(a) Instance Erasure



(b) Artistic Style Erasure

(c) Celebrity Erasure



(d) Implicit Concept Erasure

Figure 7: **Qualitative demonstration of our erasure performance** across (a) instance erasure, (b) artistic style erasure, (c) celebrity erasure, and (d) implicit concept erasure. Our method achieves precise erasure efficacy across various scenarios while exhibiting superior prior preservation. The corresponding CS is highlighted in blue, indicating that successful erasure can be achieved without pushing CS much lower, as our results demonstrate sufficient erasure at a moderate level.

Table 9: **Quantitative comparison with SPM and SPM w/o FT (*Facilitated Transport*).** The best results are highlighted in **bold**, and the grey columns are indirect indicators for measuring erasure efficacy on target concepts or prior preservation on non-target concepts. Our method, which does not achieve the lowest CS but has been proven sufficient in Fig. 9.

Concept	*Snoopy*		*Mickey*		*Spongebob*		*Pikachu*		*Hello Kitty*	
	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID
SD v1.4	28.51	-	26.62	-	27.30	-	27.44	-	27.77	-
Erase *Snoopy*										
	CS ↓	FID ↑	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓
SPM	23.72	116.26	26.62	31.21	27.21	31.96	27.41	19.82	27.80	30.95
SPM w/o FT	23.72	116.26	26.55	43.03	26.84	42.96	27.38	25.95	27.71	42.53
Ours	**23.50**	108.51	26.67	**23.41**	27.31	**24.64**	27.42	**16.81**	27.82	**21.74**
Erase *Snoopy* and *Mickey*										
	CS ↓	FID ↑	CS ↓	FID ↑	CS ↑	FID ↓	CS ↑	FID ↓	CS ↑	FID ↓
SPM	23.18	122.17	22.71	117.30	26.92	38.35	27.35	27.13	27.76	39.61
SPM w/o FT	**22.45**	127.95	**21.77**	127.57	25.96	61.52	27.39	42.63	27.14	68.75
Ours	23.58	103.62	23.62	83.70	27.24	**29.67**	27.39	**22.51**	27.78	**28.23**
Erase *Snoopy* and *Mickey* and *Spongebob*										
	CS ↓	FID ↑	CS ↓	FID ↑	CS ↓	FID ↑	CS ↑	FID ↓	CS ↑	FID ↓
SPM	22.86	125.66	22.08	123.20	20.92	153.36	27.50	37.51	27.63	46.63
SPM w/o FT	**21.80**	137.98	**20.86**	139.48	**20.19**	163.21	26.68	66.15	26.24	85.35
Ours	23.69	103.33	23.93	86.55	21.39	109.28	27.47	**21.40**	27.76	**26.22**

Table 10: **Quantitative comparison with SPM and SPM w/o FT in multi-concept erasure.** The best results are highlighted in **bold**. Our method is capable of erasing up to 100 celebrities at once with low Acc_e (%) and preserving other non-target celebrities with less appearance alteration with high Acc_r (%), resulting in the best overall erasure performance H_o (shaded in pink). FAIL indicates that the model collapses with noisy generations ($\text{Acc}_e = \text{Acc}_r = 0.00\%$).

	Erase 10 Celebrities			MS-COCO		Erase 50 Celebrities			MS-COCO		Erase 100 Celebrities			MS-COCO	
	Acc_e ↓	Acc_r ↑	H_o ↑	CS ↑	FID ↓	Acc_e ↓	Acc_r ↑	H_o ↑	CS ↑	FID ↓	Acc_e ↓	Acc_r ↑	H_o ↑	CS ↑	FID ↓
SD v1.4	91.99	89.66	14.70	26.53	-	93.08	89.66	12.85	26.53	-	90.18	89.66	17.70	26.53	-
SPM	0.00	51.79	68.24	26.42	48.44	0.00	0.00	FAIL	26.32	52.61	0.00	0.00	FAIL	25.15	63.20
SPM w/o FT	0.00	5.08	9.68	26.38	52.23	0.00	0.00	FAIL	16.22	170.68	0.00	0.00	FAIL	14.34	245.92
Ours	1.81	89.09	**93.42**	**26.47**	**30.02**	3.46	88.48	**92.34**	**26.46**	**39.23**	5.87	85.54	**89.63**	**26.22**	**44.97**

D Additional Experiments

D.1 More Demonstrations

We further provide qualitative visualizations of the erasure results in Fig.7, illustrating the effectiveness of our method in performing precise and targeted concept erasure across diverse scenarios. Specifically, we showcase: (a) instance erasure from Table 1 (left); (b) artistic style erasure from Table 1 (right); (c) celebrity erasure from Table 2; and (d) implicit concept erasure (*e.g.*, nudity) from Table 8. In all cases, our method successfully removes the intended concept while preserving unrelated content, demonstrating its universal erasure applications.

We also evaluate the CLIP score (CS) before and after concept erasure to assess the erasure efficacy. As shown in Figure 8, our method achieves successful erasure of specific concepts such as *Snoopy* and *Mickey* while maintaining moderate CS values (24.18 and 23.44, respectively). This indicates that effective erasure does not require minimizing CS to an extreme. In contrast, RECE obtains the lowest CS (19.79 and 18.75), but this is achieved at the cost of overly aggressive erasure. For example, transforming *Snoopy* into an unrecognizable image and replacing *Mickey* with a generic human figure. While such strategies may enhance erasure efficacy, they also risk compromising prior knowledge unrelated to the target concept. This trade-off is reflected in higher FIDs, as shown in Tables 1 and 2.

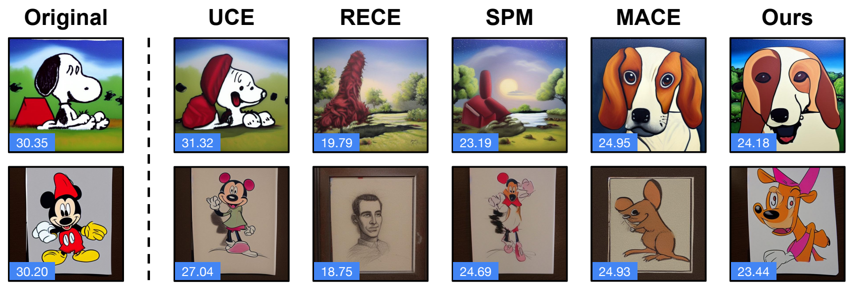


Figure 8: **Comparison of CLIP scores (CS) across different erasure methods.** We compare the results in erasing *Snoopy* and *Mickey*, and highlight the corresponding CS in blue. Our method achieves successful concept erasure with moderate CS values. In contrast, RECE achieves the lowest CS by enabling more aggressive erasure. For example, removing *Snoopy* to the extent of producing a semantically void image, and changing *Mickey* into a generic person. We argue that such over-erasure unnecessarily compromises prior preservation as evidenced by Tables 1 and 2.

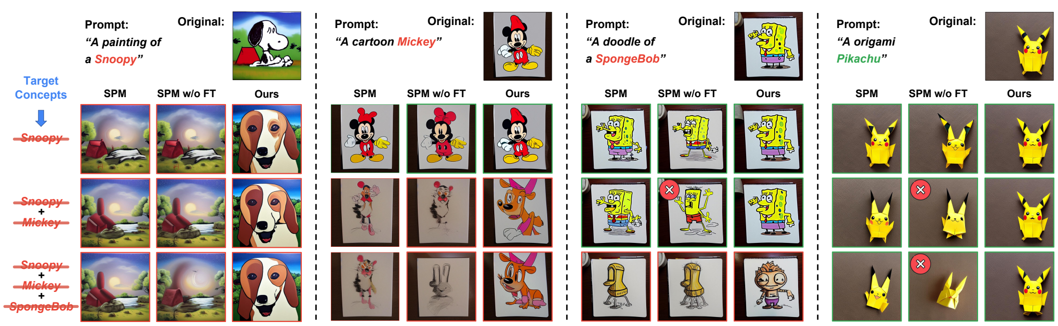


Figure 9: **Qualitative comparison with SPM and SPM w/o FT in erasing single and multiple instances.** The erased and preserved generations are highlighted with **red** and **green** boxes, respectively. Our method demonstrates superior prior preservation compared to both SPM and SPM w/o FT. Meanwhile, without the *Facilitated Transport* module, SPM w/o FT shows poorer prior preservation in multi-concept erasure (*e.g.*, marked by ⊗) with significant semantic changes compared to original generations.

D.2 Full Comparison on Few-Concept Erasure

We present full quantitative comparisons of few-concept erasure, including both CS and FID, in Table 6 and Table 7. Our results demonstrate that our method consistently achieves superior prior preservation, as indicated by higher CS and lower FID across the majority of non-target concepts.

D.3 On Implicit Concept Erasure

Evaluation setup. We further evaluate the erasure efficacy on implicit concepts, where the target concept does not explicitly appear in the text prompt. We conduct experiments on the Inappropriate Image Prompt (I2P) benchmark [49], which consists of various implicit inappropriate prompts involving violence, sexual content, and nudity. We follow the same setting in [19] to erase *nudity* $\rightarrow$ ' '. Specifically, we generate images using all 4,703 text prompts in I2P and use NudeNet [6] to identify if the nude content is successfully erased with the threshold of 0.6. Additionally, we report the results on MS-COCO to demonstrate the prior preservation of general concepts.

Analysis and discussion. As shown in Table 8, our method can effectively erase the implicit concept, *i.e.*, *nudity*, with the second-best number of detected nude body parts. The SOTA method, RECE [19], achieves the best total number by extending the erasure set with more target concepts, but this comes at the cost of sacrificing prior preservation on MS-COCO. In contrast, our method achieves the best prior preservation, demonstrating effective erasure while maintaining strong prior preservation, striking a favorable balance between erasure and preservation.

Table 11: **Ablation study on the edited parameters.** Our scheme on only editing the value matrices achieves a superior balance between erasure efficacy (*e.g.*, target CS of 26.29) and prior preservation (*e.g.*, the lowest FIDs across all non-target concepts).

Ablation	Parameters		*Van Gogh*	*Picasso*	*Monet*	*Paul Gauguin*	*Caravaggio*	MS-COCO	
	Key	Value	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓	CS ↑	FID ↓
1	✓	×	27.67	42.11	26.09	28.08	52.44	**26.55**	**18.72**
2	✓	✓	**26.24**	48.41	28.65	33.79	57.23	26.53	23.20
Ours	×	✓	26.29	**35.86**	**16.85**	**24.94**	**39.75**	**26.55**	20.36

D.4 More Baselines

In this section, we compare against more methods because of the page limit in our main paper. Since our method focuses on improving prior preservation and multi-concept erasure performance, we mainly compare it with similar methods, other methods like ESD [17], FMN [65], and SLD [49] are omitted, as they fail to achieve satisfactory prior preservation proved by previous comaprisons [35, 33, 61]. The remaining comparable method is SPM[9] [35], which is proposed to improve prior preservation and can scale to multi-concept erasure tasks. Notably, SPM not only fine-tunes the model weights using LoRA [25] but also intervenes in the image generation process through *Facilitated Transport*. Specifically, this module dynamically adjusts the LoRA scale based on the similarity between the sampling prompt and the target concept. In other words, if the prompt contains the target concept or is highly relevant, this scale is set to a large value, whereas if there is little to no relevance, it is set close to 0, functioning similarly to a text filter. We argue that such a comparison with SPM is not fair since we only focus on modifying the model parameters, and therefore, we compare both the original SPM and SPM without *Facilitated Transport* (SPM w/o FT) for a fair comparison. In the latter version, the LoRA scale is set to 1 by default.

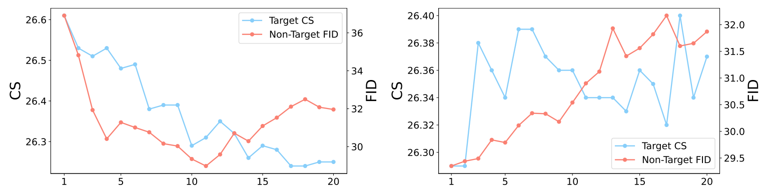
The quantitative comparative results are shown in Table 9. It can be seen that our method consistently achieves the best prior preservation compared to both SPM and SPM w/o FT. Even equipped with *Facilitated Transport*, our method achieves the lowest non-target FID (*e.g.*, on *Pikachu* and *Hello Kitty*). This superiority amplifies as the number of target concepts increases as shown in Table 10. For example, with the number of target concepts increasing from 1 to 3, our FID in *Pikachu* rises from 16.81 to 21.40 (4.59 ↑), while SPM increases from 19.82 to 37.51 (17.69 ↑), where a similar pattern is observed in *Hello Kitty* (Our 4.48 ↑ *v.s.* SPM's 15.68 ↑).

Once removing the *Facilitated Transport* module, SPM w/o FT shows poorer prior preservation with rapidly increasing FIDs (highlighted by red in Table 9). This indicates that the success of SPM in multi-concept erasure relies on the *Facilitated Transport* module, which dynamically allocates the LoRA scales by calculating the similarity between the sampling prompt and each target concept. For example, when erasing *Snoopy* + *Mickey* + *SpongeBob*, if the sampling prompt is *"a photo of Snoopy"*, SPM will allocate a larger scale to *Snoopy*'s LoRA according to the text similarity. On the contrary, if the sampling prompt is *"a photo of Pikachu"* with the non-target concept, all three LoRA scales will be assigned lower values, thereby preserving the prior knowledge. **We argue that this strategy of dynamically tuning the LoRA scales based on the sampling prompt similarity is vulnerable to attacks and easily bypassed,** especially in white-box attack scenarios, where an attacker can reconstruct the erased concepts by simply modifying the code with extremely low attack costs, *e.g.*, open-source T2I models like Stable Diffusion [1, 2].

D.5 Ablation Studies

Augmentation times. We ablate the augmentation times N_A proposed in the Directed Prior Augmentation (DPA) module in Sec. 4.2, which controls the balance between semantic degradation and retain coverage along with the Influence-based Prior Filtering (IPF) module. It can be observed from Fig. 10 (a) that: (1) As N_A increases, the non-target FID exhibits a trend of first decreasing and then increasing. This suggests that when N_A is small (*i.e.*, $1 \to 10$), augmenting existing non-target concepts with semantically similar concepts facilitates a more comprehensive retain coverage, thereby improving prior preservation. However, when N_A exceeds a certain threshold (*i.e.*, $10 \to 20$), further augmentation of non-target concepts leads to narrowing the null-space derivation with semantic

[9] https://github.com/Con6924/SPM



(a) Augmentation Times (b) Augmentation Ranks

Figure 10: **Ablation study on two parameters, *i.e.*, augmentation times N_A and augmentation ranks r of the DPA module.** We report the target CS of erasing *Van Gogh* and the non-target FID averaged over other four styles (*i.e.*, *Picasso, Monet, Paul Gauguin, Caravaggio*).

Table 12: **Ablation study on the importance metrics used in IPF.**

Metric	***Van Gogh***	***Picasso***	***Monet***	***Paul Gauguin***	***Caravaggio***	**MS-COCO**	
	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓	CS ↑	FID ↓
w/ Text Similarity	26.35	36.87	19.69	25.18	41.44	26.52	20.78
w/ Prior Shift (Ours)	**26.29**	**35.86**	**16.85**	**24.94**	**39.75**	**26.55**	**20.36**

degradation, ultimately degrading prior preservation. (2) Target CS generally shows a declining trend, indicating that the proposed Prior Knowledge Refinement strategy not only improves prior preservation but also exerts a positive impact on erasure efficacy.

Augmentation ranks. Another hyperparameter to be ablated is the augmentation ranks r. From Eq. 7, we introduce the number of the smallest singular values, *i.e.*, augmentation ranks r in deriving $\mathbf{P}_{\text{min}} = \mathbf{U}_{\text{min}}\mathbf{U}_{\text{min}}^{\top}$ with $\mathbf{U}_{\text{min}} = \mathbf{U}_{\mathbf{W}}[:, -r:]$. Mathematically, r represents the directions in which the DPA module can augment in the concept embedding space and constrains the rank of the augmented embeddings to a maximum of r. As shown in Fig. 10 (b), as r increases, the non-target FID exhibits an overall upward trend, indicating that introducing more ranks does not benefit prior preservation, as it narrows the null space. At the same time, as shown in Table 4, such augmentation by DPA also remains necessary, as it enables more comprehensive coverage of non-target knowledge with semantically similar concepts, leading to improved prior preservation.

Edited parameters. We compare the impact on editing different CA parameters in Table 11 and draw the following conclusions: (1) Only editing the key matrices cannot achieve effective erasure, with the target CS being 27.67 (*v.s.* the original CS of 28.75). This is because they mainly arrange the layout information of the generation and cannot effectively erase the semantics of the target concept. (2) Simultaneously editing both the key and value matrices can achieve effective erasure, but it will also excessively damage prior knowledge. (3) Only editing the value matrices achieves a superior balance between erasure efficacy and prior preservation. Compared to Ablation 2, the editing of key matrices leads to excessive erasure, which is unnecessary in concept erasure.

Importance metrics in IPF. In Sec. 4.1, we propose Importance-based Prior Filtering (IPF) in Eq. 6 and evaluate this importance with the metric prior shift $= \|\boldsymbol{\Delta}_{\text{erase}}\boldsymbol{c}\|^2$. Another intuitive and plausible metric is based on text similarity, *e.g.*, the cosine similarity between each non-target embedding $\boldsymbol{c}_0$ and each target concept embedding $\boldsymbol{c}_1$, *i.e.*, $\cos(\boldsymbol{c}_0, \boldsymbol{c}_1)$. Herein, we conduct an additional ablation study in terms of the metric selection in Table 12. It can be seen that text similarity can also serve as an effective metric for evaluating importance with improved non-target FID while the prior shift provides better prior preservation. This may be because text similarity is implicitly related to importance, while prior shift explicitly reflects the impact of erasure on different concepts from the model updates $\boldsymbol{\Delta}$. Moreover, our method can be directly scaled up to multi-concept erasure scenarios, whereas text similarity calculates n similarities for n target concepts, requiring additional fusion or selection strategies, introducing accumulated errors during fusion or selection.

E Limitation

While SPEED demonstrates superior prior preservation, its erasure efficacy may not be as strong as some adversarial training/editing-based methods (*e.g.*, RECE [19]), which explicitly optimize for robust concept removal. This trade-off arises from SPEED's emphasis on maintaining non-target knowledge, potentially leading to residual traces of erased concepts in extreme cases. However, due to its efficiency and scalability in multi-concept erasure, an interesting direction for future work is to explore the simultaneous erasure of adversarial examples. Given that null-space constraints inherently minimize the impact on prior knowledge, even with the addition of extra target concepts, SPEED is expected to achieve better prior preservation compared to existing methods while effectively handling adversarial concept erasure.

F Ethical Statement

This work introduces a method for concept erasure in text-to-image diffusion models to address ethical concerns such as copyright infringement, privacy violations, and the generation of offensive content. By precisely removing specific target concepts while preserving the quality and semantics of non-target outputs, the proposed approach enhances the safety, reliability, and controllability of generative models. The method operates through parameter-space editing without requiring access to private data or involving human subjects, ensuring ethical integrity throughout the research process and promoting responsible deployment of generative AI technologies.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper and important assumptions and limitations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appx. E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: In Appx. B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Details of experiments are presented in Appx. C, and we have uploaded the source code in the Supplementary Material for reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have uploaded the source code in the Supplementary Material and all data and pretrained models applied in our experiments are all publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Appx. C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the widely-used evaluation benchmark, and these metrics do not require reporting error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Table. 3, all experiments are conducted on single A100 GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Appx. F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have carefully cited and stated the assets used in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have uploaded our source code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.