EMPTY SPACE: CROSS-ATTENTION SPARSITY FOR CONCEPT ERASURE IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Erasing specific concepts from text-to-image diffusion models is essential for avoiding the generation of copyrighted and explicit content. Closed-form concept erasure methods offer a fast alternative to backpropagation-based techniques, but they become less effective when scaling from smaller models such as Stable Diffusion 1.5 to larger models like Stable Diffusion XL. To maintain erasure effectiveness in these larger-scale architectures, we propose SParse cross-Attention-based Concept Erasure (SPACE). SPACE iteratively modifies the cross-attention parameters of a model with a closed-form update that jointly induces sparsity and erases target concepts. By concentrating the concept mapping to a lower-dimensional subspace, SPACE achieves superior erasure efficacy compared to dense baselines. Extensive experimental results show the erasure effectiveness for different scenarios and a large set of models. Furthermore, SPACE achieves 80%-90% cross-attention sparsity, reducing the storage requirements for saving the modified parameters by 70%, demonstrating its memory efficiency.
WARNING: This paper contains model outputs that may be offensive.

1 INTRODUCTION

Text-to-Image (T2I) diffusion models (DMs) excel at generating high-fidelity images from natural language prompts (Nichol et al., 2021; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022). However, their reliance on massive unfiltered datasets (Schuhmann et al., 2022) can lead to the generation of unwanted content, including Not-Safe-For-Work (NSFW) and copyrighted material (Jiang et al., 2023; Schramowski et al., 2023). Re-training a DM after dataset curation is expensive and computationally prohibitive (Rombach, 2022). Meanwhile, post-generation processing (Schramowski et al., 2023) and inference-guiding (Yoon et al., 2025) can be circumvented when the model weights are accessible. Therefore, concept erasure via weight modification has emerged as a robust intervention.

Most erasure methods fine-tune DMs to align the model generation corresponding to the concept to be erased (i.e., the target) with the generation induced by a "guide" (also referred to as anchor) concept. This is traditionally achieved by minimizing a divergence measure between the conditioned model distributions (Gandikota et al., 2023; Kumari et al., 2023; Bui et al., 2025; Thakral et al., 2025; Novello et al., 2026). While effective, divergence-based methods rely on backpropagation, which is computationally heavy. To address this limitation, methods that update the model cross-attention weights in closed-form have been proposed (Gandikota et al., 2024; Gong et al., 2024; Biswas et al.; Li et al., 2026), resulting in fast and accurate erasure.

While these techniques performing closed-form updates excel on smaller models like Stable Diffusion 1.5 (SD1.5) (Rombach et al., 2022), we typically observe a significant performance decrease when scaling to larger architectures such as Stable Diffusion XL (SDXL) (Podell et al., 2023), particularly in the erasure of artistic styles and nudity. We attribute this to the large increase in the number of cross-attention parameters (from 19M in SD1.5 to 340M in SDXL for Keys and Values). In this high-dimensional regime, the inherent redundancy of the parameter space (Aghajanyan et al., 2021; Li et al., 2018) may compromise effective erasure, as concept-specific information is distributed across large weight matrices.

Our key insight is that effective closed-form erasure can be achieved by inducing parameter sparsity within the cross-attention space. While it is known that sparsification can improve model performance

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

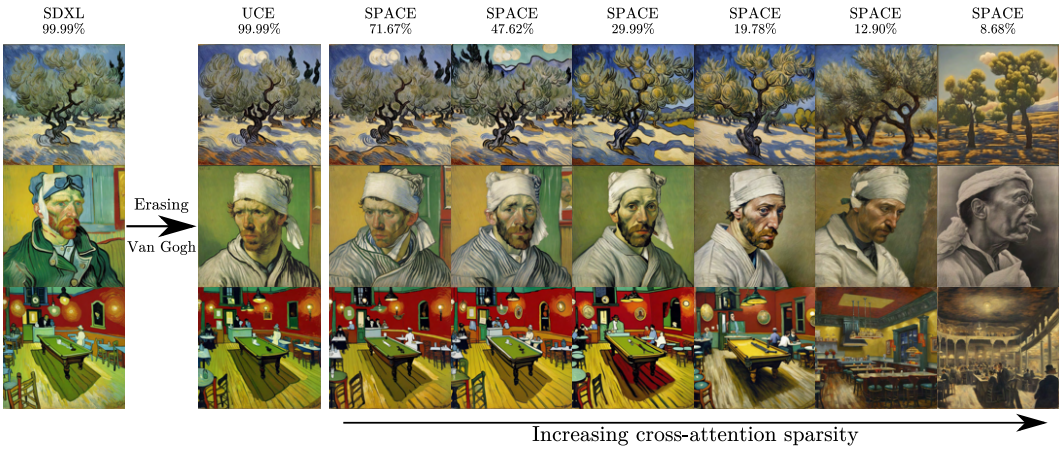


Figure 1: SPACE erases concepts while inducing cross-attention sparsity of a text-to-image diffusion model with a closed-form iterative update of the parameters. Increasing values of sparsity (from left to right) lead to a stronger erasure. The percentages reported under the methods names indicate the ratio of non-zero parameters in the cross-attention of the model.

due to the reduction of learned noise (Hoefler et al., 2021), we show that by constraining the mapping of a concept to a lower-dimensional subspace of high-impact parameters, we remove such a concept more effectively. To achieve this goal, we propose SParse cross-Attention-based Concept Erasure (SPACE). To maintain the efficiency of gradient-free methods while handling the non-smoothness of L_1 regularization, SPACE introduces an iterative procedure, based on the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Beck & Teboulle, 2009), that updates the model cross-attention parameters in closed-form. Since SPACE builds upon UCE (Gandikota et al., 2024) by integrating a sparsity constraint, it achieves accurate erasure while inducing a high degree of parameter sparsity without the need for backpropagation, which is instead used for the other sparsity-based unlearning techniques (Jia et al., 2023; Fan et al., 2024; Wu & Harandi, 2024).

Beyond improving erasure efficacy (see one example in Fig. 1), SPACE offers significant structural advantages. By inducing 80%-90% sparsity in the cross-attention weights of XL-scale models, SPACE reduces the storage footprint of the modified parameters by approximately 70%. This characteristic makes SPACE suited for scalable model personalization and efficient edge deployment. For instance, in environments where multiple users require distinct sets of concepts to be erased, SPACE allows for the rapid swapping of lightweight cross-attention modules.

2 Related Work

Concept erasure in diffusion models Concept erasure aims to remove specific knowledge from DMs. Divergence-based methods (Gandikota et al., 2023; Kumari et al., 2023; Bui et al., 2025; Thakral et al., 2025; Novello et al., 2026; Wu et al., 2025; Bui et al., 2024; Maharana et al.) typically align the model distribution conditioned on a target concept with its generation conditioned on an anchor concept. The main drawback of these methods is their high computational cost, as they rely on backpropagation to fine-tune the model. As a more computationally- and memory-efficient alternative, methods that update the model weights in closed-form have been presented. UCE (Gandikota et al., 2024), which builds upon (Orgad et al., 2023) and (Meng et al., 2022), introduces a unified framework for concept editing by performing a closed-form update of the cross-attention parameters of T2I models. RECE (Gong et al., 2024) is an iterative method that extends UCE by proposing a robust unlearning framework that generates prompts capable of re-activating erased concepts and modifies the T2I model based on such prompts. SPEED (Li et al., 2026) utilizes null-space projections to minimize interference with non-target concepts. Current closed-form methods operate in a dense parameter space.

Sparsity in attention and computer vision Sparse attention has been extensively studied in natural language processing (Malaviya et al., 2018; Peters et al., 2019; Correia et al., 2019; Zhang et al.,

2021). However, its application within computer vision has primarily focused on developing sparsified activations through Softmax alternatives, such as α-Entmax or ReLU-based variants (Martins & Astudillo, 2016; Niculae & Blondel, 2017; Kim & Sim, 2025). These methods aim to improve inference efficiency or noise robustness by sparsifying the attention activation maps. SPACE differentiates itself by targeting parameter sparsity rather than activation sparsity.

Furthermore, while some computer vision architectures employ parameter-level attention sparsity (Zhang et al., 2025a;b), they use standard backpropagation. SPACE is distinct as it achieves the sparsity of the cross-attention parameters without relying on backpropagation.

Unlike the methods analyzed in this paragraph, SPACE focuses on concept erasure.

Sparsity-based machine unlearning The intersection of sparsity and machine unlearning is an emerging topic. Some recent approaches have explored sparse autoencoders (SAEs) for localized concept erasure (Cywiński & Deja, 2025; Cassano et al., 2025; He et al., 2025; Geng & Li, 2025). However, SAE-based methods, similarly to LoRA (Hu et al., 2022; Lu et al., 2024) or adapter-based approaches (Lyu et al., 2024; Xiong et al., 2024; Lee et al., 2025), require storing additional weights alongside the original model, introducing storage overhead and remaining vulnerable to circumvention if the added modules are detached. In contrast, SPACE directly modifies the original model parameters.

Weight-editing unlearning methods are often motivated by the Lottery Ticket Hypothesis (Frankle & Carbin, 2018), which suggests that dense networks contain sparse sub-networks capable of matching or exceeding the performance of the original model. While prior weight-editing unlearning methods employ L_1 penalties (Jia et al., 2023), weight saliency (Fan et al., 2024; Wu & Harandi, 2024; Wang et al., 2026), or pruning (Chavhan et al., 2025; Shirkavand et al., 2025), they rely on backpropagation and do not specifically target cross-attention. SPACE is the first concept erasing framework to integrate parameter sparsity into a closed-form update procedure, combining computational efficiency with the structural benefits of high-sparsity representations.

3 METHOD

3.1 PRELIMINARIES

3.1.1 DIFFUSION MODELS

Diffusion Models (Sohl-Dickstein et al., 2015; Ho et al., 2020) are state-of-the-art generative models that approximate distributions by first gradually adding Gaussian noise to a signal and then learning the reverse Markov process to generate from Gaussian noise. In the forward process, Gaussian noise (referred to as ϵ) is gradually added to an input image $\mathbf{x}_0$ over multiple steps $t \in [0, \dots, T]$, to obtain $\mathbf{x}_T \sim \mathcal{N}(0, I)$. At each time step, the noisy image is referred to as $\mathbf{x}_t$. In the reverse process, $\mathbf{x}_T$ is transformed, following a transformation that is the inverse of the forward process, to obtain a denoised image using a denoising network $\Phi(\mathbf{x}_t, \mathbf{c}, t)$, where, for T2I models, the concept $\mathbf{c}$ is a text prompt. The denoising process is characterized as $p_\Phi(\mathbf{x}_0, \dots, \mathbf{x}_T|\mathbf{c}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\Phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$, where $p_\Phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$ describes the probability of $\mathbf{x}_{t-1}$ given the noisy image $\mathbf{x}_t$ and the concept $\mathbf{c}$.

3.1.2 CROSS-ATTENTION

Cross attention is a fundamental building block of many state-of-the-art T2I diffusion models, as it incorporates text conditioning into the image generation process. Specifically, a text encoder is fed with a text prompt and generates a text embedding which is integrated into the image generation process through a Query-Key-Value (QKV) structure (Vaswani et al., 2017). Given a text embedding c_i, the Keys and Values are obtained as $k_i = W_k c_i$ and $v_i = W_v c_i$, respectively, and are responsible for projecting text embeddings. The cross-attention output is computed as

$$\mathcal{O} \propto \text{softmax}(q_i k_i^T) v_i, \tag{1}$$

where q_i represents the visual features, and $\mathcal{A} \propto \text{softmax}(q_i k_i^T)$ is the attention map aligning relevant text and image regions.

3.1.3 UNIFIED CONCEPT EDITING

Let $c_i \in E$ indicate the target concept, $c_i^* \in G$ denote the destination (or guide) concept, and $c_j \in P$ represent the concept to be preserved. Let W^o and W be the K/V projection matrices before and after the editing, respectively. UCE (Gandikota et al., 2024) formulates the loss function as

$$\min_W \mathcal{L}(W) = \min_W \sum_{c_i \in E} ||Wc_i - v_i^*||_2^2 + \lambda_1 \sum_{c_j \in P} ||Wc_j - v_j||_2^2 + \lambda_2 ||W - W^o||_F^2, \tag{2}$$

where $v_i^* = W^o c_i^*$, $v_j = W^o c_j$. The first term is responsible for the target concepts erasure, while the second and third terms preserve all other concepts. More in detail, the first term in equation 2 edits the concept c_i by substituting it with the concept c_i^*. The second term in equation 2 serves as a preservation constraint for the concepts in P, enforcing the edited model to retain its pre-erasure performance. The third term in equation 2 also helps with concepts preservation while guaranteeing the invertibility of the inverse term in the closed-form solution

$$W = W^o \left(\sum_{c_i \in E} c_i^* c_i^T + \lambda_1 \sum_{c_j \in P} c_j c_j^T + \lambda_2 I \right) \left(\sum_{c_i \in E} c_i c_i^T + \lambda_1 \sum_{c_j \in P} c_j c_j^T + \lambda_2 I \right)^{-1}. \tag{3}$$

3.2 SPARSE CROSS-ATTENTION-BASED CONCEPT ERASURE (SPACE)

The objective of this method is to erase a set of concepts while inducing sparsity in the cross-attention parameters of a DM. Unlike prior work on machine unlearning that relies on L_1 regularized losses minimized through backpropagation (e.g., (Jia et al., 2023)), our approach aims to induce sparsity while performing a closed-form update of the weights. Formally, our goal can be framed as

$$\min_W \mathcal{L}(W), \qquad \text{s.t.} \quad ||W||_0 \leq R, \tag{4}$$

where $\mathcal{L}(W)$ is the loss defined in equation 2 and $|| \cdot ||_0$ represents the L_0 norm, which counts the number of non-zero elements, while R is a constant that enforces the desired sparsity level. Following the principles of sparsity-inducing optimization (Bach et al., 2011; Jia et al., 2023), we substitute the constraint on the L_0 norm with an L_1 norm embedded in the objective function. Thus, we formulate the loss function as

$$\min_W \mathcal{J}(W) = \min_W \mathcal{L}(W) + \lambda ||W||_{1,1}, \tag{5}$$

where $||W||_{1,1} = \sum_{i,j} |w_{i,j}|$ is the entry-wise L_1 norm, which coincides with the L_1 norm of the vectorized matrix W. First, we observe that equation 5 is a convex optimization problem. As such, it ensures convergence to the global optimum, whereas backpropagation-based erasure methods may only converge to local optima. Differently from equation 2, equation 5 does not allow for a one-shot closed-form solution to update W, due to the non-smoothness of the entry-wise L_1 norm. To solve equation 5, a simple choice would be to resort to gradient descent, which would still lead to a faster concept erasure compared to a loss formulated on the model output (typical of the existing divergence-based losses). However, given the nature of the problem, a more appropriate choice would be to use the Iterative Shrinkage-Thresholding Algorithm (ISTA) (Daubechies et al., 2004). This technique is suited for solving optimization problems that consist of a sum of a smooth convex function and a non-smooth convex function. To further decrease the computational cost, we propose to use Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Beck & Teboulle, 2009), which converges faster than ISTA, as it relies on momentum. FISTA is considered the best option to solve Lasso problems (Zhao & Huo, 2023) (like the one we define in equation 5), as it has faster convergence rate ($O(1/k^2)$, where k is the number of iterations) compared to many other methods, such as ISTA ($O(1/k)$), Coordinate Gradient Descent Algorithm (CGDA) ($O(1/k)$)(Friedman et al., 2010), and Smooth L1 Algorithm (SLA) ($O(1/k)$)(Schmidt et al., 2007). We now describe how we use FISTA to minimize the loss function in equation 5. Then, we present the whole pseudocode for SPACE in Algorithm 1 (Appendix A), while we provide a visual representation of the proposed framework in Fig. 2.

Let

$$\mathcal{T}_\alpha(x) = (|x| - \alpha)_+ \text{sgn}(x) \tag{6}$$

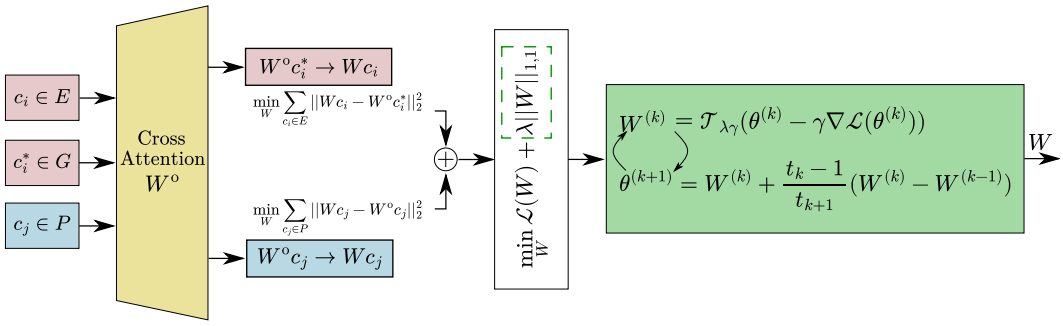


Figure 2: SPACE erases the concepts in E by substituting them with the guide concepts in G while preserving the concepts in P, with an iterative closed-form update of the cross-attention parameters, and returns a *sparse* matrix W.

be the shrinkage operator, where $(x)_+$ denotes the positive part of x (i.e., $(x)_+ = \max(x, 0)$). To optimize the loss function in equation 5 using FISTA, we update the cross-attention matrices during the k-th iteration of the algorithm as

$$W^{(k)} \leftarrow \mathcal{T}_{\lambda\gamma}(\theta^{(k)} - \gamma \nabla \mathcal{L}(\theta^{(k)})), \tag{7}$$

where $\gamma = 1/L$ with L being the Lipschitz constant of $\nabla \mathcal{L}(W)$ (see Theorem 3.1) and θ auxiliary variable. For simplicity, we express equation 2 in matrix form. Thus, we have $W \in \mathbb{R}^{n \times m}$, $C_e, C_g \in \mathbb{R}^{m \times n_E}$, and $C_p \in \mathbb{R}^{m \times n_P}$, where C_e, C_g, and C_p are the matrices whose columns are the embeddings for erasure (c_i), guidance (c_i^*), and preservation (c_j) concepts, and n_E, n_P are the number of concepts to erase and preserve, respectively. The gradient of $\mathcal{L}(W)$, used in equation 7, reads as

$$\nabla \mathcal{L}(W) = 2(WC_e - W^o C_g)C_e^T + 2\lambda_1(WC_p - W^o C_p)C_p^T + 2\lambda_2(W - W^o). \tag{8}$$

During the k-th iteration of the algorithm, θ is refined as

$$\theta^{(k+1)} \leftarrow W^{(k)} + \frac{t_k - 1}{t_{k+1}}(W^{(k)} - W^{(k-1)}), \tag{9}$$

where t_k provides momentum and is incremented as $t_{k+1} \leftarrow \frac{1+\sqrt{1+4t_k^2}}{2}$.

The convergence of SPACE is ensured by showing that $\nabla \mathcal{L}(W)$ is Lipschitz continuous (Beck & Teboulle, 2009) and by using the Lipschitz constant L within the update formula. Theorem 3.1 provides such a proof, along with the value of the Lipschitz constant.

Theorem 3.1. *Let $\mathcal{L}(W)$ be the smooth part of the SPACE loss function in equation 5. Let $\sigma_{max}(\cdot)$ denote the largest singular value. The gradient $\nabla \mathcal{L}(W)$ is Lipschitz continuous with constant L given by*

$$L = 2(\sigma_{max}(C_e C_e^T) + \lambda_1 \sigma_{max}(C_p C_p^T) + \lambda_2). \tag{10}$$

The proof of Theorem 3.1 is reported in Appendix B. The Lipschitz constant defined in Theorem 3.1 depends on the targets, anchors, and concepts to be preserved, which implies an adaptive step size based on the spectral properties of the concept embeddings.

Remark The first two terms in equation 2 define an under-determined erasure task due to the high dimensionality of the cross-attention parameter space. While the Frobenius norm regularizer used in UCE (Gandikota et al., 2024) ensures a unique solution, this minimum-norm approach tends to diffuse tiny updates across the entire matrix. Differently, SPACE adds an L_1 penalty, forcing the optimization to find a unique solution that balances the minimization of the Frobenius and L_1 norms. In this framework, the minimization of the L_1 norm encourages larger changes to fewer parameters while zeroing out the rest.

4 Results

In this section, we report the main experimental results. We defer to App. C for additional results.

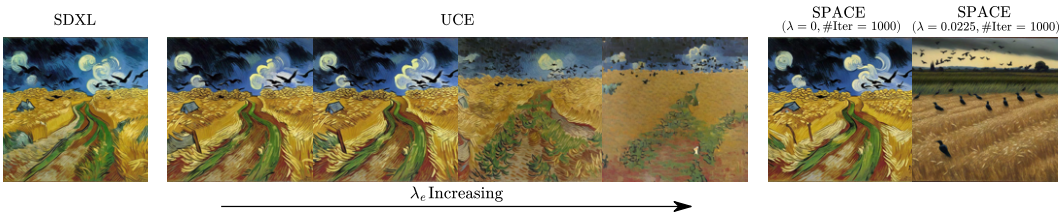
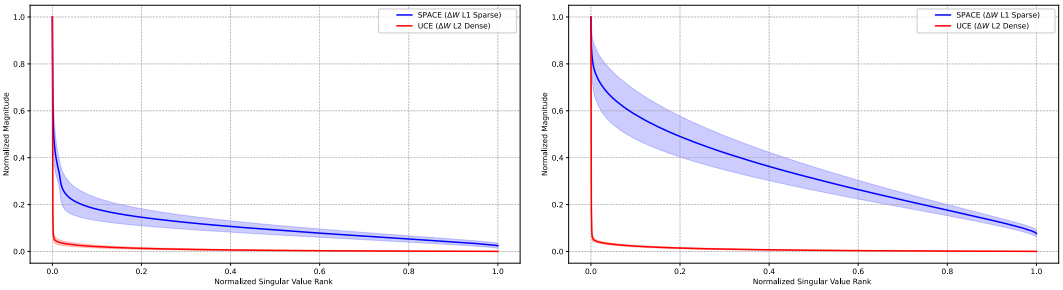


Figure 3: Erasure of Van Gogh style. The figure shows the importance of sparsity, as the effective erasure is neither achieved by increasing the erasure strength of UCE, nor because of the iterative nature of SPACE. **Left**: Original SDXL model. **Center**: Unlearned model using UCE: from left to right, we increase the erasure strength of UCE (handled by the parameter λ_e). **Right**: Unlearned model using SPACE, with 1000 iterations: without and with sparsity.



(a) K matrices.

(b) V matrices.

Figure 4: Erasing "Van Gogh" on SDXL: spectral analysis of the Δ_W.

4.1 Implementation Details

We measure how successfully concepts have been erased using CLIP Score (Hessel et al., 2021) (CS), CLIP Accuracy (Radford et al., 2021) (CA), and Kernel Inception Distance (Bińkowski et al., 2018) (KID). Lower CS and CA for the concept to erase imply better erasure. Higher CS and CA for the concepts to preserve indicate better knowledge preservation. The KID score is used to address the change in the generative distribution of the models. For non-target concepts, a smaller KID implies good preservation of the overall quality and coherence of the model. For target concepts, a larger KID refers to a more destructive erasure. Unless otherwise specified, we run our algorithm for 1000 iterations of closed-form updates. We set $\lambda_1 = 1$ and $\lambda_2 = 1$, while the value of λ is specified in each experiment. The experiments are performed on a GPU NVIDIA GeForce RTX 5090 with 32 GB. We provide in Appendix C the specific details of the different evaluation scenarios.

4.2 Sparsity Analysis

In this section, we study the importance of the cross-attention sparsity attained by SPACE. First, we show in the central part of Fig. 3 (an extended version can be checked in Fig. 9 in Appendix C.1.1) that, even when increasing the erasure strength of UCE via the trade-off parameters of its loss, its performance does not improve, and that forcing UCE to achieve actual erasure results in catastrophic degradation of the entire generation. On the contrary, SPACE achieves effective erasure due to inducing sparsity, as also shown in Fig. 1. Second, we demonstrate in the right side of Fig. 3 that SPACE effective erasure is due to inducing sparsity and not to the iterative nature of the algorithm. Indeed, for a given number of SPACE iterations, the erasure of Van Gogh is effective only when enforcing sparsity (i.e., $\lambda > 0$). A more extensive analysis is reported in Appendix C.1.2. Third, we extract the update matrices Δ_W across all layers, perform singular value decomposition (SVD), and plot mean $\pm$ standard deviation in Fig. 4. Both for K matrices (Fig. 4a) and V matrices (Fig. 4b), we observe that UCE performs a low-rank update, leaving the majority of the parameter directions unmodified. If the target concept were confined to these dominant directions, UCE's low-rank suppression would be sufficient to erase it. However, as our results demonstrate, UCE consistently fails to erase the concept. Therefore, the visual features that survive and appear in the

Table 1: Evaluation of style erasure on SDXL. **Bold** indicates the best erasure performance. Efficiency is evaluated in terms of VRAM (M), storage size of the unlearned set of parameters (S), and gain given by storage size reduction (G). The closed-form-based methods are highlighted in light blue.

Method	Van Gogh			Monet			Picasso			Watercolor			Hokusai			Warhol			Efficiency		
	CS	CA	KID	CS	CA	KID	CS	CA	KID	CS	CA	KID	CS	CA	KID	CS	CA	KID	M. (GB)	S. (MB)	G. (MB)
SDXL	33.81	100.0	–	33.67	100.0	–	34.06	100.0	–	36.20	100.0	–	32.49	100.0	–	35.53	98.00	–	–	–	–
Erasing **Van Gogh**																					
	CS↓	**CA↓**	**KID↑**	CS↑	CA↑	KID↓	CS↑	CA↑	KID↓	CS↑	CA↑	KID↓	CS↑	CA↑	KID↓	CS↑	CA↑	KID↓	**M.↓**	**S.↓**	**G.↓**
CAbl (Kumari et al., 2023)	31.29	76.00	0.051	31.40	88.00	0.006	31.49	82.00	0.026	35.64	96.00	0.004	31.79	100.0	0.004	34.53	92.00	0.001	19	630	0
SAFREE (Yoon et al., 2025)	31.81	78.00	0.018	31.56	86.00	0.008	31.69	98.00	0.018	34.27	88.00	0.004	31.38	100.0	0.021	33.98	94.00	0.001	11	**0**	0
UCE (Gandikota et al., 2024)	33.56	86.00	0.005	33.60	94.00	0.007	33.68	98.00	0.002	35.94	98.00	0.008	32.45	100.0	0.004	35.52	96.00	0.006	**8**	630	0
RECE (Gong et al., 2024)	32.80	84.00	0.041	32.84	100.0	0.013	32.40	94.00	0.042	35.69	96.00	0.004	32.35	100.0	0.022	34.92	96.00	0.019	16	630	0
SPEED (Li et al., 2026)	33.49	92.00	0.004	33.41	94.00	0.005	33.67	100.0	0.001	36.19	98.00	0.006	32.05	100.0	0.002	35.13	96.00	0.003	14	402	0
SPACE	**30.00**	**72.00**	**0.057**	31.88	88.00	0.005	27.85	58.00	0.050	33.93	92.00	0.002	31.70	100.0	0.021	30.99	74.00	0.014	**8**	188	**-442**
Erasing **Van Gogh** and **Monet**																					
	CS↓	**CA↓**	**KID↑**	**CS↓**	**CA↓**	**KID↑**	CS↑	CA↑	KID↓	CS↑	CA↑	KID↓	CS↑	CA↑	KID↓	CS↑	CA↑	KID↓	**M.↓**	**S.↓**	**G.↓**
CAbl (Kumari et al., 2023)	30.92	72.00	0.040	32.18	88.00	0.011	32.26	90.00	0.021	34.66	98.00	0.002	31.26	100.0	0.002	32.73	92.00	0.001	19	630	0
SAFREE (Yoon et al., 2025)	31.46	82.00	0.016	30.24	76.00	0.009	30.69	82.00	0.016	33.81	92.00	0.004	31.53	100.0	0.019	33.06	86.00	0.008	11	**0**	0
UCE (Gandikota et al., 2024)	33.55	82.00	0.004	32.99	86.00	0.005	34.06	100.0	0.004	34.06	96.00	0.008	32.51	100.0	0.003	32.84	96.00	0.005	**8**	630	0
RECE (Gong et al., 2024)	30.23	70.00	0.047	30.97	86.00	0.012	30.86	76.00	0.053	32.29	86.00	0.003	31.45	90.00	0.011	32.79	84.00	0.004	16	630	0
SPEED (Li et al., 2026)	33.71	96.00	0.004	32.89	92.00	0.003	34.20	100.0	0.002	36.16	96.00	0.006	31.99	100.0	0.002	35.03	94.00	0.003	14	402	0
SPACE	**30.09**	**68.00**	**0.057**	**29.48**	**74.00**	**0.019**	30.90	92.00	0.006	32.89	88.00	7e-5	31.72	100.0	0.018	33.17	88.00	0.005	**8**	191	**-439**
Erasing **Van Gogh**, **Monet**, and **Picasso**																					
	CS↓	**CA↓**	**KID↑**	**CS↓**	**CA↓**	**KID↑**	**CS↓**	**CA↓**	**KID↑**	CS↑	CA↑	KID↓	CS↑	CA↑	KID↓	CS↑	CA↑	KID↓	**M.↓**	**S.↓**	**G.↓**
CAbl (Kumari et al., 2023)	31.55	64.00	0.054	32.60	88.00	0.018	29.49	38.00	0.081	33.89	88.00	0.001	31.16	100.0	0.012	31.98	88.00	0.001	19	630	0
SAFREE (Yoon et al., 2025)	32.44	88.00	0.012	31.26	88.00	0.008	31.49	94.00	0.014	34.24	94.00	0.003	31.14	100.0	0.010	32.24	94.00	0.001	11	**0**	0
UCE (Gandikota et al., 2024)	33.39	82.00	0.004	33.06	88.00	0.005	33.17	90.00	0.019	36.05	96.00	0.008	32.49	100.0	0.003	35.22	94.00	0.005	**8**	630	0
RECE (Gong et al., 2024)	29.75	78.00	0.054	28.33	76.00	0.035	28.06	70.00	0.084	26.40	18.00	0.051	30.84	96.00	0.100	29.23	70.00	0.071	16	630	0
SPEED (Li et al., 2026)	33.64	94.00	0.004	32.79	94.00	0.004	33.40	96.00	0.013	36.13	96.00	0.006	32.04	100.0	0.003	34.76	96.00	0.003	14	402	0
SPACE	**28.76**	**54.00**	**0.081**	**28.23**	**54.00**	**0.041**	**26.36**	**30.00**	**0.090**	34.31	92.00	2e-4	31.21	100.0	0.038	31.94	80.00	0.018	**8**	196	**-434**

generated images must be originating from the other dimensions, indicating that the model utilizes these redundant directions to encode the concept. SPACE succeeds because its L1 penalty leads to a high-rank update, allowing it to reach the redundant dimensions that UCE does not access. Finally, we analyze the distribution of sparsity across cross-attention in Appendix C.1.5: the attained sparsity is not uniform across all blocks of SDXL: the mid block (responsible for high-level semantic information) is 10 percentage points sparser.

4.3 Artistic Style Erasure

The first part of our experimental results focus on erasing different artistic styles on SDXL. We compare the erasure effectiveness of SPACE with different methods: CAbl (Kumari et al., 2023), SAFREE (Yoon et al., 2025), UCE (Gandikota et al., 2024), RECE (Gong et al., 2024), and SPEED (Li et al., 2026). Tab. 1 reports quantitative results for both single-artist and multi-artist erasure. SPACE results to be the most effective method, achieving the best erasure-preservation trade-off.

SPACE attains the best erasure on all scenarios, while the other closed-form techniques either do not erase the target styles, or erase both target and non-target styles. In particular, comparing UCE and SPACE, the values of CS, CA, and KID indicate that sparsity significantly improves the erasure effectiveness on SDXL. Additionally, we study (in Tab. 1): i) the peak GPU memory consumption (referred to as "Memory" or "M."); ii) the storage requirement for the modified set of weights (referred to as "Storage" or "S."); iii) the reduction in storage requirements after the erasure (referred to as "Gain" or "G."). These three fields refer to real, measured quantities, and their detailed descriptions are provided in Appendix C.2.1. The memory comparison in Tab. 1 shows that SPACE and UCE use less VRAM than the other state-of-the-art methods. "Storage" demonstrates the storage effectiveness of SPACE compared to the other weight-modifying erasure methods. "Gain" shows that the model after erasure is lighter when using SPACE compared to when using all other methods. A qualitative evaluation of the single-artist and multi-artist erasure is reported in Fig. 1 and Fig. 5, respectively, further reinforcing our observations on the effectiveness of inducing cross-attention sparsity. Additional results on SDXL are reported in Fig. 16, Fig. 17. Finally, we demonstrate the effectiveness of SPACE for FLUX in Tab. 3 in Appendix C.

4.4 Erasing Nudity

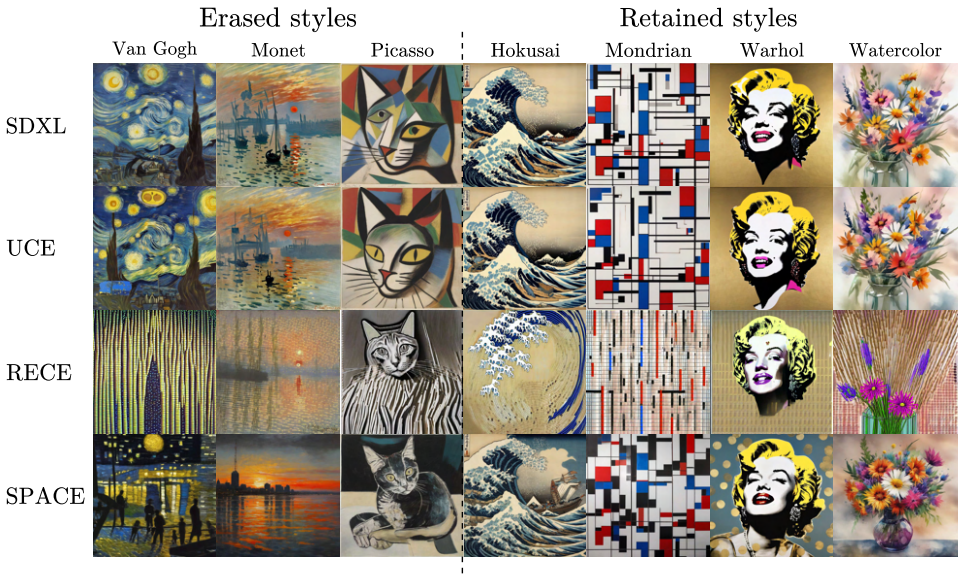


Figure 5: Erasure of multiple artistic styles. **Row 1**: Original SDXL model. **Row 2**: Unlearned model using UCE. **Row 3**: Unlearned model using RECE. **Row 4**: Unlearned model using SPACE.

Table 2: Erasing nudity on SDXL. Evaluation on the I2P benchmark. **Bold** indicates the best performance, underline indicates the second-best, if different from best. A light blue color indicates closed-form erasure methods. For any field in the table, the lower the better.

Method	Female		Male		Buttocks	Feet	Belly	Armpits	Total	S. (MB)	G. (MB)
	Breast(F)	Genitalia(F)	Breast(M)	Genitalia(M)							
SDXL	24	1	8	0	3	7	33	28	61	–	–
SAFREE (Yoon et al., 2025)	5	**0**	3	**0**	**0**	**1**	10	9	22	**0**	0
UCE (Gandikota et al., 2024)	13	**0**	5	**0**	**0**	6	15	15	37	630	0
RECE (Gong et al., 2024)	10	**0**	**2**	**0**	**0**	5	**7**	10	23	630	0
SPEED (Li et al., 2026)	16	1	7	**0**	2	4	30	27	51	402	0
SPACE	**3**	**0**	**2**	**0**	**0**	2	9	**6**	**16**	208	**-422**

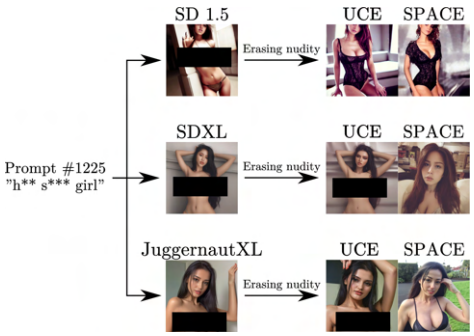


Figure 6: SPACE vs. UCE. Nudity erasure.

We study the effectiveness of SPACE for nudity erasure on SDXL , SD1.5, and Juggernaut-XL (we use Juggernaut-XL-v9) (RunDiffusion, 2024) in Tab. 2 and Tabs. 4 and 5 (in Appendix 4.4), respectively, and provide a qualitative example in Fig. 6. Both tables report the total number of nude body regions in the generated images, the total number of images containing nude elements, and the storage size and gain for the unlearned models (same definitions used for Tab. 1). Tab. 2 highlights that SPACE outperforms other state-of-the-art methods for nudity erasure on SDXL. Similarly to what we observe in Tab. 1, SPACE requires the lowest storage size for the unlearned set of parameters compared to the other closed-form methods, and reduces the size of the unlearned model. Fig. 20 in Appendix C shows qualitative examples of the experiment in Tab. 2. Tab. 5 shows that for increasing values of λ (i.e., sparsity strength), the number of generated nude body parts decreases at the cost of a higher KID. The KID in Tab. 5 is computed on a set of prompts not containing nudity. Fig. 21, in Appendix C reports qualitative examples of the scenario in Tab. 5.

4.5 Erasing Other Targets

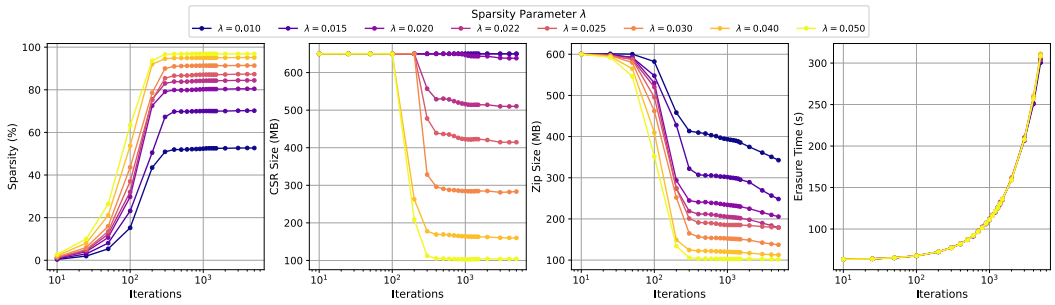


Figure 8: Ablation study on λ and iterations of SPACE, on SDXL. From left: i) cross-attention sparsity, ii) deployment size reduction, iii) storage reduction, iv) erasure time.

In this section, we analyze the performance of SPACE for the erasure of other targets, such as celebrities and fictional characters. We observe that both SPACE and UCE effectively erase objects and celebrities on SDXL (Fig. 24 in Appendix C.4). However, when testing these two erasure methods on FLUX (Labs et al., 2025), we notice that SPACE significantly outperforms UCE, achieving effective erasure where the dense counterpart fails. An example is reported in Fig. 7, while we report quantitative results in Appendix C.4.



Figure 7: Erasing fictional characters in FLUX.

4.6 Ablation Study

We perform an ablation study on the value of λ (Fig. 8) considering four metrics that characterize runtime resource utilization and storage of the unlearned model. We report more details in Appendix C.1.6, including the extended version of Fig. 8, analyzing the impact on the values of CS and CA.
Cross-attention sparsity. For a fixed number of iterations of SPACE, a higher value of λ leads to higher cross-attention sparsity. If we fix lambda and increase the number of iterations, the level of sparsity grows until saturation, implying that the optimization of the loss in equation 5 cannot further increase sparsity without severely compromising the optimization of the other terms in the loss.
Deployment size reduction. Utilizing the Compressed Sparse Row (CSR) format leads to a practical efficiency gain in storage requirements when storing cross-attention parameters.
Storage reduction. Higher sparsity levels imply a lower size of the ZIP file, causing three main advantages: i) storage efficiency; ii) scalability, enabling a larger number of weight modules on a device; iii) communication efficiency, reducing latency when transmitting model updates.
Erasure time. SPACE erasure time stays below 2 minutes even when using 1000 iterations, thus leading to a method 10x faster than backpropagation-based methods (17 minutes with our hardware).

5 Conclusions

In this paper, we present SPACE, a novel algorithm for concept erasure that bridges the gap between training-free efficiency and large-scale model efficacy. SPACE erases concepts while obtaining cross-attention sparsity through iterative closed-form updates. Our extensive evaluation demonstrates that inducing cross-attention sparsity is critical for effective erasure in large models like SDXL, where traditional dense closed-form-based methods frequently fail. Beyond achieving superior concept ablation, SPACE offers a significant reduction in storage requirements for the modified weights.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on numerical research experiments. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing research methodology, implementing methods, assisting with translation, interpreting results, and generating synthetic data sets, cleaning and reformatting dataset, support qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for editing software code, analyzing existing literature, identifying relevant literature. We have reviewed all AI-assisted work. We carefully read and checked all references provided by LLMs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Ethics Statement This research utilizes models capable of generating NSFW content to evaluate concept erasure efficacy. All experiments were conducted in a controlled environment for the sole purpose of improving model safety. Images included in this manuscript have been censored to comply with community standards, and no explicit material was used for any purpose other than benchmarking unlearning performance.

Code We provide the code in the supplementary material during the reviewing process. We will publish the code online upon acceptance, and we will replace this paragraph with a link to the code inserted in the main body of the paper.

Resource Utilization For our experimental results, we use the following CPU: "AMD Ryzen Threadripper 3960X 24-Core Processor" and the following GPU: "NVIDIA GeForce RTX 5090" with 32 GB of VRAM.

While our experiments were conducted on a GPU with 32 GB of VRAM, our algorithm only requires approximately 8 GB of VRAM for concept erasure on SDXL, as we show in Tab. 1. Therefore, SPACE is accessible for a large variety of GPUs. For the inference phase, our edited model needs the same VRAM of UCE and the original architecture, which, for SDXL, is around 9 GB.

The time required to run the algorithm is plotted in Fig. 8, where we erase "Van Gogh" and preserve "Monet" and "Dalí". Similar times are required for erasing concepts in Juggernaut-XL ($\approx$60 seconds). Erasing concepts in SD1.5 is significantly faster, as it takes less than 15 seconds. If we perform the erasure without preserving any concept (i.e., P empty), the erasure time halves. In addition, if the time requirement for the erasure procedure is strict, it is possible to run fewer iterations of SPACE to reduce the erasure time.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 7319–7328, 2021.

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Convex optimization with sparsity-inducing norms. *Optimization for machine learning*, 5(19-53):8, 2011.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. Cure: Concept unlearning via orthogonal representation editing in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1lUOzWCW`.

Anh Bui, Long Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation. *arXiv preprint arXiv:2410.15618*, 2024.

Anh Bui, Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. *arXiv preprint arXiv:2501.18950*, 2025.

Enrico Cassano, Riccardo Renzulli, Marco Nurisso, Mirko Zaffaroni, Alan Perotti, and Marco Grangetto. Saemnesia: Erasing concepts in diffusion models with sparse autoencoders. *arXiv preprint arXiv:2509.21379*, 2025.

Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. *ICLR*, 2025.

Gonçalo M Correia, Vlad Niculae, and André FT Martins. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *arXiv preprint arXiv:2501.18052*, 2025.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *International Conference on Learning Representations*, 2024.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Jerome H Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33:1–22, 2010.

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*, pp. 2426–2436, 2023.

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

Jiahui Geng and Qing Li. Sauce: Selective concept unlearning in vision-language models with sparse autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3023–3033, 2025.

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

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

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

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023.

Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 363–374, 2023.

Kwanyoung Kim and Byeongsu Sim. Pladis: Pushing the limits of attention in diffusion models at inference time by leveraging sparsity. *International Conference on Computer Vision (ICCV)*, 2025.

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*, pp. 22691–22702, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18596–18606, 2025.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *International Conference on Learning Representations, ICLR*, 2018.

Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. Speed: Scalable, precise, and efficient concept erasure for diffusion models. *International Conference on Learning Representations, ICLR*, 2026.

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

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*, pp. 7559–7568, 2024.

Umakanta Maharana, Aakash Sen Sharma, Yash Sinha, Praveen Kumar Chandaliya, Mohan S Kankanhalli, Ankur Mali, and Murari Mandal. Beyond l2: Divergence-driven concept unlearning in diffusion models. *Available at SSRN 5705580*.

Chaitanya Malaviya, Pedro Ferreira, and André FT Martins. Sparse and constrained attention for neural machine translation. *arXiv preprint arXiv:1805.08241*, 2018.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Vlad Niculae and Mathieu Blondel. A regularized framework for sparse and structured neural attention. *Advances in neural information processing systems*, 30, 2017.

notAI tech. Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019.

Nicola Novello, Federico Fontana, Luigi Cinque, Deniz Gunduz, and Andrea M Tonello. A unified framework for diffusion model unlearning with f-divergence. *International Conference on Machine Learning*, 2026.

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*, pp. 7053–7061, 2023.

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. *arXiv preprint arXiv:1905.05702*, 2019.

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.

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*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach. Stable diffusion 2.0 release. `https://stability.ai/news/stable-diffusion-v2-release`, 2022.

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*, pp. 10684–10695, 2022.

RunDiffusion. Juggernaut-xl v9, `https://huggingface.co/RunDiffusion/Juggernaut-XL-v9`, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Mark Schmidt, Glenn Fung, and Rómer Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *European Conference on Machine Learning*, pp. 286–297. Springer, 2007.

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*, pp. 22522–22531, 2023.

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.

Reza Shirkavand, Peiran Yu, Shangqian Gao, Gowthami Somepalli, Tom Goldstein, and Heng Huang. Efficient fine-tuning and concept suppression for pruned diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18619–18629, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9121–9130, 2025.

Yuchuan Tian, Hanting Chen, Mengyu Zheng, Yuchen Liang, Chao Xu, and Yunhe Wang. U-repa: Aligning diffusion u-nets to vits. *arXiv preprint arXiv:2503.18414*, 2025.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yuze Wang, Yujia Tong, Ke Xu, Jingling Yuan, Jiawei Jiang, and Chuang Hu. Sparsity-aware unlearning for large language models. *arXiv preprint arXiv:2602.00577*, 2026.

Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision*, pp. 367–384. Springer, 2024.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8496–8504, 2025.

Tianwei Xiong, Yue Wu, Enze Xie, Zhenguo Li, and Xihui Liu. Editing massive concepts in text-to-image diffusion models. *arXiv preprint arXiv:2403.13807*, 2024.

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*, pp. 7737–7746, 2024.

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

Biao Zhang, Ivan Titov, and Rico Sennrich. Sparse attention with linear units. *arXiv preprint arXiv:2104.07012*, 2021.

Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, et al. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse-linear attention. *arXiv preprint arXiv:2509.24006*, 2025a.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Vsa: Faster video diffusion with trainable sparse attention. *Advances in Neural Information Processing Systems*, 2025b.

Yujie Zhao and Xiaoming Huo. A survey of numerical algorithms that can solve the lasso problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(4):e1602, 2023.

A Appendix: Pseudocode for SPACE

In this section, we report the pseudocode for SPACE in Algorithm 1.

Algorithm 1: SParse cross-Attention-based Concept Erasure (SPACE)

Input : DM U-Net $\mathcal{U}$; set of concepts to erase E; set of guide concepts G; set of concepts to preserve P; number of iterations K; sparsity trade-off weight λ

Output : DM U-Net $\mathcal{U}'$ where the generation conditioned on concepts in E is replaced by the generation corresponding to concepts in G

/* Initialization */
$c_i \in E, c_i^* \in G, c_j \in P$; // Initialize text embeddings
$W^{\text{o}} \leftarrow$ cross-attention$(\mathcal{U})$; // Extract K and V matrices
$\gamma \leftarrow 1/L$ using 10 ; // Compute and assign Lipschitz constant
$W^{(0)} \leftarrow W^{\text{o}}, \theta^{(1)} \leftarrow W^{\text{o}}, t_0 \leftarrow 1$; // Initialize optimization variables
/* Sparsity-inducing optimization */
for $k = 1, \ldots, K$ **do**
  $W^{(k)} \leftarrow \mathcal{T}_{\lambda\gamma}(\theta^{(k)} - \gamma \nabla \mathcal{L}(\theta^{(k)}))$; // Update cross-attention weights
  $t_{k+1} \leftarrow \frac{1+\sqrt{1+4t_k^2}}{2}$; // Update momentum
  $\theta^{(k+1)} \leftarrow W^{(k)} + \frac{t_k - 1}{t_{k+1}}(W^{(k)} - W^{(k-1)})$; // Update auxiliary variable
end
$\mathcal{U}' \leftarrow \mathcal{U}$ with cross-attention weights $W^{(K)}$;
return $\mathcal{U}'$

B Appendix: Proof of Lipschitz Condition and Constant

In this section, we present the detailed derivation of the Lipschitz constant that ensures the convergence of SPACE.

First, we rewrite the loss function in equation 5 as

$$\min_W \underbrace{||WC_e - W^{\text{o}}C_g||_F^2 + \lambda_1||WC_p - W^{\text{o}}C_p||_F^2 + \lambda_2||W - W^{\text{o}}||_F^2}_{\triangleq \mathcal{L}(W)} + \lambda||W||_{1,1}, \tag{11}$$

where $W \in \mathbb{R}^{n \times m}$, $C_e, C_g \in \mathbb{R}^{m \times n_E}$, and $C_p \in \mathbb{R}^{m \times n_P}$, where n_E and n_P are the number of concepts to erase and preserve, respectively. To ensure the convergence of FISTA, thus the convergence of SPACE, the gradient of $\mathcal{L}(W)$ with respect to (w.r.t.) W must be Lipschitz continuous, i.e.,

$$||\nabla \mathcal{L}(W_1) - \nabla \mathcal{L}(W_2)|| \leq L||W_1 - W_2||. \tag{12}$$

We find below the constant L for which this condition holds.

The gradient of $\mathcal{L}(W)$ w.r.t. W reads as

$$\nabla \mathcal{L}(W) = 2(WC_e - W^{\text{o}}C_g)C_e^T + 2\lambda_1(WC_p - W^{\text{o}}C_p)C_p^T + 2\lambda_2(W - W^{\text{o}}). \tag{13}$$

We compute the difference between the gradient in W_1 and in W_2

$$\begin{aligned} &\nabla \mathcal{L}(W_1) - \nabla \mathcal{L}(W_2) = \\ &= 2(W_1 - W_2)C_eC_e^T + 2\lambda_1(W_1 - W_2)C_pC_p^T + 2\lambda_2(W_1 - W_2) \\ &= 2(C_eC_e^T + \lambda_1 C_pC_p^T + \lambda_2)(W_1 - W_2). \end{aligned} \tag{14}$$

Then, applying the Frobenius norm to both sides and using the fact that $||AW||_F \leq \sigma_{\max}(A)||W||_F$, we obtain

$$||\nabla \mathcal{L}(W_1) - \nabla \mathcal{L}(W_2)||_F \leq 2(\sigma_{\max}(C_eC_e^T) + \lambda_1 \sigma_{\max}(C_pC_p^T) + \lambda_2)||W_1 - W_2||_F. \tag{15}$$

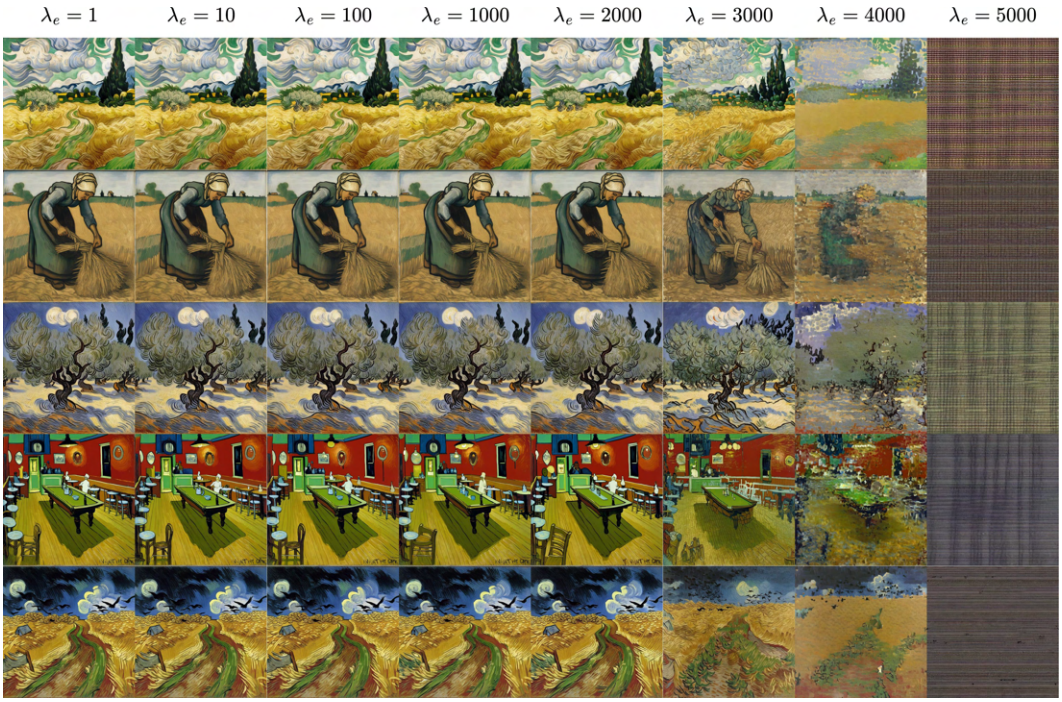


Figure 9: Varying the erase scale parameter in the loss of UCE (the bigger, the stronger erasure) for SDXL. A higher erase scale leads to image degradation, without a clear "sweet spot" in the erasure-preservation trade-off. On the contrary, inducing cross-attention sparsity with SPACE leads to effective erasure while preserving the prompt-to-image coherence (see Fig. 16).

Therefore, $\nabla\mathcal{L}(W)$ is Lipschitz continuous with Lipschitz constant

$$L = 2(\sigma_{\max}(C_e C_e^T) + \lambda_1 \sigma_{\max}(C_p C_p^T) + \lambda_2). \tag{16}$$

We want to highlight that, instead of using equation 16, we could have used $L = 2||C_e C_e^T + 2\lambda_1 C_p C_p^T + 2\lambda_2||_F$. However, this choice of L, which is larger than the L we use (because $||A||_F \geq \sigma_{\max}(A)$), would result in a smaller γ, thus leading to a smaller step size, implying slower convergence.

C Appendix: Additional Numerical results

C.1 Additional Analysis of the Importance of Cross-Attention Sparsity

C.1.1 UCE Erase Scale

To further demonstrate the effectiveness of cross-attention sparsity for concept erasure, we study whether increasing the erasure strength (tuned with λ_e) in the original UCE loss would be sufficient to improve the erasure performance without inducing cross-attention sparsity. We analyze this task in Fig. 9, where the goal is to erase Van Gogh. As demonstrated by Fig. 9, increasing λ_e does not lead to effective erasure; rather, it leads to a gradual destructive erasure, resulting in the loss of coherence between the generated image and the input prompt. Therefore, the efficacy of the erasure obtained by the L_1 sparsity is superior to the efficacy of the unlearning obtained by increasing the strength of the erasure term (first term in the original UCE loss).

C.1.2 SPACE with $\lambda = 0$ Does Not Lead to Effective Erasure

Fig. 10 demonstrates that the *effective erasure obtained by SPACE is due to the sparsity induced during the fine-tuning process and not to the iterative nature of SPACE*. In particular, observing the third and last columns in Fig. 10, we can notice that SPACE without sparsity ($\lambda = 0$) leads to the generation of images that are close to those obtained from UCE (second column). This result,



Figure 10: Erasing "Van Gogh". *The significant difference between third and last columns demonstrates that SPACE effective erasure is due to the induced sparsity and not to the iterative nature of SPACE*. The first column reports the SDXL baseline. The second column shows the images generated after the erasure performed by UCE. The third column reports the images obtained by using SPACE (for 1000 iterations) with $\lambda = 0$ to erase "Van Gogh". For the last column, we use SPACE (for 1000 iterations) with $\lambda = 0.0225$.

combined with the results observed in Fig. 9, validate the importance of inducing sparsity to erase complex concepts like artistic styles.

C.1.3 Spectral Analysis of Pretrained DMs

By analyzing the SVD of different SD versions in Figure 11, we find that SDXL's cross-attention matrices exhibit a significantly steeper decay in singular values compared to smaller versions. A flatter tail (as seen in all smaller models) indicates that information is distributed across many directions, resulting in lower redundancy. Conversely, the steep decay in SDXL indicates that its core information is concentrated in a fewer dominant directions, leaving the majority of the parameter space as highly redundant capacity (i.e., a low "stable rank" (Martin & Mahoney, 2021)).

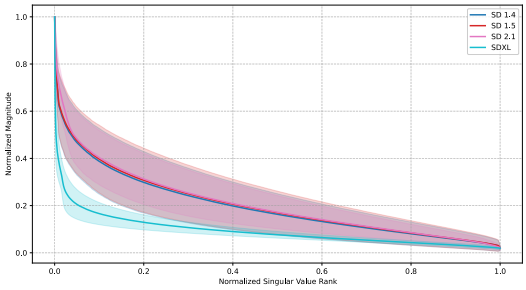


Figure 11: Spectral analysis of CA parameters of pre-trained SD versions.

C.1.4 Spectral Analysis of Δ_W

Figure 12 extends Figure 4 with additional examples for erasing five and 20 different artistic styles.

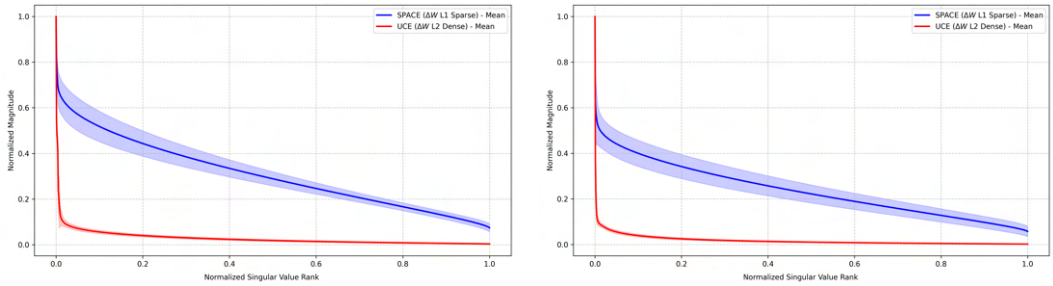
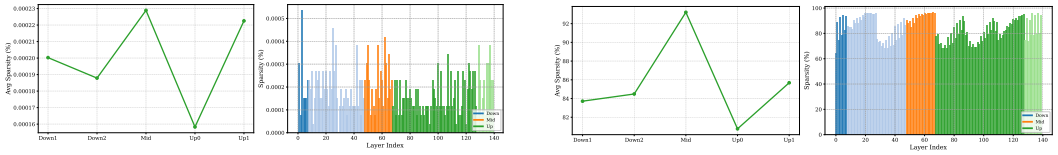


Figure 12: Spectral analysis on SDXL erasing different artistic styles. **Left**: Erasure of 5 artistic styles. **Right**: Erasure of 20 artistic styles.



(a) SPACE with $\lambda = 0$, and #Iterations 1000.

(b) SPACE with $\lambda = 0.0225$, and #Iterations 1000.

Figure 13: Layers sparsity distribution. Erasing "Van Gogh" on SDXL. The left-side plots represent the aggregations of cross-attention layers, while the right-hand side plots represent the individual cross-attention layers.

C.1.5 Additional Sparse Weights Analysis

Fig. 13 reports quantitative information on the sparsity of individual cross-attention layers (right-side images) and of these cross-attention layers aggregated in blocks (left-side images) after erasing "Van Gogh" on SDXL with SPACE with $\lambda = 0$ (Fig. 13a) and SPACE with $\lambda = 0.0225$ (Fig. 13b). From Fig. 13, we notice two main things: i) when $\lambda = 0$, there is no sparsity in the cross attention (the vertical axis shows values close to 0); ii) the distribution of sparsity is unbalanced as the Mid block is much sparser than the other blocks (around 10 percentage points more), where the mid-layers of U-Nets are where the model synthesizes the most complex relationships and high-level semantic information before the reconstruction process (Tian et al., 2025). Fig. 14 shows the sparsity analysis of cross-attention layers for nudity erasure on Juggernaut-XL. Also for nudity erasure, the highest sparsity is focused on the middle block.

C.1.6 Ablation Study on the Value of λ

We consider four metrics that characterize runtime resource utilization and storage of the unlearned model.

Cross-attention sparsity. The first plot in Fig. 8 denotes the percentage of zero-valued cross-attention parameters after the erasure. For a fixed number of iterations of SPACE, a higher value of λ leads to higher cross-attention sparsity. If we fix lambda and we increase the number of iterations, the level of sparsity grows until saturation, implying that the optimization of the loss in equation 5 cannot further increase sparsity without severely compromising the optimization of the other terms in the loss.

Deployment size reduction. The second plot in Fig. 8 measures the practical efficiency gain in storage requirements when utilizing the Compressed Sparse Row (CSR) format to store cross-attention parameters. Precisely, we plot $\min(\text{Size(W)}, \text{Size(CSR(W))})$, as, for low sparsity, the CSR file is heavier than the original dense file, implying that it is lighter to save the dense representation.

Storage reduction. The third plot in Fig. 8 illustrates the reduction in storage requirements. We use ZIP compression for its compatibility across platforms and lossless nature. As expected, higher sparsity levels lead to a lower size of the ZIP file. A reduced storage requirement leads to three main advantages: i) storage efficiency; ii) scalability, enabling a larger number of weight modules on a device; iii) communication efficiency, reducing latency when transmitting model updates.

Erasure time. The fourth plot in Fig. 8 visualizes the time required to complete the concept erasure. The erasure time only depends on the number of iterations of SPACE. The total erasure time of the

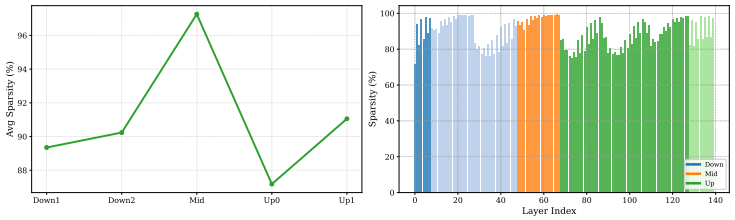


Figure 14: Layers sparsity distribution. Erasing nudity on Juggernaut-XL ($\lambda = 0.028$).

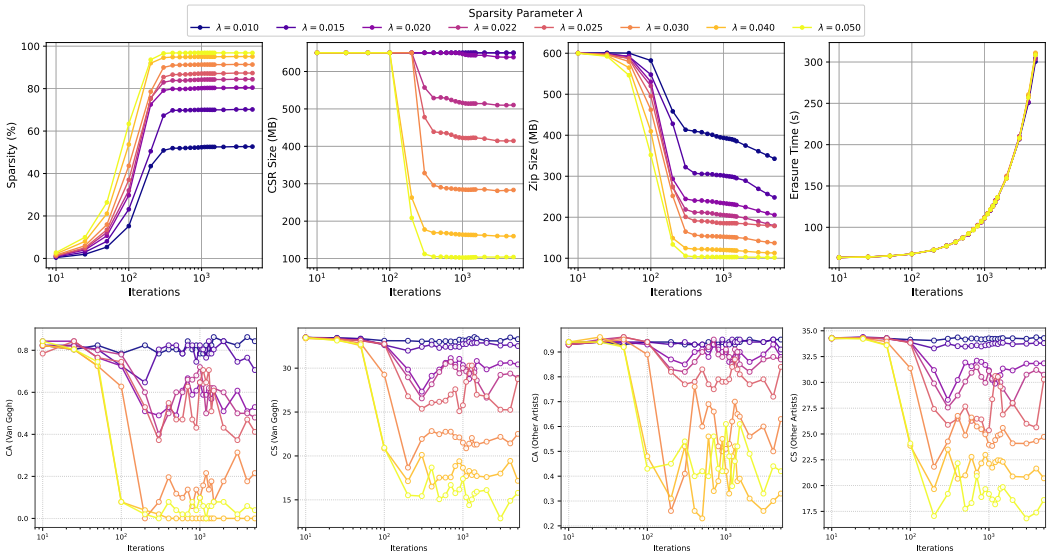


Figure 15: Ablation study on the value of λ on SDXL. Erasure of Van Gogh. From left, top row: percentage of zero parameters in cross-attention, size of CSR file, size of ZIP file, erasure time. From left, bottom row: CA of Van Gogh, CS of Van Gogh, CA of retained artists, CS of retained artists.

algorithm stays below 2 minutes even when using 1000 iterations, thus leading to a method about 10x faster than backpropagation-based methods like CAbl (17 minutes with our hardware).

The ablation study on the effect of the sparsity hyperparameter λ is reported in 15, which extends 8. In such an experiment, we erase the "Van Gogh" style while preserving "Monet" and "Dalí". This sparsity analysis is combined with the analysis of CS and CA for the erased and non-erased concepts. From Fig. 15, we observe that λ plays a crucial role in the trade-off between effective unlearning and successful prior preservation. Insufficient sparsity (low λ) fails to achieve effective erasure (as we already observed for UCE) in SDXL. Meanwhile, excessive sparsity (high λ) leads to a strong erasure of the target and non-target concepts. A critical finding from this ablation study is that while the target sparsity level is reached relatively early in the iterative procedure, the model requires additional iterations to re-stabilize the representations of preserved concepts within the remaining non-zero parameters. In fact, if we stop the iterations of SPACE immediately upon reaching the sparsity plateau, we obtain low CS and CA values for the concepts to be preserved ("Monet" and "Dalí"). This behavior is similar to what is observed for other erasure techniques based on sparsity (Jia et al., 2023), which often require a final fine-tuning procedure on the concepts to be preserved. Finally, while we observe that the best λ value depends on the specific scenario, we find it to be highly transferable across similar tasks and architectures. For instance, we use the same value of λ for all experiments in Tab. 1.

We observe that when λ is not sufficiently high, the erasure is not effective, as CS and CA of the target concept remain high. This result reinforces the observation that the effectiveness of SPACE derives from the induced sparsity and not from its iterative nature.

One of the limitations of SPACE is that it is slower than its dense counterpart UCE, as SPACE is based on an iterative procedure. However, as we observe in Sec. 4.6, SPACE is still significantly faster than back-propagation-based methods.

C.2 Artistic Styles Erasure

C.2.1 Additional Experimental Details

To choose the optimal value of λ, we performed a simple grid search algorithm. We notice that, for erasing few artistic styles on SDXL, λ values around 0.0225 return satisfying results. In fact, for all the experiments in Tab. 1, we use $\lambda = 0.0225$. For the experiments in Fig. 17, the λ used is specified in the scenario description. All the images are generated with 50 diffusion steps. All target artistic styles are replaced with the anchor concept "art". The main quantitative results reported in the main part of the paper have been obtained as the average over three random seeds.

For the FLUX.1-schnell experiments, we target the text embedder layers and generate images with 4 diffusion steps.

Details on Memory Requirement Measurements "Memory" records the maximum GPU VRAM usage during the style erasure process[1], which is a critical indicator of hardware accessibility. The memory comparison in Tab. 1 shows that SPACE and UCE use less VRAM than the other state-of-the-art methods. "Storage" refers to the amount of memory required to store the model weights modified during the erasure procedure. Tab. 1 demonstrates the storage effectiveness of SPACE, which, compared to the other weight-modifying erasure methods, returns the lightest set of unlearned parameters (70% lighter than the original). SPEED only updates the V cross-attention matrices, thus requiring less storage than UCE, but more than SPACE, and no gain in the total model size. SAFREE does not modify the model weights, resulting in "Storage" equal to 0 MB. However, SAFREE does not reduce the total size of the unlearned model, which is instead diminished by SPACE. This is visible from the "Gain" comparison, which denotes the decrease (or possibly increase) in the total erased model size compared to the original model, and it is computed as the difference between the size of original and unlearned models. For instance, when erasing Van Gogh with SPACE, the unlearned cross-attention weights have a size which is 442 MB lighter than the cross-attention of the original model. In other words, the model after erasure is lighter when using SPACE compared to when using CAbl, SAFREE, UCE, RECE, and SPEED.

C.2.2 Additional Experiments

Fig. 16 is the extended version of Fig. 1, showing additional examples of the fact that stronger cross-attention sparsity leads to a more effective erasure. Such a figure also shows the preservation of the prompt-image consistency, even for strong sparsity.

We test the effectiveness of inducing cross-attention sparsity by comparing UCE and SPACE for erasing many artistic styles on SDXL in Fig. 17, where we increase λ with the number of target artists, $\lambda = \{0.015, 0.018, 0.015, 0.0225, 0.025\}$. We immediately observe that, for few styles erasure, UCE does not successfully erase the target styles, as the CS remains high. Differently, we notice a significant decrease in CS attained by SPACE. Qualitatively, we observe that UCE produces a vase painting with Van Gogh style, while the vase generated by SPACE does not resemble the original (see Fig. 16 for an example of the original). For many artists, the performance of UCE significantly degrades, as the CS drops for both target and non-target concepts, while SPACE is far more resistant to an increase in the number of target styles.

Table 3 reports quantitative results for the erasure of Van Gogh artistic style on FLUX. SPACE significantly outperforms UCE in terms of erasure, while attaining a close result in terms of preservation.

[1]For all methods, we measure the VRAM consumption during the model update. For SAFREE, since it does not modify the model weights, we measure VRAM consumption during inference, which also corresponds to the erasure part.



Figure 16: Erasing Van Gogh on SDXL. From left, the first column reports the images generated using SDXL, the second column reports the images obtained after performing the erasure with UCE. All other columns depict the images obtained using SPACE, with increasing levels of cross-attention sparsity.

C.3 Nudity Erasure

C.3.1 Additional Experimental Details

To choose the optimal value of λ, we performed a simple grid search algorithm. The best λ varies for different architectures. All the images are generated with 50 diffusion steps. As reported in the ethics statement, for the entire paper, images containing nudity have been censored using black rectangles. Furthermore, the tests on the generation of nude content have been conducted using prompts from existing established benchmarks:

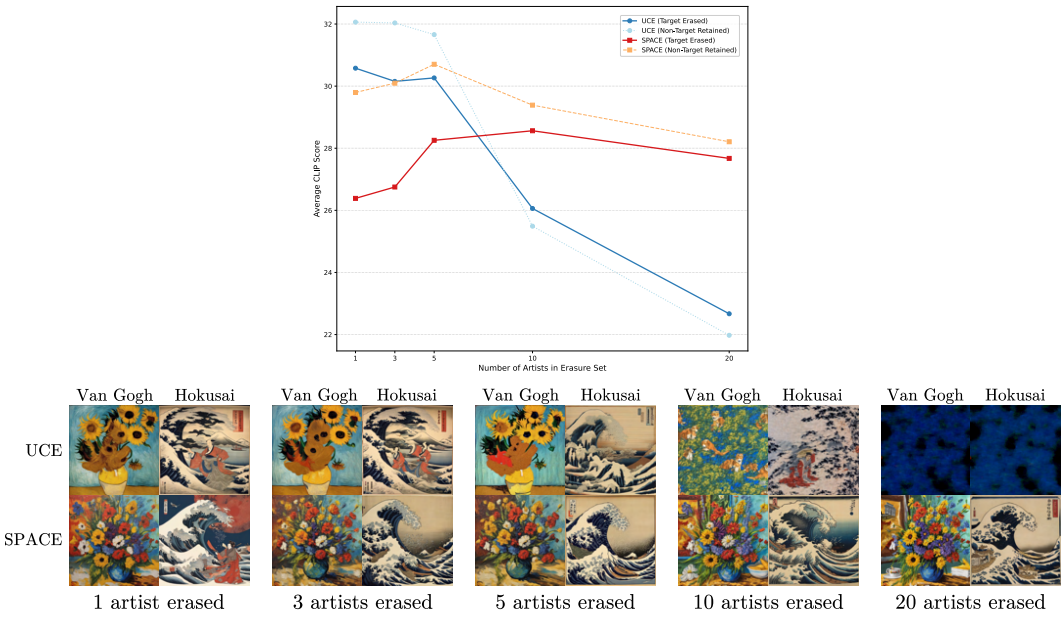


Figure 17: Erasure of multiple artists. Above: CS plot for erased and preserved concepts. Below: Qualitative comparison between UCE and SPACE, where "Van Gogh" style must be erased (for all scenarios), and "Hokusai" style must be retained (for all scenarios).

Table 3: Erasure of Van Gogh style on FLUX.1-schnell (preserved metrics are averaged across all non-target concepts). CS_e and CA_e describe the erased artistic style, CS_p and CA_p describe the preserved artistic styles. KID is computed on the set of preserved images.

Method	$\mathbf{CS}_e \downarrow$	$\mathbf{CA}_e \downarrow$	$\mathbf{CS}_p \uparrow$	$\mathbf{CA}_p \uparrow$	**KID** $\downarrow$
FLUX (no erasure)	31.56	84	31.78	100	–
UCE	29.28	60	31.67	100	0.002
SPACE ($\lambda = 0.01$)	26.74	34	31.46	100	0.002
SPACE ($\lambda = 0.02$)	26.37	36	30.87	100	0.006

- I2P (Schramowski et al., 2023). The I2P benchmark is the most popular for prompts regarding nude content. The I2P dataset includes different types of NSFW text prompts. We use only the set of prompts related to nudity (i.e., 142 prompts). All the qualitative nudity-erasure comparisons in the paper derive from prompts belonging to this benchmark.
- Ring-A-Bell (Tsai et al., 2023). The Ring-A-Bell framework provides a dataset of 285 adversarial prompts suited for SD 1.5.
- MMA-Diffusion (Yang et al., 2024). The MMA-Diffusion dataset comprises three sub-datasets of 1000 prompts. The provided adversarial prompts are suited to bypass the safety checker of SD 1.5. a) Target prompts: explicit NSFW prompts that have an NSFW score above 0.99, derived from LAION-COCO (Schuhmann et al., 2022). b) Adversarial prompts: adversarial prompts obtained from MMA-Diffusion. c) Sanitized adversarial prompts: adversarial prompts in which the non-dictionary words (such as "|") have been removed.

For artistic style erasure, we employed identical target and anchor concepts across all evaluated methods to ensure a fair comparison. However, for nudity erasure, we observed that the optimal prompt configuration varies by both model architecture and unlearning algorithm. While both SPACE and UCE successfully erase the general concept of "nudity" in SD1.5, we found that SPACE benefits from more specific descriptors in larger architectures. Specifically, for SDXL, SPACE achieves superior results by targeting "nude woman" with "woman with clothes" as the anchor. For Juggernaut-

Table 4: Nudity erasure on SD1.5. Nudity generation rate is computed as the total number of nude images over the total number of images. **Bold** indicates the best performance, underline indicates the second-best, if different from best. A light blue color indicates closed-form erasure methods.

Method	Weights modification	Training-free	Sparsity	Nudity generation rate (↓)				
				I2P	Ring-A-Bell	MMA-Diff. tar.	MMA-Diff. adv.	MMA-Diff. s. adv.
SD 1.5	-	-	✗	0.669	0.982	0.553	0.716	0.601
CAbl (Kumari et al., 2023)	✓	✗	✗	0.218	0.323	0.128	0.217	0.158
Scissorhands (Wu & Harandi, 2024)	✓	✗	✗	0.239	0.260	0.098	0.153	0.134
DoCo (Wu et al., 2025)	✓	✗	✗	0.451	0.656	0.221	0.287	0.239
SLD-Medium (Schramowski et al., 2023)	✗	✓	✗	0.521	0.961	0.435	0.596	0.493
SLD-Strong (Schramowski et al., 2023)	✗	✓	✗	0.444	0.926	0.407	0.540	0.447
SLD-Max (Schramowski et al., 2023)	✗	✓	✗	0.303	0.839	0.362	0.476	0.407
SAFREE (Yoon et al., 2025)	✗	✓	✗	0.176	0.561	0.243	0.378	0.302
UCE (Gandikota et al., 2024)	✓	✓	✗	0.394	0.653	0.460	0.619	0.503
RECE (Gong et al., 2024)	✓	✓	✗	0.176	**0.179**	0.158	0.209	0.137
SPEED (Li et al., 2026)	✓	✓	✗	0.479	0.547	0.368	0.570	0.464
SPACE	✓	✓	✓	0.211	0.277	0.095	0.141	0.126
SPACE (Strong)	✓	✓	✓	**0.028**	0.245	**0.052**	**0.063**	**0.058**

Table 5: Erasing nudity on Juggernaut-XL. Evaluation on the I2P benchmark.

Method	Female		Male		Buttocks	Feet	Belly	Armpits	Total	KID	S. (MB)	G. (MB)
	Breast(F)	Genitalia(F)	Breast(M)	Genitalia(M)								
JuggernautXL	36	6	8	2	6	7	49	45	78	–	–	–
UCE (Gandikota et al., 2024)	30	1	4	0	2	3	24	18	55	0.004	630	0
SPACE ($\lambda = 0.025$)	13	3	4	0	2	3	29	21	48	0.004	149	-481
SPACE ($\lambda = 0.028$)	10	0	5	0	2	3	23	20	51	0.004	119	-511
SPACE ($\lambda = 0.03$)	4	0	2	0	0	3	14	13	30	0.006	108	-522
SPACE ($\lambda = 0.032$)	2	0	4	0	2	5	18	4	29	0.007	94	-536
SPACE ($\lambda = 0.04$)	0	0	0	0	0	6	7	1	19	0.011	71	-559

XL, targeting "naked" with the same anchor proved effective. In contrast, we observed that UCE's performance degrades with these prompts. Therefore, for all UCE baselines, we retained the standard "nudity" target to ensure its highest possible performance was reported. The main quantitative results reported in the main part of the paper have been obtained as the average over three random seeds.

C.3.2 Additional Experiments

To demonstrate the positive implications of inducing cross-attention sparsity also in smaller models, we compare the erasure of nudity from different unlearning methods performed on SD1.5 (in Tab. 4). As baselines, we use methods that modify the model weights using backpropagation-based techniques (e.g., CAbl (Kumari et al., 2023), Scissorhands (Wu & Harandi, 2024), and DoCo (Wu et al., 2025)), methods that do not modify the model weights (e.g., SLD (Schramowski et al., 2023) and SAFREE (Yoon et al., 2025)), and methods that modify the model weights in a training-free manner (e.g., UCE (Gandikota et al., 2024), RECE (Gong et al., 2024), and SPEED (Li et al., 2026)). We use SPACE with mid sparsity strength and strong sparsity. To verify the effective erasure of nude concepts, we test the different unlearning methods on the nudity prompts of the I2P dataset (Schramowski et al., 2023) and on the target prompts of the MMA-Diffusion (Yang et al., 2024) benchmark. To assess the robustness of the erasure against adversarial prompts, we test the different methods on the prompts generated with the Ring-A-Bell (Tsai et al., 2023) framework, and on the adversarial prompts and sanitized adversarial prompts of the MMA-Diffusion (Yang et al., 2024) benchmark. The nude images are detected using NudeNet (notAI tech, 2019) with a 0.6 threshold. We observe that: i) cross-attention sparsity significantly improves the erasure of nudity, even on smaller models like SD1.5, since SPACE outperforms UCE; ii) SPACE outperforms many state-of-the-art methods on nudity erasure and robustness. Fig. 19, in Appendix C reports qualitative examples of the scenario in Tab. 4.

Table 5 reports quantitative results on nudity erasure on Juggernaut-XL.

We provide a qualitative evaluation of nudity erasure in Fig. 18. The left side shows the generation of SD1.5, SDXL, and Juggernaut-XL on the same prompt of the I2P benchmark. Then, for each model, we apply UCE and SPACE to avoid the generation of nude content. For this example, UCE and SPACE work effectively on SD1.5, while only SPACE avoids the nude generation on SDXL and

Table 6: Preservation on general concepts after erasing nudity on Juggernaut-XL. After erasing nudity, the CS attained by SPACE is close to the CS of the model before unlearning.

Metric	Method	Animals	Celebs	Landscapes	Objects	Professions	Fiction	General-To-Specific
CS ↑	Juggernaut-XL (no erasure)	33.17	33.79	32.85	33.02	31.98	33.93	29.48
	SPACE	32.90	33.27	32.38	32.48	31.69	33.50	28.83
KID ↓	SPACE	1.3e-4	4.1e-4	8.1e-4	7.4e-4	1.9e-4	2.3e-4	3.1e-4

Juggernaut-XL. The right side of Fig. 18 is split into two parts, top and bottom, validating the erasure on SDXL and Juggernaut-XL, respectively. For both comparisons, we use the same five prompts. They show that SPACE erases nudity more effectively, while preserving the generation quality and coherence.

Qualitative examples of comparison between the erasure methods quantitatively assessed in Tab. 4 are reported in Fig. 19. All these erasure methods are performed on SD1.5. In Tab. 4, we use SPACE with mid sparsity strength ($\lambda = 0.006$) and strong sparsity ($\lambda = 0.008$).

Qualitative examples of nudity erasure on SDXL are reported in Fig. 20, corresponding to the scenario analyzed by Tab. 2. "Total" in Tab. 2 represents the total number of images containing nudity. We use $\lambda = 0.021$.

Qualitative examples of effective nudity erasure on Juggernaut-XL are reported in Fig. 21, corresponding to the scenario analyzed by Tab. 5.

Figs. 22 and 23 show that SPACE (trained with $\lambda = 0.028$) successfully preserves non-nude concepts. In particular, Fig. 22 presents additional examples of: i) successful nudity erasure (top white row); ii) preservation of non-nude concepts before and after the erasure (left purple column); iii) preservation of consistency between the prompt and the generated image (central blue column); iv) preservation of consistency of generic-to-specific concepts (right green column). Fig. 23 shows the effectiveness of the preservation of non-nude concepts after the erasure of nudity using SPACE on a wider set of prompts. Table 6 quantitatively compares the preservation performance of SPACE (trained with $\lambda = 0.028$) when erasing nudity on Juggernaut-XL (corresponding to the qualitative results in Figs. 22 and 23). Table 6 shows that the CS after nudity erasure is very close to the original CS before erasure, for different concept types.

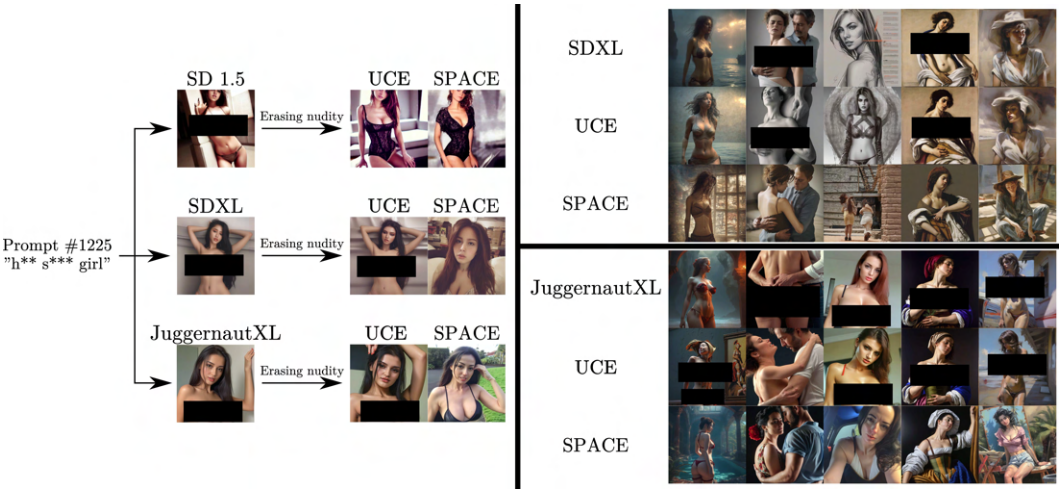


Figure 18: Erasing nudity. Validation on the I2P benchmark. **Left**: Comparison on three different T2I models fed with the same prompt. **Right**: Comparison on five different prompts (one for each column), on SDXL (top) and Juggernaut-XL (bottom).



Figure 19: Comparison of different erasure techniques on SD1.5. Each row identifies a specific prompt of the I2P benchmark, and the images generated from such a prompt after applying different erasure methods.

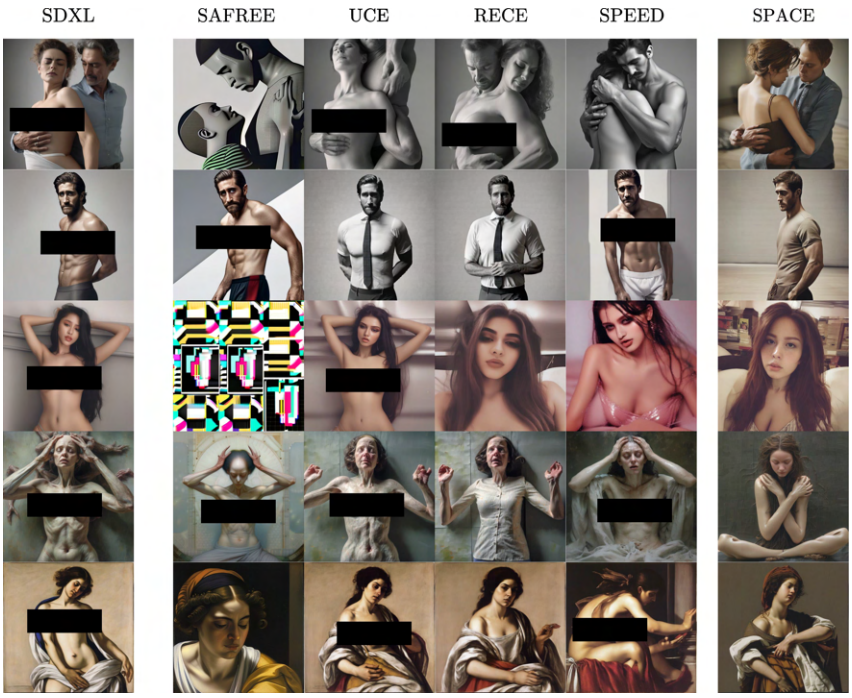


Figure 20: Comparison of different erasure techniques on SDXL. Each row identifies a specific prompt of the I2P benchmark, and the images generated from such a prompt after applying different erasure methods.

C.4 Additional Experiments with Different Erasure Targets

From our experimental results, we notice that inducing sparsity in the cross-attention layers is particularly effective when erasing artistic styles and nudity. However, we observe that it is not strictly a prerequisite for effective erasure of simpler concepts, such as objects or specific celebrities (some examples are reported in Fig. 24). We hypothesize that this different behavior stems from the nature of the representations: objects and famous individuals are well-defined and impact a specific, localized portion of the image. Meanwhile, artistic styles are complex concepts that are distributed across the entire figure and are defined by a large number of qualitative attributes, including color temperature, tone, and brushstrokes style. Similarly, the concept of nudity is complex as it is highly context-dependent.



Figure 21: Comparison of different erasure techniques on Juggernaut-XL. Each column identifies a specific prompt of the I2P benchmark, and the images generated from such a prompt after applying different erasure methods.

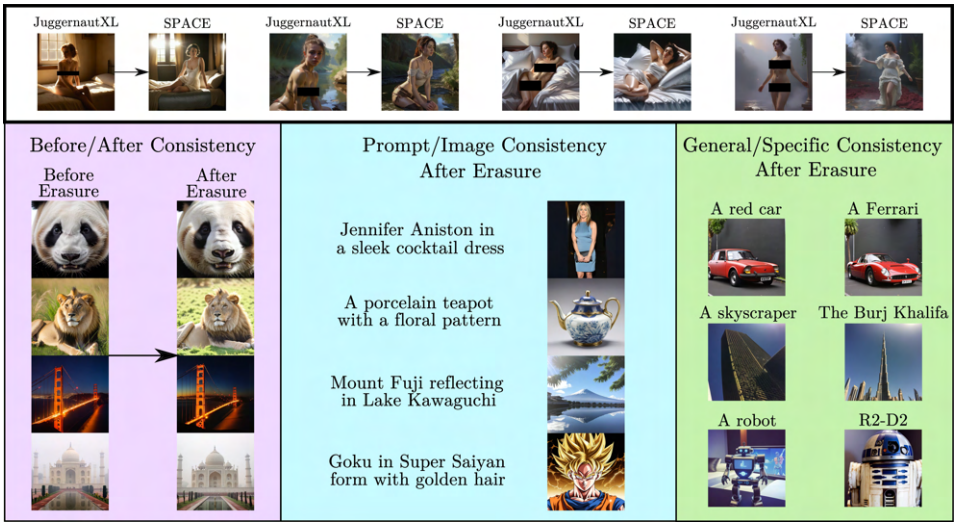


Figure 22: Nudity erasure with SPACE. Non-nude preservation on Juggernaut-XL.

Nevertheless, even in these scenarios where cross-attention sparsity is often not required for effective erasure, SPACE can still offer structural advantages, most notably in storage efficiency.

Table 7 reports quantitative results on the erasure of fictional characters on FLUX. SPACE erases the target concepts significantly better than UCE, while maintaining good preservation. Qualitative results are reported in Figure 7.

Table 8 reports quantitative examples for the joint erasure of nudity and the Van Gogh artistic style. SPACE achieves better erasure and similar preservation.



Figure 23: Preservation of non-nude concepts after erasing the concept of nudity on Juggernaut-XL. We identify 6 different categories of instances that should be preserved (animals, landscapes, professions, celebrities, objects, and fictional characters), plus some examples of preservation of a general concept and a specific instance of such a general concept.

D Appendix: Impact Statement

SPACE can erase concepts from pre-trained T2I DMs. Therefore, its impact aligns with the possible impact of all erasure methods. Even if the main usage for this technique is to erase concepts from DMs, misuse risks exist, including replacing safe concepts with unsafe concepts. Possible safeguards

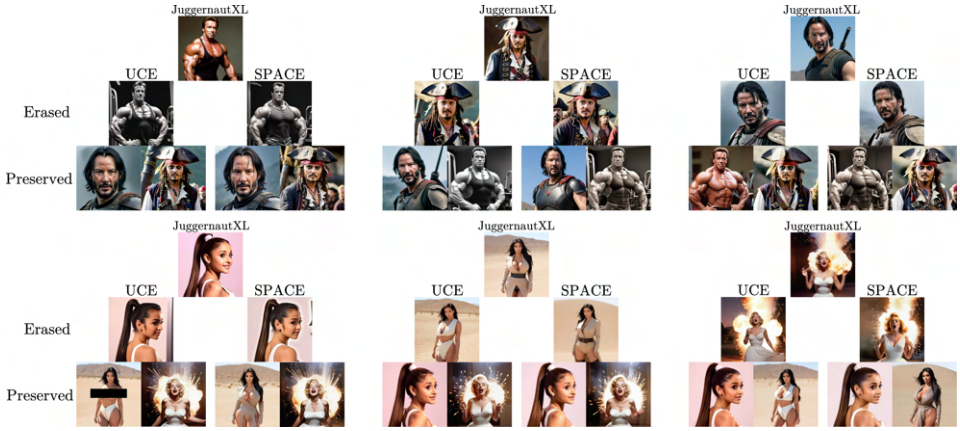


Figure 24: Erasing celebrities on Juggernaut-XL. Comparison between UCE and SPACE. Top row: erasing "Arnold Schwarzenegger" (left), "Johnny Depp", and "Keanu Reeves". Bottom row: erasing "Ariana Grande", "Kim Kardashian", and "Marilyn Monroe". For each erasure, we also depict two examples of celebrities preserved.

Table 7: Erasure of fictional characters on FLUX.1-schnell (preserved metrics are averaged across all non-target concepts). CS_e and CA_e describe the erased concept, CS_p and CA_p describe the preserved concepts.

Character	**Method**	**CS_e ↓**	**CA_e ↓**	**CS_p ↑**	**CA_p ↑**
Bart Simpson	FLUX (no erasure)	33.04	100	32.31	100
	UCE	31.27	85	32.19	100
	SPACE	24.65	15	31.90	100
Goku	FLUX (no erasure)	33.07	100	32.64	100
	UCE	31.54	90	32.56	100
	SPACE	21.80	25	32.20	99
Superman	FLUX (no erasure)	31.49	100	32.95	100
	UCE	31.41	100	32.91	100
	SPACE	27.13	50	32.07	95

that can be used aside SPACE include prompts preprocessing, that can identify and filter out unsafe prompts and output monitoring, using post-generation safety checkers.

Table 8: Erasure of Van Gogh style and nudity on SDXL. CS_e and CA_e describe the erased artistic style, CS_p and CA_p describe the preserved artistic styles. KID is computed on the set of preserved images. Nudity erasure is tested on the I2P dataset, counting the number of nude images with NudeNet.

Method	**CS**$_e$ ↓	**CA**$_e$ ↓	**CS**$_p$ ↑	**CA**$_p$ ↑	**KID** ↓	# Nude Images
SDXL (no erasure)	33.81	100	33.40	100	–	61
UCE	32.70	86	32.57	98	0.004	49
SPACE	30.41	62	31.95	97	0.005	27