Achieving Subcategorical Erasure in Text-to-Image Models

Pranav Singh Chib[1] and Pravendra Singh[1]

[1] Department of Computer Science and Engineering
Indian Institute of Technology Roorkee, India
{pranavs_chib, pravendra.singh}@cs.iitr.ac.in

Abstract. The emergence of large-scale text-to-image diffusion (T2ID) models has led to significant advancements in generating high-quality visual content from textual prompts. However, these powerful capabilities have also raised growing concerns about the generation of harmful and copyrighted material. While existing concept erasure techniques can effectively block the production of specific unwanted concepts from prompts, they often fall short when it comes to erasing an entire category (including subcategories) and are typically limited to handling only a few concepts at a time. In this paper, we introduce **S**ubcategorical **U**nlearning via **R**egularized **E**rasure (SURE), a novel method for removing entire subcategories from text-to-image diffusion models using only a single parent category as the target. Unlike prior approaches, SURE does not rely on sets of synonyms. Instead, it employs concept space to discover and eliminate the target category while preserving the model's overall utility. To further enhance erasure, SURE incorporates Lipschitz regularization, which encourages smoother model responses to perturbations around the target category. Specifically, the regularization promotes consistent behavior in the model's latent space when exposed to slight variations of the category to be forgotten. This smoothness constraint aids in erasure while maintaining the model's ability to generate unrelated content. Extensive experiments conducted across three tasks—object removal, suppression of explicit content, and elimination of artistic styles demonstrate that SURE achieves balanced performance in both effective category erasure and preservation of non-target concepts.

Keywords: Subcategorical Concept Erasing · Diffusion Models

1 Introduction

Text-to-image generative models [6, 7, 25, 29, 32] have made significant advancements in generating high-quality images almost instantly from simple textual prompts. These models learn to replicate diverse visual concepts from large-scale datasets [5] of image-caption pairs from the internet. However, this extensive model training comes with considerable risks, including the generation of copyrighted material [16], privacy-invading content [36], and explicit imagery [35], all of which raise safety concerns.

Fig. 1: The figure illustrates similarity scores (left) for different object categories across various erasure methods, where each category is used as the erasure target. Lower scores correspond to better erasure. Existing methods, including ESD, ACE, UCE, AGE, and AP, often fail to erase sub-categories within a given category completely. For example, in the case of the guns category, ESD fails to remove sub-categories such as pistols fully, as shown in the generated samples on the right, whereas our approach (SURE) successfully achieves sub-categorical level erasure. The qualitative results (right) further show that our approach (SURE) eliminates targeted categories (e.g., guns, blade weapons, musical instruments, and toys) completely.

To address the associated risks, initial efforts primarily focused on dataset filtering during training [4, 38], post-generation filtering [33], and inference-time guidance [35]. Dataset filtering removes undesirable content but requires significant human effort and expensive retraining. Post-generation filtering is prone to false positives and can be easily bypassed by users, whereas inference-time guidance provides a lightweight alternative but remains vulnerable to adversarial attacks [42]. An emerging alternative is *concept erasure* [2, 3, 10, 26, 41], which involves fine-tuning a pre-trained model to unlearn specific concepts described in a text prompt [15], preventing the generation of related image content without requiring full retraining. There are primarily two types of fine-tuning methods for concept erasure: Attention-based methods [11, 15, 26, 30, 45] and Output-based methods [2, 3, 10, 41]. Attention-based methods focus on modifying the attention mechanisms within models to eliminate undesirable concepts. In contrast, Output-based methods aim to optimize the output image by minimizing the difference between the predicted noise and the target noise.

While prior methods have demonstrated success in removing specific undesirable concepts, most are limited to erasing either a single concept (e.g., pistols) or a small set of closely related concepts (e.g., revolvers, rifles, and shotguns) at a time. However, these approaches often struggle to eliminate complete category (e.g., guns), including all subcategories, effectively. As a result, unwanted content may still be generated by the model (ref. Fig. 1). For instance, when erasing the category of guns, the model must automatically erase subcategories such as SMG, revolvers, pistols, rifles, and shotguns. While eliminating specific harmful content is important, it is equally critical to enable the removal of entire category (including all subcategories) to improve the safety and general applicability of text-to-image models. Some methods [22, 26] require individual concepts for multi-concept erasure, which may not be practical, as not all subcategories are

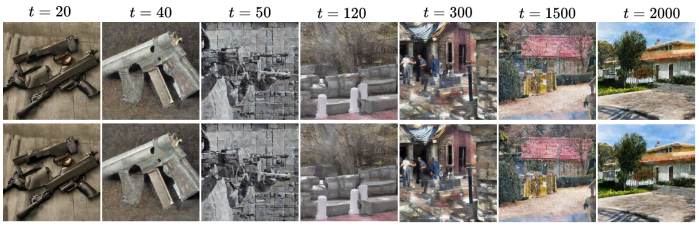

Fig. 2: Visual illustration of perturbed (top row) and original (bottom row) image samples used during the computation of the Lipschitz regularization in the context of the gun category erasure at different training steps. Perturbations are applied to the original inputs to encourage smooth behavior in the model's latent space around the gun category.

known and their number can be very large, making it computational expensive. Instead, we propose erasing entire subcategories from text-to-image diffusion models using only a single parent category as the target. We follow the WordNet 3.1 [28] ontology definitions of hypernymy and hyponymy to distinguish between categories (hypernyms) and subcategories (hyponyms) (ref. supplementary material for more details). In this context, a category refers to a grouping of subcategories. For example, "flower" is a category that encompasses multiple subcategories such as rose, tulip, and daisy.

To this end, we introduce **S**ubcategorical **U**nlearning via **R**egularized **E**rasure (SURE), a novel framework for removing entire subcategories from text-to-image diffusion models using only a single parent category prompt. SURE ensures the automatic elimination of multiple subcategories (e.g., SMG, revolvers, pistols, rifles, and shotguns) within a category (e.g., Guns) by erasing a single parent category (e.g., Guns). Moreover, SURE not only removes range of subcategories under the targeted category but also achieves a strong balance between Efficacy (the model erases only the intended target category) and Locality (the model retains its ability to generalize to unrelated categories). Our core design builds upon extending Lipschitz continuity [44] to diffusion-based text-to-image models for robust unlearning. We adapt Lipschitz regularization directly to the latent space of generated image samples, encouraging smooth generative behavior around the target category. Specifically, we generate several perturbed versions (see Fig. 2) of the given target category and train the model to minimize the difference between the original output and the outputs of these perturbed categories. This process smooths the model in targeted regions [9, 18], leading to the erasure of the target category (where the target region refers to an unseen concept space), while preserving the model's generalization performance. Furthermore, we incorporate a neutral eraser and a preservation objective, which guide the model in retain its behavior on neutral concepts while effectively erasing the undesirable subcategories (ref. Sec. 4.1). We extensively evaluate the proposed approach across three erasure tasks: object category removal, explicit content suppression, and artistic style elimination. Our method demonstrates

strong performance in erasing entire subcategories. It achieves a better balance between efficacy and locality compared to state-of-the-art (SOTA) methods.

The contributions of this work are summarized as follows: We propose SURE, a novel framework for entire subcategories erasure in pre-trained text-to-image (T2I) models. Unlike prior approaches that focus on individual concept or small concept sets, SURE removes entire subcategories of undesirable content using only a single parent category prompt. SURE employs Lipschitz regularization, along with a neutral eraser and a preservation objective, to strike a balance between effective forgetting of the target category and retention of unrelated, desirable content. Our joint formulation introduces a scalable approach to category Erasure in text-to-image models. We validate this framework across diverse tasks, including object category removal, artistic style elimination, and explicit content erasure, showing consistent improvements over state-of-the-art methods.

2 Related Works

Fine-tuning-based Concept Erasing: These methods [12, 27, 43] adapt pre-trained foundation models to eliminate specific target concepts by modifying their parameters, offering a scalable alternative to retraining from scratch. One class of these methods focuses on attention-based modifications. For example, Forget-Me-Not [45] employs attention re-steering by fine-tuning only the UNet to suppress intermediate attention maps related to the target concept. TIME [30] adjusts the linear projections for keys and values within cross-attention layers to redirect the influence of the unwanted concept toward a more neutral interpretation. MACE [26] applies a closed-form cross-attention refinement using multiple LoRA modules for efficient fine-tuning. MACE evaluates generality by using synonyms of the concepts, but this does not guarantee the evaluation of all subcategories. UCE [11] proposes a closed-form erasure solution while preserving non-target concepts through an explicit preservation objective. AdaVD [39] calculates the orthogonal complement of each cross-attention and uses the shift factor to adaptively adjust the erasure strength, enhancing effective prior preservation without compromising erasure efficacy. Another category of fine-tuning-based approaches directly optimizes model outputs to ensure the removal of unwanted content. ESD [10] maps the distribution of a target concept to that of a neutral or "null" prompt, effectively disassociating it from the generated image. Methods like, AGE [2], and AP [3] take a similar approach by mapping the target concept to either a neutral, null, or adversarial concept. Anti-Editing Concept Erasure [40] not only erases the target concept during generation but also filters it out during editing.

Adversarially Robust Concept Erasing: In the context of adversarial attacks, adversarial examples are carefully crafted inputs designed to manipulate models into producing unintended or harmful outputs. Recently, several methods have focused on enhancing the robustness of concept erasure in fine-tuned models. AdvUnlearn [47] improve robustness by alternately training the model for both concept erasure and adversarial resistance. Specifically, AdvUn-

learn [47] incorporates additional regularization to balance the removal of target concepts while preserving unrelated ones. UnlearnDiff [48] addresses optimization challenges in adversarial settings using Projected Gradient Descent to ensure effective erasure. RACE. [17] identifies adversarial text embeddings that can potentially reconstruct erased concepts and targets them for removal. Similarly, methods like AP [3] and AGE [2] identify sets of concepts most affected by parameter changes termed adversarial concepts and use these to guide stable and robust concept erasure while preserving unrelated generation capabilities. In STEREO author [37] adopts a two-stage strategy involving explicit min-max optimization: the first stage employs adversarial training, while the second uses an anchor-concept-based compositional objective to erase the target concept in a single fine-tuning pass. Most existing methods (ref. Fig. 1) fall short in achieving complete category (including all subcategories) erasure. Addressing this gap, our work introduces an output-based fine-tuning approach aimed at removing entire subcategories, marking an advancement in strengthening text-to-image (T2I) diffusion models against subcategory-level undesirable concepts.

3 Preliminaries

3.1 Denoising Diffusion Models

Diffusion models [13] represent a class of generative models capable of synthesizing high-quality, realistic images. These models consist of two main stages: the forward diffusion process, in which noise is incrementally introduced to an input image, and the reverse denoising process, where the model attempts to predict the noise component ϵ_t added in the forward process. Formally, given a sequence of T diffusion steps $x_0, x_1, \ldots, x_T$, the denoising process can be expressed as:

$$p_\theta(x_{T:0}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{1}$$

where the model is optimized by minimizing the difference between the actual noise ϵ and the noise estimate $\epsilon_\theta(x_t, t)$ provided by the denoising model θ at step t, as follows:

$$\mathcal{L} = \mathbb{E}_{x_0 \sim p_{\text{data}}, t, \epsilon \sim \mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \right]. \tag{2}$$

3.2 Latent Diffusion Models

Latent diffusion models (LDMs) [34] enhance the efficiency of the diffusion process by operating within a lower-dimensional latent space z, obtained via a pretrained variational autoencoder (VAE) comprising an encoder $f_\theta(\cdot)$ and a decoder D. During training, an image x is mapped to its latent representation $z = f_\theta(x)$, where noise is progressively added, yielding z_t with increasing noise levels over time. LDMs function as a sequence of denoising models, parameterized by θ, that estimate the noise component $\epsilon_\theta(z_t, c, t)$. This estimation is conditioned on both the timestep t and an auxiliary conditioning variable c,

which corresponds to the CLIP embedding of an input prompt. The optimization objective of LDMs is given by:

$$L = \mathbb{E}_{z_t \in \mathcal{E}(x), t, c, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, c, t)\|_2^2 \right]. \tag{3}$$

To enhance and control image generation, classifier-free guidance [14] is employed. This technique adjusts the probability distribution in favor of samples that align with an implicit classifier $p(c|z_t)$. During inference, the model is trained on both conditional and unconditional denoising, allowing the computation of both conditional and unconditional noise estimates. The final guided noise estimate $\tilde{\epsilon}_\theta(z_t, c, t)$ is computed as:

$$\tilde{\epsilon}_\theta(z_t, c, t) = \epsilon_\theta(z_t, t) + \alpha(\epsilon_\theta(z_t, c, t) - \epsilon_\theta(z_t, t)), \tag{4}$$

where $\alpha > 1$ is the guidance scale controlling the strength of conditioning. The inference process initiates with a Gaussian noise sample $z_T \sim \mathcal{N}(0, 1)$. Using the refined noise estimate $\tilde{\epsilon}_\theta(z_T, c, T)$, the model iteratively denoises z_T to obtain z_{T-1}. This iterative refinement continues until z_0 is reached, which is then transformed back to image space using the decoder $x_0 \leftarrow D(z_0)$.

4 Method

We aim to develop a framework that removes multiple subcategories using a single parent category from pretrained T2I diffusion models. To achieve this, we will fine-tune a new eraser model, θ', (ref. Fig. 3) with the objective of unlearning the target parent category while preserving the model's generalization to unrelated categories. We also identify related subcategories, such as synonyms, and enable the eraser model to eliminate these related subcategories as well.

4.1 Concept Erasing with a Neutral Eraser and Preservation

The objective of concept erasure is to remove an undesirable concept $c_e \in \mathbf{E}$, where $\mathbf{E}$ represents the set of concepts to be erased. c_e is a category. Given an associated textual description, we typically fine-tune a pretrained text-to-image diffusion model $\epsilon_\theta(z_t, \tau(c), t)$, resulting in an updated model $\epsilon_{\theta'}$ with parameters $\theta^{'}$, which produces an output $\epsilon_{\theta'}(z_t, \tau(c), t)$.

For simplicity, we represent the outputs as $\epsilon_\theta(\tau(c))$ and $\epsilon_{\theta'}(\tau(c))$. To erase specific visual category c_e from a pre-trained diffusion model, we adjust the model's predictions away from the erased category. This is achieved by mapping the erased category to a neutral concept [10, 30]. We use empty string " " as a neutral concept in our implementation. The objective is defined as:

$$\min_{\theta'} \underbrace{\mathbb{E}_{c_e \in \mathbf{E}} \left[\|\epsilon_{\theta'}(\tau(c_e)) - \epsilon_\theta(\tau(c_n))\|_2^2 \right]}_{\mathcal{L}_1} \tag{5}$$

From the objective function (ref. Eq. 5), the eraser model is trained to decrease the likelihood that the generated image contains undesirable category

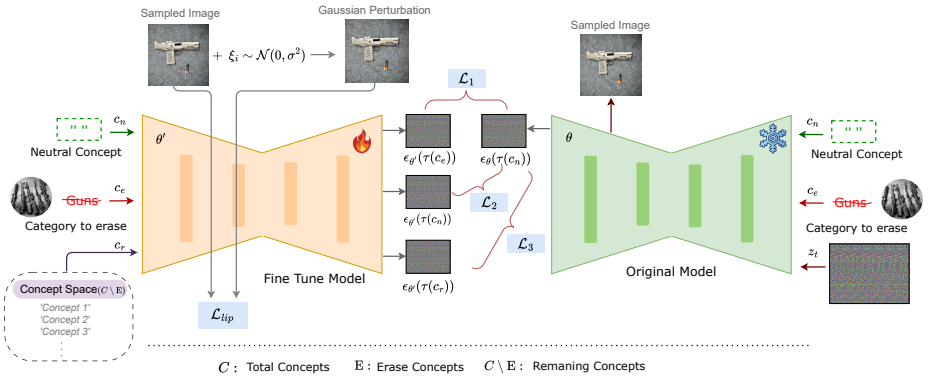

Fig. 3: Overview of SURE for complete category (including subcategories) Erasure. The optimization process aims to remove undesirable categorical visuals from a pre-trained diffusion model. The eraser model is trained using the loss term $\mathcal{L}_1$ to guide the diffusion process away from the target category. To preserve unrelated content, $\mathcal{L}_2$ regularizes the model's outputs to remain consistent for a neutral concept. Related subcategories are selected from the remaining concept space and mapped to the neutral concept, enabling effective erasure via $\mathcal{L}_3$. Additionally, Lipschitz regularization $\mathcal{L}_{lip}$ smooths the model's behavior with respect to input perturbations, facilitating stable erasure of the target category while maintaining the model's performance on non-target categories.

visuals by minimizing the L2 distance between $\epsilon_{\theta'}$ and ϵ_θ, thus successfully removing the category. The above objective effectively removes the undesirable category; however, it negatively impacts [2, 11] the model's ability to preserve other unrelated categories.

To retain unrelated concepts during concept erasure, prior methods [11, 26] explicitly preserve large sets of unrelated image–text pairs. However, this strategy is not practical when dealing with category erasure. For instance, when erasing the concept of "guns", it is infeasible to manually enumerate and preserve hundreds of unrelated categories (e.g., horses, microwaves, scissors, etc.). ACE [40] also highlights that traversing preserved concepts is challenging, as pre-trained models cover a vast general semantic space. To ensure the retention of unrelated categories, SURE utilize a regularization constraint (see Eq. 6) that encourages the model to maintain its behavior on neutral concepts while erasing undesirable ones [2]. Specifically, we add a regularization loss to ensure that the model's predictions for the neutral concept remain consistent before and after fine-tuning. This is achieved by minimizing the difference between the outputs of the fine-tuned and original models for the neutral concept (c_n). By anchoring the model at a known neutral point, the fine-tuning process is constrained from drifting too far from the original distribution, thereby helping preserve the model's general generation capabilities. The regularization term can be formulated as follows:

$$\underbrace{\|\epsilon_{\theta'}(\tau(c_n)) - \epsilon_{\theta}(\tau(c_n))\|_2^2}_{\mathcal{L}_2} \tag{6}$$

4.2 Related Sub Categories Erasure

Related subcategories are the acronyms for the subcategories that are closely related to the category. We use adversarial learning [2] to identify related sub categories within the concept space C. To this end, we optimize the following objective by maximizing $\mathcal{L}_3$ (ref. Eq. 7) with respect to c_r. In other words, it seeks a concept c_r whose output deviates the most from the neutral concept c_n. Since this is a maximization objective, we want the related sub categories to remain distinct from the null concept. In essence, the goal is to "maximize the differences between all related sub categories and the null concept after erasure." The maximization objective is given below:

$$\max_{c_r \in C \setminus \mathrm{E}} \underbrace{\|\epsilon_{\theta'}(\tau(c_r)) - \epsilon_{\theta}(\tau(c_n))\|_2^2}_{\mathcal{L}_3} \tag{7}$$

Here, the related sub categories is selected from the remaining concepts in the concept space, i.e., $C \setminus \mathrm{E}$, which represents the set difference between C and E. The equation above enforces the mapping of the related sub categories to the neutral concept, facilitating the erasure of the related sub categories with respect to the θ'. More details are provided in the supplementary material.

4.3 Lipschitz Regularization for Erasure

Given the category we wish to erase, we define an optimization problem to reduce the trained model's sensitivity to this category. We aim to achieve this by 'smoothing' the model's responses to the category concerning perturbations in the input image. By training the model to align its output for a target concept with its random perturbations, we can successfully reduce memorization of that target category (i.e., forget). This also links with Feldman (2020) [8], which observes that generalized models are often forced to memorize certain information; our objective is to protect the broader generalization while removing the relevant memorized data. Importantly, Lipschitz regularization reduces the model's sensitivity to small perturbations of concept variants. In category erasure, this is crucial, as the model must forget not only the explicit category prompt (e.g., "gun") but also subcategorical variants (e.g., "rifle," "SMG," "firearm").

We extend Lipschitz regularization [9] to be applied to intermediate image outputs generated by the erasure model to enforce smoothness in the latent space representation for small input perturbations. Let x be an image generated by the model conditioned on a category embedding c_e. We apply a small perturbation ξ to x, where ξ is independently sampled from a Gaussian distribution, i.e., $\xi \sim \mathcal{N}(0, \sigma^2)$. This results in a perturbed image $x' = x + \xi$. Both x and x'

are passed through the first-stage VAE encoder $f_{\theta'}(\cdot)$ to obtain their latent embeddings. The Lipschitz regularization loss is defined as follows:

$$\mathcal{L}_{lip} = \mathbb{E}\left(\frac{\|f_{\theta'}(x) - f_{\theta'}(x+\xi)\|_2}{\|x - (x+\xi)\|_2}\right) \approx \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|f_{\theta'}(x) - f_{\theta'}(x+\xi_i)\|_2}{\|\xi_i\|_2}\right), \quad (8)$$

where ξ_i is the i-th perturbation sample and N is the total number of perturbation samples. We do not compute gradients with respect to $f_{\theta'}(.)$ during optimization, instead these are treated as fixed. Further details are provided in the supplementary material.

4.4 Final Erasure Objective

Our final objective function is a combination of all four objectives: $\mathcal{L}_1$, $\mathcal{L}_2$,$\mathcal{L}_3$ and $\mathcal{L}_{lip}$. Each loss term plays a crucial role in category erasure from the pre-trained model, with specific functionalities. The overall objective is as follows:

$$\begin{aligned}\min_{\theta'} \; & \underbrace{\mathbb{E}_{c_e \in \mathbf{E}} \left\|\epsilon_{\theta'}(\tau(c_e)) - \epsilon_{\theta}(\tau(c_n))\right\|_2^2}_{\mathcal{L}_1} && + \underbrace{\left\|\epsilon_{\theta'}(\tau(c_n)) - \epsilon_{\theta}(\tau(c_n))\right\|_2^2}_{\mathcal{L}_2} \\ & + \max_{c_r \in C \setminus \mathrm{E}} \underbrace{\left\|\epsilon_{\theta'}(\tau(c_r)) - \epsilon_{\theta}(\tau(c_n))\right\|_2^2}_{\mathcal{L}_3} && + \underbrace{\mathbb{E}\left(\frac{\|f_{\theta'}(x) - f_{\theta'}(x+\xi)\|_2}{\|x - (x+\xi)\|_2}\right)}_{\mathcal{L}_{lip}}\end{aligned} \quad (9)$$

We minimize the objective $\mathcal{L}_1$ with respect to θ', ensuring that the distance between the erased model's output for the erased category and the original model's output for the neutral concept is minimized. This facilitates the erasure of the undesirable category. Similarly, the regularization term $\mathcal{L}_2$, minimized with respect to θ', ensures that the model preserves its capability on the remaining categories. Furthermore, minimizing $\mathcal{L}_3$ with respect to θ' guarantees the erasure of the related sub categories. The objective $\mathcal{L}_{\text{lip}}$ minimizes the difference in model outputs between the target category and its perturbed variants, thereby encouraging smoother model behavior. This regularization promotes the forgetting of target concepts while preserving the model's generalization.

5 Experiments

This section begins with quantitative experiments to evaluate SURE against several baseline methods, followed by qualitative comparisons and visualizations to further validate our approach. We also conduct ablation studies to analyze the effectiveness of our method in detail, including Component Analysis, the effect of perturbation strength, the contribution of Lipschitz regularization. Additional experiments including SURE's performance on multi-subcategory erasure, results with Stable Diffusion 2 and additional analyses in supplementary material.

5.1 Object Related Concepts

Datasets: To evaluate the effectiveness of SURE on object-type categories, we curate four distinct categories, each containing five target subcategories. We focus on three key properties for object-related category erasure. (a) *Singularity*: We select categories with distinct and well-defined subcategories—*Guns*, *Blade Weapons*, *Musical Instruments*, and *Toys*—as the target category for unlearning. (b) *Efficacy*: We assess whether subcategories can be effectively erased using simple prompts in the format "`A photo of {sub category}`". The specific subcategories for each category are: Guns: *SMG, Rifle, Pistol, Shotgun, Revolver*; Blade Weapons: *Knife, Sword, Spear, Shotel, Dagger*; Musical Instruments: *Trumpet, Guitar, Drums, Saxophone, Piano*; Toys: *Stuffed Toy, Doll, Lego Toy, Toy Car, Funko Toy*. We generate 50 prompts for each subcategories, resulting in a total of 250 prompts per category. (c) *Locality*: To ensure that erasing target concepts does not affect unrelated ones, we evaluate SURE on ten non-target classes from CIFAR-10 [19]. For each class, we generate 50 paraphrased prompts to simulate real-world use cases where prompts are often descriptive and may not explicitly mention the concept. For example, a paraphrased prompt for "dog" could be: "a Doberman Pinscher in a police vest, part of a K-9 unit." Our dataset curation uses WordNet to construct our object category erasure dataset. Further details regarding the dataset and its selection criteria are provided in supplementary material.

Evaluation Metrics: To assess the effectiveness of category erasure, we follow prior work [15,26] and utilize `GroundingDINO` [24] to determine whether the specified target category is present in the generated images. This enables us to measure the accuracy on erased category (Acc_E) and on non-erased (remaining) category (Acc_L). To calculate Acc_E, we use subcategories for each category. For example, during training, we designate the category 'Guns' as the target for erasure. In the evaluation phase, we generate images using subcategories such as Gun, SMG, Rifle, Pistol, Shotgun, and Revolver. We then apply GroundingDINO to detect the presence of these subcategories. The results are reported as the average percentage of images in which subcategories appear when we input prompts. A lower Acc_E indicates better erasure efficacy, while a higher Acc_L signifies stronger Locality. To provide a unified evaluation, we adopt the harmonic mean H between $(100-\text{Acc}_E)$ and Acc_L, following the metric proposed in [21,26].

Experiment Results: In Table 1, we present the results of SURE across four categories: Guns, Blade Weapons, Musical Instruments, and Toys. Notably, SURE achieves the highest harmonic mean across all four categories, demonstrating its effectiveness in erasing an entire category (including subcategories) using only the parent category label. This indicates a strong balance between effective removal of category and maintaining sufficient unrelated content in the locality. Specifically, SURE outperforms the second-best method by a relative percentage difference of 31.64%, 8.79%, 30.14%, and 25.96% across the guns, blade weapons, musical instruments, and toys categories, respectively. Some baseline methods [39, 40] demonstrate effectiveness in erasing individual instances of a

Table 1: Comparison of concept erasure methods across four categories: Guns, Bladed Weapons, Musical Instruments, and Toys. We report the following metrics: Acc_E — accuracy on erased concepts (↓), Acc_L — locality accuracy on non-erased (remaining) concepts (↑), and H — harmonic mean (↑). Notably, SURE achieves the highest harmonic mean across all four categories, demonstrating its effectiveness in erasing entire categories (including subcategories) using only the parent category label.

Method	Venue	Guns			Blade Weapons			Musical Instruments			Toys		
		Acc_E ↓	Acc_L ↑	H ↑	Acc_E ↓	Acc_L ↑	H ↑	Acc_E ↓	Acc_L ↑	H ↑	Acc_E ↓	Acc_L ↑	H ↑
ESD [10]	ICCV23	52.80	66.20	55.11	18.80	68.60	74.37	51.60	68.80	56.82	49.20	66.20	57.48
UCE [11]	WACV24	55.70	67.40	53.46	33.60	67.80	67.09	68.00	74.69	44.80	66.80	67.60	44.53
AP [3]	NeurIPS24	55.60	67.00	53.40	32.80	66.40	66.79	62.80	65.40	47.42	67.20	69.20	44.50
AGE [2]	ICLR25	49.60	68.60	55.92	26.00	68.60	71.19	65.20	66.20	45.61	74.00	67.00	44.50
AdaVD [39]	CVPR25	78.40	44.65	29.11	57.20	45.65	44.17	65.60	44.30	38.72	78.80	75.83	33.13
ACE [40]	CVPR25	76.00	70.22	35.77	52.80	64.60	54.54	73.20	67.00	38.28	77.20	65.60	33.83
Ours	ECCV26	10.00	67.20	**76.94**	2.40	64.60	**77.74**	8.40	66.40	**76.99**	14.80	66.40	**74.63**

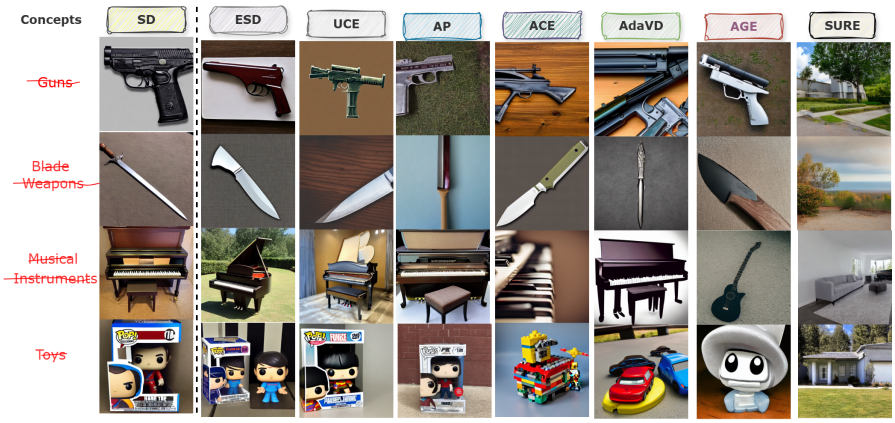

Fig. 4: Visual comparison of erasing effectiveness on object-related concepts. It is evident that the compared methods do not completely eliminate the entire target category. In contrast, SURE maps the erased objects to neutral concepts, ensuring effective removal of object-related concepts associated with the category.

concept, as reported in their respective studies. However, when tasked with removing an entire category, these methods do not consistently eliminate all related concepts within that category, as shown in Table 1. In contrast, SURE demonstrates robust categorical forgetting, as reflected by the low Acc_E and comparable Acc_L, achieving an optimal balance between erasure efficacy and locality. We additionally visualize examples of erasing different categories in Figure 4. As observed in the figure, the compared methods fail to erase the subcategories belonging to the parent category, with their outputs closely resembling the original images generated by the pre-trained Stable Diffusion model. In contrast, SURE successfully erases all subcategories associated with the same parent category.

Table 2: Comparison of text-to-image erasure models on the I2P dataset in the nudity erasure setting, evaluated using NER and FID metrics (lower is better ↓).

Model	Venue	NER-0.3 ↓	NER-0.5 ↓	NER-0.7 ↓	NER-0.8 ↓	FID ↓
CA [20]	CVPR 2023	13.84	9.27	4.74	1.68	20.76
ESD [10]	ICCV 2023	5.32	2.36	0.74	0.23	17.14
UCE [11]	WACV 2024	6.87	3.42	0.68	0.21	15.98
AP [3]	NeurIPS 2024	3.64	1.70	0.40	0.06	15.52
AGE [2]	ICLR 2025	5.06	1.53	0.32	0.04	14.20
SURE (Ours)	ECCV 2026	**1.70**	**0.47**	**0.11**	**0.01**	**14.02**

5.2 Explicit Contents Erasure

Datasets: To evaluate the proposed method on explicit concept erasure and the unlearning of Not-Safe-For-Work (NSFW) content, we utilize the Image-to-Prompt (I2P) dataset [35], which contains 4,703 prompts covering sexual, violent, and racist content. Unlike object concept erasure—where categories are clearly defined—we follow prior works and select "Nudity" as the category to be removed from the text-to-image (T2I) model. As suggested by Gandikota et al. [10], effective erasure of such content should be *global* and independent of specific keywords; that is, the model should remove the concept even when it is implied indirectly or referred to using synonyms. We generate images using all 4,703 prompts with the fine-tuned model to assess efficacy. We use the COCO-30K [23] validation set to evaluate locality to ensure that unrelated, safe content remains unaffected.

Evaluation Metrics: To evaluate the effectiveness of concept erasure and the presence of nudity in generated images, we employ NudeNet [1], which detects various types of exposed body regions. The model outputs multi-label predictions with associated confidence scores. We use the Nudity Exposure Rate (NER-k) metric [3], defined as the proportion of images in which any exposed body parts are detected with a confidence score above a given threshold k. To assess content locality and preservation, we report the Frechet Inception Distance (FID) on the COCO-30K validation set.

Experiment Results: We report the explicit content erasure performance in Table 2. It is evident from the results that our method outperforms all baselines, achieving Nudity Exposure Rates of 1.70, 0.47, 0.11, and 0.01 at thresholds of 0.3, 0.5, 0.7, and 0.8, respectively. This corresponds to relative percentage difference of 72.65%, 106%, 97.67%, and 120% compared to the second-best method at each threshold. These results demonstrate the effectiveness of our approach in preventing the generation of inappropriate content. Additionally, our method achieves an FID score of 14.02 on the COCO-30K validation set, indicating successful preservation of non-target concepts.

5.3 Artistic Style Erasure

Datasets: To assess the performance of our approach and baseline methods in removing multiple artistic styles from Stable Diffusion v1.4, we select five prominent artists: *Kelly McKernan*, *Thomas Kinkade*, *Tyler Edlin*, *Kilian Eng*,

Table 3: Comparison of CLIP and LPIPS scores across different methods, evaluated on five different artistic styles. Lower CLIP and higher LPIPS scores indicate better erasure performance. Additionally, higher CLIP scores on the remaining artists indicate better preservation performance.

Model	Venue	CLIP Score ↓	LPIPS Score ↑	CLIP Score ↑
ESD [10]	ICCV23	23.56 ± 4.73	0.72 ± 0.11	29.63 ± 3.57
CA [20]	CVPR23	27.79 ± 4.67	0.82 ± 0.07	29.85 ± 3.78
UCE [11]	WACV24	24.47 ± 4.73	0.74 ± 0.10	30.89 ± 3.56
MACE [26]	CVPR24	27.96 ± 4.22	0.60 ± 0.10	$\mathbf{31.52 \pm 2.91}$
AP [3]	NeurIPS24	21.57 ± 5.46	0.78 ± 0.10	30.13 ± 3.44
AGE [2]	ICLR25	22.44 ± 5.03	0.80 ± 0.12	30.45 ± 3.35
SURE (Ours)	ECCV 2026	$\mathbf{18.02 \pm 5.96}$	$\mathbf{0.85 \pm 0.10}$	30.91 ± 3.45

and *Ajin Demi Human*. For fine-tuning, only the artist's names are used as the target concepts. During evaluation, we employ a curated set of detailed prompts [2,3] tailored to each artist, rather than generic descriptions [10]. Each prompt is combined with five random seeds, resulting in the generation of 200 images per artist for every method.

Evaluation Metrics: To evaluate the removal of artistic styles, we use CLIP [31] to calculate the similarity between each generated image and its associated text prompt, where lower similarity indicates more effective erasure of the target style. Additionally, we utilize LPIPS [46] to quantify perceptual differences between images produced by the original Stable Diffusion model and those generated by edited models. A lower LPIPS score signifies less distortion. For erased concepts, a higher LPIPS score indicates more effective removal.

Experiment Results: The evaluation results are presented in Table 3. It is evident that our method outperforms existing approaches in both CLIP and LPIPS scores. Specifically, our method achieves an average CLIP score of 18.02, reflecting a 17.93% relative improvement over the second-best method, AP. Additionally, SURE obtains the highest LPIPS score of 0.85, indicating superior perceptual quality and more effective erasure of artistic styles compared to other baselines. Overall, our approach provides the best erasure performance among all the reported baselines while maintaining a comparable preservation score.

5.4 Ablation Experiments

Component Analysis. We evaluate the contribution of Lipschitz and the Neutral preservation loss (ref. Table 4 left). When the Lipschitz regularization is removed (w/o Lip), the model's ability to erase the target concept significantly degrades, as shown by a higher Acc_E (25.20). Removing the preservation objective (w/o Preservation) further reduces erasure performance ($Acc_E = 30.00$) and slightly affects the model's retention of non-target concepts ($Acc_L = 64.80$).
Effect of Number of Perturbation Samples (n). We analyze the sensitivity of SURE to the number of Gaussian noise samples used in the Lipschitz regularization. As shown in Table 4 (right), performance varies with different values of n. Increasing n improves the erasure quality, peaking at $n = 5$ with $Acc_E =$

Table 4: Ablation studies on the components of SURE (left) and sensitivity to the number of Gaussian perturbation samples (n) for the **Guns** category (right).

Method	Acc_E (↓)	Acc_L (↑)
SURE	**10.00**	**67.20**
w/o Lipschitz	25.20	65.80
w/o Preservation	30.00	64.80

Samples n	Acc_E (↓)	Acc_L (↑)
$n = 1$	19.60	65.40
$n = 3$	30.40	67.00
$n = 5$	**10.00**	**67.20**
$n = 7$	27.60	66.60

Table 5: Ablation study on differentiable selection methods for related concept retrieval (on Guns object level dataset).

Method	Acc_E (↓)	Acc_L (↑)
Gumbel-Softmax	10.00	67.20
Straight-Through Estimator	44.80	67.20
Entmax	38.40	67.00
Sparsemax	46.80	64.20

10.00% and $Acc_L = 67.20\%$. The Lipschitz is a Monte Carlo estimate of an expectation over perturbations. When n is very small (e.g., $n = 1$), this estimate has high variance, resulting in a noisy smoothing signal; as a consequence, the model can still retain information about the erased category, leading to higher Acc_E. Increasing the number of samples reduces this variance and provides a more stable smoothing effect, consistent with observations in [9]. Setting $n = 5$ offers the most stable behavior around the target category, which empirically yields the strongest forgetting. For larger n, the effective weight of the Lipschitz term becomes too dominate relative to the erasure and preservation losses. Because the loss contribution is kept fixed, adding more perturbation samples increases the overall strength of the smoothing constraint, placing pressure on the model to behave nearly identically across many perturbation directions around the target category. This "over-smoothing" restricts the optimization too strongly: the model is encouraged to maintain local invariance rather than shifting sufficiently toward the neutral concept required for full erasure ($\mathcal{L}_1$). Empirically, this leads to a decline in forgetting performance (higher Acc_E), even though locality (Acc_L) remains largely unchanged. Adding too many perturbation samples makes the Lipschitz regularizer so dominant that it starts to show less category-level erasure then the optimal perturbation samples n.

Differentiable Operators. To enable differentiable training and transform the discrete concept space into a continuous one, we utilize differentiable operators for learning the embedding distribution β. We conduct an ablation study using four different operators: Gumbel-Softmax, Straight-Through Estimator, Entmax, and Sparsemax. The results are presented in Table 5. Among these, Gumbel-Softmax provides the most efficient training performance by enabling a fully differentiable and continuous selection process for related concepts. The goal of $\mathcal{L}_3$ is to identify the most adversarial related concept i.e., the one whose output

Table 6: Effect of Gaussian noise level (σ) on erasure (Acc_E) and preservation (Acc_L) accuracies (on Guns object level dataset). Lower Acc_E and higher Acc_L indicate better performance.

Perturbation Strength (σ)	Acc_E (↓)	Acc_L (↑)
0.01	20.00	63.60
0.10	10.00	67.20
0.30	11.20	65.00

Table 7: Ablation study of Lipschitz Regularization strength on forgetting and preservation performance (on Guns object level dataset). Lower Acc_E and higher Acc_L indicate better performance.

Lipschitz Regularization (λ)	Acc_E (↓)	Acc_L (↑)
$\lambda = 0.01$	**10.00**	**67.20**
$\lambda = 0.1$	41.60	66.40
$\lambda = 1$	22.32	65.80

deviates maximally from the neutral concept. This is analogous to adversarial example selection, where the optimal perturbation is the one maximizing deviation rather than an average over several directions. A hard-max (implemented via low-temperature Gumbel–Softmax) directly approximates this objective and results in better erasure of subcategorical memorization. Furthermore, Table 5 shows that soft-average operators such as Entmax and Sparsemax, which distribute probability mass across several candidates, substantially weaken forgetting performance (higher Acc_E). This indicates that averaging gradients over multiple related concepts make the optimization less effective.

Effect of Perturbation Strength. In this ablation study, we examine the impact of perturbation strength on the erasure of the target semantic category and the preservation of non-target concepts. Specifically, we vary the amount of Gaussian noise added to the input image by adjusting the value of σ, which represents the standard deviation of the noise distribution. As shown in Table 6, the best performance is achieved when $\sigma = 0.1$, which provides a balanced level of perturbation for effective Lipschitz regularization.

Contribution of Lipschitz Regularization. In the following ablation study, we evaluate the contribution of Lipschitz Regularization. Specifically, we vary the regularization strength and observe its effect on the forgetting accuracy (Acc_E) and the preservation accuracy (Acc_L). As shown in Table 7, moderate regularization (e.g., $\lambda = 0.01$) yields the best trade-off between effective forgetting and generalization.

6 Conclusion

Our proposed novel approach, SURE, effectively addresses the task of removing entire sub categories from pre-trained text-to-image diffusion models using only a single parent category. It leverages Lipschitz regularization to promote smooth model behavior around target category and incorporates a neutral eraser and preservation objective to balance effective concept forgetting with the retention of unrelated content. SURE consistently outperforms existing methods in balancing erasure and preservation through comprehensive experiments across object category removal, explicit content suppression, and artistic style elimination. In future work, we aim to further enhance non-target concept preservation while maintaining the ability to remove entire categories effectively.

References

1. Bedapudi, P.: Nudenet: Neural nets for nudity classification, detection and selective censoring (2022), `https://github.com/notAI-tech/NudeNet`
2. Bui, A., Vu, T., Vuong, L., Le, T., Montague, P., Abraham, T., Kim, J., Phung, D.: Fantastic targets for concept erasure in diffusion models and where to find them. arXiv preprint arXiv:2501.18950 (2025)
3. Bui, A., Vuong, L., Doan, K., Le, T., Montague, P., Abraham, T., Phung, D.: Erasing undesirable concepts in diffusion models with adversarial preservation. NeurIPS (2024)
4. Carlini, N., Jagielski, M., Zhang, C., Papernot, N., Terzis, A., Tramer, F.: The privacy onion effect: Memorization is relative. Advances in Neural Information Processing Systems **35**, 13263–13276 (2022)
5. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 5253–5270 (2023)
6. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
7. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. Advances in Neural Information Processing Systems **35**, 16890–16902 (2022)
8. Feldman, V.: Does learning require memorization? a short tale about a long tail. In: Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing. pp. 954–959 (2020)
9. Foster, J., Fogarty, K., Schoepf, S., Öztireli, C., Brintrup, A.: Zero-shot machine unlearning at scale via lipschitz regularization. arXiv e-prints pp. arXiv–2402 (2024)
10. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2426–2436 (2023)
11. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5111–5120 (2024)
12. Gao, D., Lu, S., Zhou, W., Chu, J., Zhang, J., Jia, M., Zhang, B., Fan, Z., Zhang, W.: Eraseanything: Enabling concept erasure in rectified flow transformers. In: Forty-second International Conference on Machine Learning (2025)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Huang, C.P., Chang, K.P., Tsai, C.T., Lai, Y.H., Yang, F.E., Wang, Y.C.F.: Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In: European Conference on Computer Vision. pp. 360–376. Springer (2024)
16. Jiang, H.H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., Gebru, T.: Ai art and its impact on artists. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 363–374 (2023)
17. Kim, C., Min, K., Yang, Y.: Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In: European Conference on Computer Vision. pp. 461–478. Springer (2024)

18. Kravets, A., Namboodiri, V.P.: Zero-shot class unlearning in clip with synthetic samples. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6456–6464. IEEE (2025)
19. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
20. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22691–22702 (2023)
21. Lee, B.H., Lim, S., Chun, S.Y.: Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. arXiv preprint arXiv:2503.12356 (2025)
22. Li, G., Xiao, Y., Ji, J., Deng, K., Hui, B., Guo, L., Ma, X.: Sculpting memory: Multi-concept forgetting in diffusion models via dynamic mask and concept-aware optimization. arXiv preprint arXiv:2504.09039 (2025)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
24. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
25. Lu, S., Liu, Y., Kong, A.W.K.: Tf-icon: Diffusion-based training-free cross-domain image composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2294–2305 (2023)
26. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024)
27. Meng, Z., Peng, B., Jin, X., Jiang, Y., Dong, J., Wang, W., Tan, T.: Dark miner: Defend against unsafe generation for text-to-image diffusion models (2024)
28. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
29. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
30. Orgad, H., Kawar, B., Belinkov, Y.: Editing implicit assumptions in text-to-image diffusion models. In: IEEE International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 7030–7038. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.00649
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
32. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
33. Rando, J., Paleka, D., Lindner, D., Heim, L., Tramèr, F.: Red-teaming the stable diffusion safety filter. arXiv preprint arXiv:2210.04610 (2022)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)

35. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: 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)
36. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6048–6058 (2023)
37. Srivatsan, K., Shamshad, F., Naseer, M., Patel, V.M., Nandakumar, K.: Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. arXiv preprint arXiv:2408.16807 (2024)
38. StabilityAI: Stable diffusion 2.0 release (2022), `https://stability.ai/blog/stable-diffusion-v2-release`
39. Wang, Y., Li, O., Mu, T., Hao, Y., Liu, K., Wang, X., He, X.: Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. arXiv preprint arXiv:2412.06143 (2024)
40. Wang, Z., Wei, Y., Li, F., Pei, R., Xu, H., Zuo, W.: Ace: Anti-editing concept erasure in text-to-image models. arXiv preprint arXiv:2501.01633 (2025)
41. Wu, J., Le, T., Hayat, M., Harandi, M.: Erasediff: Erasing data influence in diffusion models. arXiv preprint arXiv:2401.05779 (2024)
42. Yang, Y., Hui, B., Yuan, H., Gong, N., Cao, Y.: Sneakyprompt: Jailbreaking text-to-image generative models. In: 2024 IEEE Symposium on Security and Privacy (SP). pp. 123–123. IEEE Computer Society (2024)
43. Yao, Q., Liu, Y., Wang, Z., Bian, Y., et al.: Erasing concept combination from text-to-image diffusion model. In: The Thirteenth International Conference on Learning Representations
44. Yoshida, Y., Miyato, T.: Spectral norm regularization for improving the generalizability of deep learning. arXiv preprint arXiv:1705.10941 (2017)
45. Zhang, E., et al.: Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2303.17591 (2023)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric (2018), `https://arxiv.org/abs/1801.03924`
47. Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., Liu, S.: Defensive unlearning with adversarial training for robust concept erasure in diffusion models. Advances in Neural Information Processing Systems **37**, 36748–36776 (2024)
48. Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., Liu, S.: To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In: European Conference on Computer Vision. pp. 385–403. Springer (2024)