Selective Fine-Tuning for Targeted and Robust Concept Unlearning

Mansi
Imperial College London
London, United Kingdom
m.-24@imperial.ac.uk

Avinash Kori
Imperial College London
London, United Kingdom
a.kori21@imperial.ac.uk

Francesca Toni
Imperial College London
London, United Kingdom
f.toni@imperial.ac.uk

Soteris Demetriou
Imperial College London
London, United Kingdom
s.demetriou@imperial.ac.uk

Abstract

Text-guided diffusion models are used by millions of users, but can be easily exploited to produce harmful content. Concept unlearning methods have emerged as a principled model-level defence, but state-of-the-art methods remain brittle: they are vulnerable to adversarial prompts, degrade generation quality for benign concepts, and lack interpretability. Concept localization approaches improve on concept unlearning through more targeted and interpretable selective fine-tuning, but current techniques are still limited: they are static, which results in suboptimal utility and weak robustness. In order to tackle these challenges, we propose TRuST (**T**argeted **R**ob**u**st **S**elective fine-**T**uning), a novel approach for dynamically localizing target *concept neurons* and unlearning them demonstrated with two new optimisation objectives and a selective fine-tuning method. We evaluate TRuST under both static and adaptive adversaries, and show experimentally, against a number of state-of-the-art baselines, that TRuST achieves up to 14× reduction in Attack Success Rate (ASR) across both adversary types, preserves generation quality (ΔFID = 0.016, CLIP = 30.95), and converges in up to 22× fewer fine-tuning steps than the state-of-the-art. Our method achieves robust and interpretable unlearning of not only individual concepts but also combinations of concepts and conditional concepts, without any task-specific regularization. Content Warning: This paper contains examples of unsafe and explicit visual content and language included solely for research and evaluation purposes.

CCS Concepts

• **Computing methodologies** → **Feature selection**; **Knowledge representation and reasoning**; **Computer vision**.

Keywords

Unlearning, Stable Diffusion, Alignment, Model Safety

ACM Reference Format:
Mansi, Avinash Kori, Francesca Toni, and Soteris Demetriou. 2026. Selective Fine-Tuning for Targeted and Robust Concept Unlearning. In *Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS '26), November 15–19, 2026, The Hague, Netherlands.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3830454.3846806

1 Introduction

Text-guided diffusion models are the most prevalent class of text-to-image (T2I) generative models and are used by millions of users to generate and distribute photorealistic images. However, their growing popularity has sparked ethical and safety concerns. T2I models are already exploited to generate harmful content, including explicit or extortionate images, biased outputs, and manipulative content aimed at influencing public opinion, such as during elections[47], thus eroding trust from digital media. Recent works [3] have identified 22 potential risks posed by T2I diffusion models. This behavior predominantly arises because these models are trained on diverse datasets that inadvertently include harmful or undesirable concepts.

Therefore, developing effective safeguards against such misuse poses an urgent challenge. Existing approaches, though effective in the removal of targeted concepts, are vulnerable to adversarial attacks [58] and tend to degrade the generation quality of the non-targeted concepts [42]. Importantly, they also undermine a critical aspect of harmfulness. Harmful concepts could be generated out of the combination of different simple harmless concepts, for example, concepts like "*child*" and "*beer*" could be harmless by themselves, but could be harmful when used in a combination such as "*child drinking beer*". Due to the compositional ability of these models [34], these models are capable of generating a harmful concept by combining benign concepts. Only recently [33] introduced a method (CoGFD) for addressing Concept Combination Erasure (CCE). CoGFD unlearns the harmful concept combination while preserving the benign constituent concepts. The method is effective, but only prevents accidental unsafe generations in very confined scenarios used in its logic graph, thus confining the adversarial robustness. It also relies on an LLM for generating the logic graph which is then used for formulating the unlearning loss function. Notably, CoGFD fine-tunes the entire model, making the process less targeted leading to suboptimal utility preservation. Robustly unlearning a concept combination without impacting related concepts demands precise editing of the model.

Prior research [1, 26, 30] established prominent results regarding the presence of groups of neurons and layers in the cross-attention (CA) layers of the T2I models, which primarily contribute towards the generation of a *concept*. Subsequent concept unlearning techniques leverage this *localization* for developing selective unlearning techniques with minimal impact on non-targeted concepts [1, 9].

CCS '26, The Hague, Netherlands

ACM ISBN 979-8-4007-2871-6/2026/11
https://doi.org/10.1145/3830454.3846806

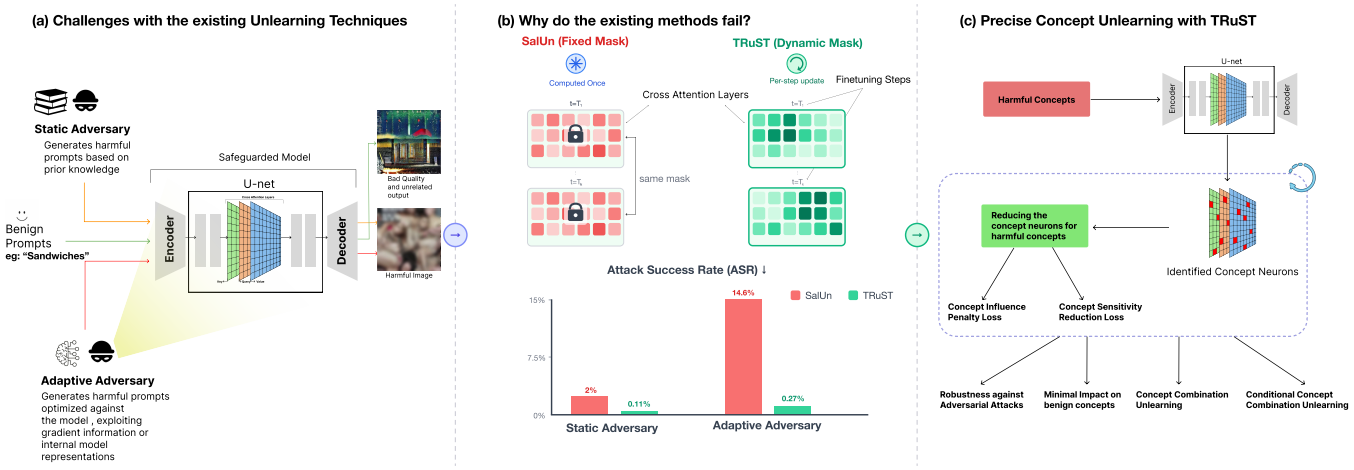


Figure 1: (a) Safety-unlearned T2I models remain vulnerable to both static and adaptive adversaries, while degrading image quality and faithfulness on benign concepts. (b) Existing methods rely on static neuron localization, which becomes outdated during fine-tuning and yields weak adversarial robustness. (c) TRuST dynamically re-identifies concept neurons at each step and suppresses them via CIP or CSR regularization, achieving robust unlearning under both adversary types.

However, these methods assume that the localization of relevant neurons remains static throughout fine-tuning. While this simplifies the parameter selection process, it overlooks the fact that salient neuron localization is dynamic and reconfigures as the model adapts to the unlearning objective. In our experiments, we observed that neuron activations and gradient magnitudes associated with a target concept change significantly at each step, suggesting that a fixed saliency determination becomes outdated early in training and results in suboptimal unlearning. Moreover, state-of-the-art (SOTA) unlearning techniques such as SalUn [9] rely on localization based on the predicted noise during the denoising process in diffusion models, which does not have any grounding based on the actual concept.

In our work, we found that predicted noise based grounding is not robust and requires more than 5x finetuning steps and 8x more wall clock time to achieve unlearning of the targeted concepts even if we improve it with dynamic localization of tunable parameters. Furthermore, we empirically observe that the existing methods are less effective in capturing how multiple concepts interact. Although interpretability-based approaches using model activations as the grounding[1, 7, 23, 46] are particularly useful for isolating features tied to discrete concepts, they tend to struggle to represent fine-grained relationships or contextual dependencies between concepts.

Robust unlearning requires disentangling *concepts* and their relations which demands precise editing of the model. In this work, we propose TRuST a selective and fine-grained neuron-level model editing approach. TRuST is based on two core ideas. Firstly, neuron saliency is dynamic. TRuST identifies the neurons responsible for encoding both simple and complex concepts and relations, using a new saliency mask driven by the alignment of the input prompt with the generated image which renders the approach more effective compared to existing noise-based masks. Secondly, unlearning effectiveness is improved with more direct disentanglement of target from benign concepts. TRuST performs a selective unlearning of specific concepts based on two novel objectives which can target the salient neurons with minimal impact on the generation of other non-targeted benign concepts or other safe concept combinations.

Contributions. Here we summarize our contributions:

(1) We propose a novel CLIP Score based gradient-based method for identifying the concept neurons which encode target semantic concepts within the cross-attention layers of T2I diffusion models. This enables fine-grained localization of not just individual concepts but also nuanced concept combinations.
(2) We introduce a selective fine-tuning method and two novel unlearning objective functions that *dynamically* update identified neurons to robustly unlearn target concepts while preserving the generation quality for unrelated concepts and selectively disentangling harmful/undesired associations between concepts.
(3) We embody these into a new end–to–end unlearning framework (TRuST) which we rigorously benchmark against SOTA concept unlearning techniques and characterize through ablation studies. Our results strongly support the improved effectiveness and precision of our approach.

2 Related Work

Most of the safeguarding T2I diffusion models can be broadly categorized into: data update methods and model-level interventions. From the data update perspective, several approaches aim to provide formal guarantees of unlearning, such as differential privacy-based unlearning and certified data removal techniques [5, 14]. While these methods offer strong theoretical guarantees, they typically require retraining the model from scratch after removing sensitive data from the training set [39, 48], making them computationally expensive and often impractical, especially given the compositional generalization abilities of generative T2I models [34].

An alternative direction is post-hoc safeguarding, which avoids retraining altogether. These methods operate by sanitizing the prompt before it is fed into the model [52], denying generation based on prompt analysis, or applying filtering or safety checks to the generated images [8, 54]. While these safeguards can be effective in certain scenarios, they remain vulnerable to adversarial prompt attacks that can bypass safety filters [11, 49, 53, 58].

Recently, a growing number of methods have been proposed from the perspective of model-level interventions in T2I diffusion models. These approaches can be broadly classified into two categories: model/knowledge editing and model steering. To provide a structured overview of these methods, we introduce an abstract comparative framework spanning seven dimensions, four methodology-driven (columns [1-4]) and three use case-oriented (columns [5-7]), as summarized in Table 1. This analysis highlights the key design choices and applications that differentiate existing works, and also shows how TRuST is positioned.

2.1 Model steering

Model steering techniques aim to guide the generation process at inference time by computing steering vectors, typically in the latent space, that suppress or amplify certain concepts without modifying the model's parameters. Some approaches[23, 55], operate in the text embedding space, either switching off offensive concept features or projecting embeddings away from unsafe directions. Others[21, 43], focus on computing steering directions in the latent noise or unconditional sampling space, often leveraging negative prompting to identify vectors that suppress undesired concepts during sampling. Recent methods inspired from mechanistic interpretability [1], rely on causal tracing to identify the cross attention (CA) layers within U-Net [40], that are most responsible for specific concepts by perturbing text embeddings, and use them to steer. Similarly, sparse autoencoders (k-SAE) have been used to isolate concept-specific attention heads[7], while CA saliency maps have been employed to extract concept-relevant parameters [32]. Although these steering methods offer practical advantages such as being non-destructive to model weights, they increase inference time by a multifold. Additionally, these methods are brittle and very sensitive to the choice of guidance/steering parameters.

2.2 Model/Knowledge Editing

This category of methods focuses on editing model parameters to reduce the likelihood of generating specific concepts by modifying only those parameters that store concept-relevant knowledge. Prior studies [11, 25] have demonstrated that such information is primarily embedded within the CA layers of DDPM [17] models. Several methods [11, 15, 31, 50, 56] leverage this insight by fine-tuning the entire set of CA layers to achieve concept unlearning. In contrast, other techniques[12, 13], perform closed-form post-hoc edits of CA weights, eliminating the need for retraining. Building on these findings, recent works have attempted to localize specific layers [1] or identify individual neurons [9] within the CA layers that are most responsible for encoding the undesired concepts. These identified components are then edited either via anchor-based techniques, replacing unsafe generations with aligned alternatives [9] or by ablating them until the undesired concept is suppressed [28]. While these methods have shown promising results, they often suffer from key limitations: they require extensive fine-tuning, are vulnerable to adversarial prompts, and may degrade the model's performance on unrelated (non-targeted) concepts [55].

While existing methods effectively unlearn individual concepts, limited progress has been made on combinations of concepts [33] or conditional associations [50]. Our approach, TRuST, addresses this by dynamically estimating concept neurons using a CA saliency-based method. Unlike SalUn[9], which relies on *static* neurons (which means concept neurons are identified at the beginning of the fine-tuning process and never updated), we update concept neurons continuously during the fine-tuning stage with an efficient batch-based sampling strategy. This enables scalability and effective unlearning of multiple concepts and relationships.

3 Preliminaries

Diffusion models. Diffusion models have emerged as a dominant paradigm in generative modeling, especially for high-fidelity image synthesis. These models define a generative process by learning to invert a fixed, stochastic forward noising process. Specifically, given a data distribution $x_0 \sim p_{\text{data}}(x)$, a forward process is constructed by gradually corrupting x_0 into a Gaussian noise sample x_T over T time steps. The noising process is defined as a Markov chain:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{1}$$

where $\alpha_t \in (0,1)$ denotes the variance schedule controlling the noise magnitude at each step t [17], and $\mathcal{N}$ is the normal distribution. Under the assumption that $x_T \sim \mathcal{N}(0,\mathbf{I})$, the goal of the reverse process is to denoise this Gaussian noise back to a data sample by estimating the conditional probabilities $p_\theta(x_{t-1}|x_t)$. In practice, this reverse denoising process is equipped with a denoiser network ϵ_θ typically parameterized using conditional U-Nets [40], trained to predict the added noise ϵ instead of directly modeling x_{t-1}:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t,t)\|_2^2\right], \tag{2}$$

where $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$.

Latent Text-to-Image Diffusion. To make diffusion models more computationally efficient and semantically controllable, recent works introduce latent-space diffusion models conditioned on text [39, 41]. Here, an autoencoder is first used to encode an image x into a latent representation $z = \mathcal{E}(x)$, and the diffusion process is performed on z rather than the pixel space, learning a posterior distribution. Additionally, the generative process is conditioned on a text embedding c, typically obtained from a pretrained language-image model such as CLIP [37] or T5 [38]. The training objective for the latent diffusion model follows:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_t,\epsilon,t,c}\left[\|\epsilon - \epsilon_\theta(z_t,c,t)\|_2^2\right], \tag{3}$$

where z_t denotes the noisy latent at step t.

Cross-Attention (CA).. A crucial component enabling semantic alignment between text and image in text-to-image diffusion models is the use of CA mechanisms. At each layer of the denoiser network ϵ_θ, the intermediate feature map attends to a text embedding $c \in \mathbb{R}^{L\times d}$, where L is the number of tokens and d is the embedding dimension, typically obtained from a pretrained language model like CLIP [37]. Formally, the CA is computed with the *query* $Q \in \mathbb{R}^{N\times d}$

Method	Weights Modification[1]	Training Free[2]	Anchor Free[3]	CN/Layer Targeted Tuning[5]	Multiple Concepts[5]	Concept Combination[6]	Conditional Concepts[7]
SAFREE [55]	✗	✓	✓	✗	✗	✗	✗
Concept Steerers [23]	✗	✓	✓	✗	✗	✗	✗
SLD-Max [43]	✗	✓	✗	✗	✗	✗	✗
TraSCE [21]	✗	✓	✗	✗	✗	✗	✗
LOCOEDIT [1]	✗	✓	✗	✓	✗	✗	✗
SAeUron [7]	✗	✓	✓	✓	✓	✗	✗
Concept Correctors [32]	✗	✓	✗	✓	✓	✗	✗
UCE [12]	✓	✓	✗	✗	✓	✗	✗
RECE [13]	✓	✓	✗	✗	✗	✗	✗
SSD [10]	✓	✓	✓	✓	✗	✗	✗
SLUG [4]	✓	✓	✗	✓	✗	✗	✗
ESD [11]	✓	✗	✓	✗	✓	✗	✗
SA [15]	✓	✗	✗	✗	✗	✗	✗
CA [25]	✓	✗	✗	✗	✗	✗	✗
MACE [31]	✓	✗	✗	✗	✓	✗	✗
SDID [27]	✓	✗	✗	✗	✗	✗	✗
All but One [19]	✓	✗	✗	✗	✗	✗	✗
ANT [28]	✓	✗	✓	✓	✓	✗	✗
SalUn [9]	✓	✗	✗	✓	✗	✗	✗
AdvUnlearn [56]	✓	✗	✓	✗	✗	✗	✗
Moderator [50]	✓	✗	✓	✗	✓	✓	✗
CoGFD [33]	✓	✗	✓	✗	✗	✓	✗
TRuST (Ours)	✓	✗	✓	✓	✓	✓	✓

Table 1: A structured overview of prior methods using a comparative framework across seven dimensions: four methodology-driven (columns [1–4]) and three use case-oriented (columns [5–7]). This analysis highlights key design choices and applications that distinguish existing works and contextualizes the positioning of our approach.

derived from the image latents (with N being the number of spatial locations in z_t), and the *key* $K \in \mathbb{R}^{L\times d}$ and *value* $V \in \mathbb{R}^{L\times d}$ are linear projections of the text embedding c, as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{4}$$

This formulation allows each spatial position in the image latent to attend to all tokens in the text prompt, enabling fine-grained semantic alignment. The resulting aligned features are then used in U-Net to guide the generation.

Saliency Maps. Saliency Maps are essentially a mask $\mathcal{M}$ defined on CA layers, representing the targeted parameters in the CA layers, which predominantly hold information regarding the concept under investigation c. Prior works [9] defined the saliency map based on the gradient of the difference of the predicted noise $\epsilon_\theta(z_t, c, t)$ and actual noise ϵ_θ at a random timestep t for a given concept prompt c and image latent z:

$$\mathcal{M} = \eta\left(\sum_{i=1}^{n}\left|\nabla_{\theta=\theta_0}(\epsilon_\theta(z_t, c, t) - \epsilon_\theta)\right| > \gamma\right), \tag{5}$$

where $\eta(g > \gamma)$ is an element-wise indicator that returns 1 if $g_i \geq \gamma$, and 0 otherwise; $|\cdot|$ is the element-wise absolute value; $\gamma > 0$ is a threshold.

4 Problem Statement and Motivation

4.1 Problem Statement and Definitions

Let θ denote the parameters of a pre-trained text-to-image diffusion model. The objective of machine unlearning (MU) is to compute an updated parameter set θ' such that the resulting model $f_{\theta'}$ satisfies the following properties:

(1) **Effectiveness (Concept Erasure)**: $f_{\theta'}$ reliably suppresses the generation of a target unsafe concept c_u.
(2) **Utility Preservation**: $f_{\theta'}$ retains its ability to generate high-quality, semantically faithful images for prompts $c_r \in C_r$, where C_r denotes a retained (benign) prompt set.
(3) **Erasure Efficiency**: the update from θ to θ' should be achieved with minimal computational cost and data requirements.

We define *effectiveness* in terms of robustness to both *static* and *adversarial* unsafe generations. This is quantitatively measured by Attack Success Rate(ASR), Unlearning Accuracy(UA)[7] and CLIPScore. Refer to Appendix **??** for a more detailed discussion on the metrics.

Utility measures the extent to which the model's generation capabilities for safe prompts are preserved. It is evaluated using FID, CLIPScore, TIFA, and the Retaining Accuracy (RA)[7](refer to Appendix **??**).

Finally, we define *Erasure Efficiency* as the ability to achieve low ASR (or high UA) using minimal unsafe data and compute budget.

Efficient methods require fewer fine-tuning steps and less exposure to unsafe concepts, making them more practical for real-world deployment in safety-critical settings.

4.2 Challenges

TRuST aims to tackle three major challenges in machine unlearning.

Challenge 1: Catastrophic Interference: Similar to catastrophic forgetting, standard fine-tuning introduces unintended concept interference, whereby changes made to remove an unsafe concept also perturb its neighboring representations in latent space. SOTA unlearning techniques, though good at unlearning the target concept, impact the overall utility of the model towards benign concepts as shown in Figure 2(a).

Challenge 2: Saliency shift during optimization: To address the above challenge others localized parts of the model to enable selective fine-tuning. Recent concept localization works such as SalUn (by)[9]) assume for simplicity that the set of concept neurons remains unchanged across all the finetuning steps. However, the set changes after every step. Figure 2(b) shows how the total number of concept neurons vary in SalUn after every finetuning step.

Challenge 3: Sample and Step Efficiency: Existing unlearning methods often require hundreds to thousands of update steps and substantial data exposure to converge, particularly when applied to semantically entangled or combinatorial concepts as shown in Figure 2(c).

5 Threat Model

Here we formalize the threat model governing our evaluation of concept unlearning in text-to-image (T2I) diffusion models. Let $f_{\theta'}$ denote a safety-edited model whose parameters θ' are obtained by applying a concept-unlearning algorithm to a pre-trained model f_θ, with the goal of suppressing the generation of a target unsafe concept c_u.

Adversary's Goal. The adversary seeks to recover the suppressed concept c_u from $f_{\theta'}$. Formally, the adversary crafts a set of adversarial prompts $\mathcal{A}(C_u)$, where C_u denotes the set of prompts that leads to the generation of c_u, such that:

$$\exists\, c \in \mathcal{A}(C_u) : D(I_c) = 1, \quad I_c = f_{\theta'}(c), \tag{6}$$

where $D(I_c) \in \{0, 1\}$ is a binary detector (*e.g.*, NudeNet when c_u = 'Nudity') that returns 1 if the erased concept is present in the generated image I_c. The effectiveness of such attacks is quantified by the Attack Success Rate (ASR), defined in Equation 7.

Adversary's Knowledge. For a realistic evaluation, we consider two distinct threat levels based on the adversary's knowledge of the deployed model $f_{\theta'}$ and the unlearning algorithm, a static adversary and an adaptive adversary.

Static Adversary. A static adversary has *no knowledge* of the deployed model $f_{\theta'}$, its architecture, parameters, or the specific unlearning algorithm applied. Adversarial prompts in $\mathcal{A}(C_u)$ are constructed *a priori*, independently of the target model, relying solely on semantic knowledge of the concept c_u or transferability from surrogate models. We consider the following static adversaries based on both well–established (facilitating comparisons) and SOTA works:

- **I2P** [43]: a black-box attack that generates adversarial prompts by paraphrasing and obfuscating the original concept c_u using a large language model, without any access to $f_{\theta'}$.
- **Ring-A-Bell** [49]: a model-agnostic red-teaming tool that extracts unsafe concept vectors and employs a genetic algorithm over a proxy visual encoder to construct adversarial prompts, requiring no access to $f_{\theta'}$.
- **MMA-Diffusion** [53]: a multi-modal attack that jointly perturbs text prompts and input images using a surrogate model to bypass prompt filters and post-hoc safety checkers, without querying $f_{\theta'}$ directly.

Adaptive Adversary. Despite the multiple approaches considered in the static adversary model, the robustness of any defense mechanism should be evaluated against adaptive adversaries as well, since we cannot assume the adversary did not or will not have access to the defense mechanism. This could happen through deliberate open-sourcing, reverse-engineering or model extraction attacks, online breaches or insider threats. We consider an adaptive adversary with *full knowledge* of the deployed model $f_{\theta'}$, including its architecture, weights θ', and the unlearning procedure used to produce them. To evaluate TRuST against this most challenging case, we consider SOTA attacks where adversarial prompts are optimized *specifically against* $f_{\theta'}$, exploiting gradient information or internal model representations. The adaptive adversary may inspect and exploit the released model, but its attack operates through prompts and does not modify the model parameters. While this is the primary adversary TRuST is designed to defend against, for completeness, we additionally evaluate robustness to post-release model-side recovery which we discuss separately in Section 8.5.

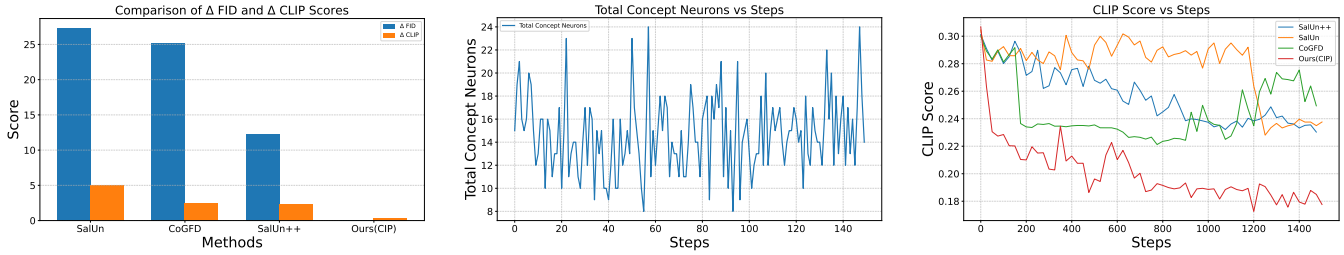
- **P4D** [6]: optimizes continuous adversarial prompts by minimizing the discrepancy between the noise predicted by $f_{\theta'}$ conditioned on the adversarial prompt and that predicted by an unprotected reference model, requiring white-box access to $f_{\theta'}$.
- **UnlearnDiffAtk** [58]: exploits the intrinsic classification capability of $f_{\theta'}$ itself to directly optimize discrete adversarial prompts via the PGD algorithm, with full gradient access to the target model.

Adversary's Capabilities. The adversary can craft prompts through obfuscation, paraphrasing, semantic indirection, or compositional manipulation of c_u. Both static and adaptive adversary effectiveness are formalized via the Attack Success Rate (ASR) metric:

$$\text{ASR} = \frac{1}{|\mathcal{A}(C_u)|} \sum_{c \in \mathcal{A}(C_u)} D(I_c), \tag{7}$$

where a lower ASR indicates stronger robustness of the unlearning method against the respective adversary type.

Defender's Goal. Given these two threat levels, the unlearned model $f_{\theta'}$ (defender) must simultaneously satisfy the three machine unlearning properties, i.e. *effectiveness*, *utility preservations*, and *efficiency* (Section 4.1) for all $c \in C_u \cup \mathcal{A}(C_u)$ in the ideal case. TRUST is designed to optimize for all three properties under both static



(a) Effect on the generation quality of retained concepts.

(b) The number of concept neurons changes after every finetuning step in SalUn.

(c) Number of finetuning steps taken to unlearn the target concept.

Figure 2: Challenges in MU for image generation. We compare CoGFD, SalUn and SalUn++ which is a stronger version of SalUn where the saliency map is recomputed after each finetuning step. Stable Diffusion 1.5 is considered as a reference for computation of Δ CLIP and Δ FID.

and adaptive adversaries, as demonstrated in our experimental evaluation (Section 7.2).

6 Methodology

TRuST tackles the above challenges and achieves fine-grained and robust unlearning of individual concepts and combinations of concepts through three main techniques: (a) identification of concept neurons; (b) design of concept unlearning objectives; and (c) an adaptive mask–guided fine-tuning process. An overview of TRuST is provided in Figure 3. We devise a novel concept neuron identification process that relies on dynamically estimating concept neurons with the help of gradients of an *alignment* objective. During fine-tuning, we apply two complementary regularization strategies. The first, Concept Influence Penalty (CIP): directly targets the concept neurons and encourages sparsity in the set of activated concept neurons by penalizing the number of high-influence parameters, which can be treated as *hard* unlearning. The second, Concept Sensitivity Reduction (CSR) is an indirect and more generalizable method which minimizes the sensitivity (updates the local curvature with hessian of the noise predictor) of the predicted noise to the concept neuron parameters, thereby weakening their influence, which improves the utility of the model when targeting concept combinations, resulting in a *soft* unlearning method. In essence, the first approach minimizes the number of concept neurons, while the second approach minimizes the sensitivity of concept neurons. As in all previous works in this domain, here, we assume the text-encoder is optimally trained and is robust to perturbations, and only work with a conditional diffusion to achieve unlearning. Below, we provide details on the discovery of concept neurons method, the two concept unlearning methods, and the selective fine-tuning process. The overall framework of our method is shown in Figure 3.

6.1 Discovering Concept Neurons

We first define a *concept neuron*. A concept neuron θ_{c_u} is a parameter within the CA projection matrices, specifically, the key, query, or value projections, that encodes information $\mathcal{I}_{c_u}$ relevant to a concept c_u. We consider a parameter θ_{c_u} to hold such information if perturbing its value results in a measurable change in the model's output, quantified by an alignment objective $\mathcal{L}c_u$. Unlike prior work [9], which relies on correlations within the noise predictor, we define $\mathcal{L}c_u$ as the CLIP score [37] between the concept prompt c_u and the image I_{c_u} generated by the model conditioned on c_u. Formally, the resulting concept neuron mask $\mathcal{M}_r(c_u)$ over projection type $r \in k, q, v$ is defined as:

$$\mathcal{L}_{c_u} = \text{CLIPScore}(I_{c_u},\ c_u) \tag{8}$$

$$\mathcal{M}_r(c_u) = \eta\left(\mathbb{E}[\left|\nabla_{\theta=\theta_0}\mathcal{L}_{c_u}\right|] > \gamma\right)_r, \tag{9}$$

where $\eta\,(g \geq \gamma)_r$ is an element-wise indicator function, returning 1 for the i^{th} element if $g_i \geq \gamma$, and 0 otherwise. The operator $|\cdot|$ denotes the element-wise absolute value. θ_0 is the current value of the parameter at the i^{th} location. The threshold γ is defined as: $\gamma = \xi \cdot \sigma_G + \mu_G$, where μ_G and σ_G represent the mean and standard deviation, respectively, of the gradient matrix G computed over a selected data subset. The gradient matrix G for a projection matrix r is given by: $G = \left|\nabla_{\theta=\theta_0}\mathcal{L}_c\right|_r$ and ξ is a tunable hyperparameter controlling sensitivity to outliers in the gradient distribution. We set $\xi = 2.0$ for the experiments (refer Appendix 8.1 to study the ablations for this selections). Also, we accumulate the gradients over the batch, instead of adding them individually.

Intuitively, a parameter is considered as a concept neuron, if by varying its values, we observe changes in L_c, *i.e.,* if the gradient $\frac{\partial \mathcal{L}_c}{\partial \theta} \neq 0$, it implies that this parameter has influence over the concept c_u and is used in generating the desired image I_{c_u}. The algorithm for computing the concept neuron mask is shown in Algorithm 1.

6.2 Concept Unlearning Objectives

TRuST leverages the above method for unlearning targeted concepts. TRuST's core idea for selective *hard- or soft*-unlearning is to minimize the number or sensitivity of concept neurons as detailed before. We devise $\mathcal{L}_{\text{CIP}}$ for hard unlearning, which directly minimizes the number of targeted parameters, encouraging sparsity and $\mathcal{L}_{\text{CSR}}$ for soft unlearning, which minimizes the sensitivity of these parameters by minimizing gradients of targeted parameters.

Concept Influence Penalty (CIP). CIP aims at directly minimizing the total number of concept neurons for the targeted concept c_u, encouraging sparsity, while preserving the influence of non-targeted concepts. Formally,

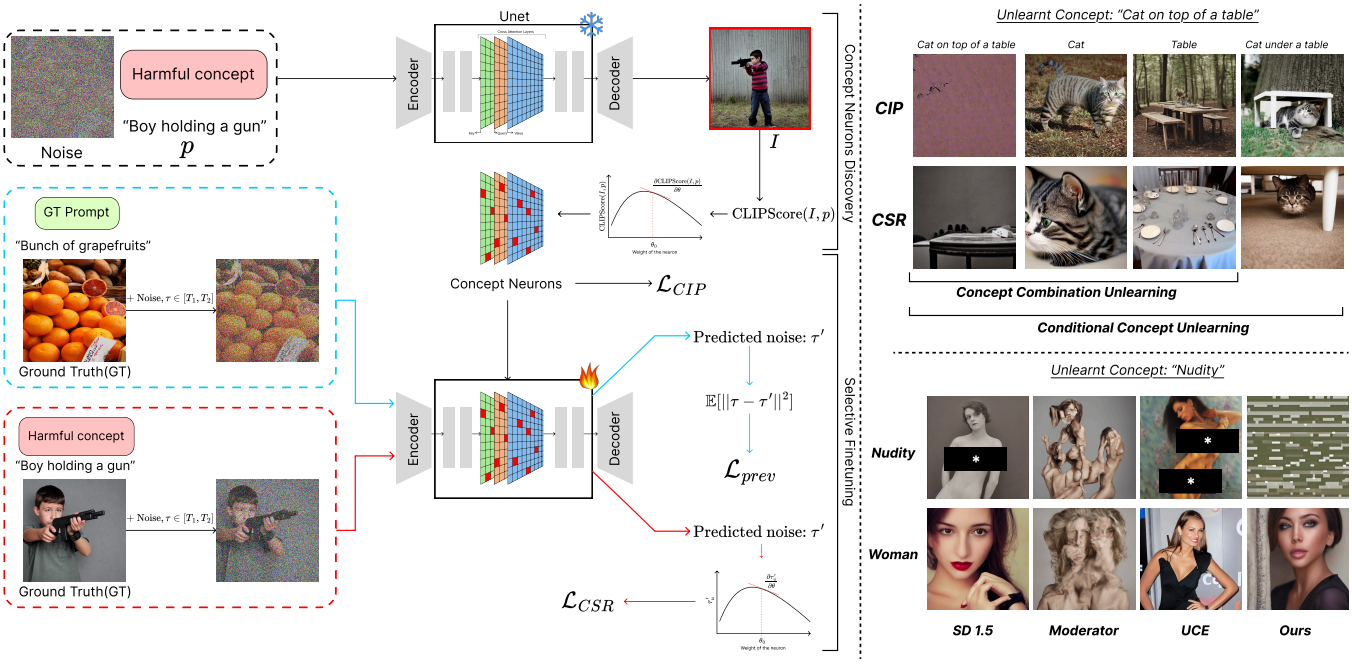


Figure 3: Overview of TRuST: The pipeline (left), depicts the concept neurons discovery and selective finetuning with both CSR and CIP. The right half showcases TRuST's ability to unlearn both concept combinations and conditional concepts, along with comparisons against well established concept erasure methods for "Nudity" unlearning against adversarial prompts (P4D [6]). Sections of image with "*" have been intentionally hidden for safety purposes.

Algorithm 1 Compute Concept Neuron Mask $M_r(c_u) \leftarrow$ ComputeMask

Require Concept c_u, model weights θ_0, projection matrices $r \in \{k, q, v\}$, threshold scale ξ
Ensure Concept neuron mask $M_r(c_u)$
  Generate I_{c_u} using the model conditioned on c_u
  Compute CLIP score: $L_{c_u} \leftarrow \text{CLIPScore}(I_{c_u}, c_u)$
  Compute gradient: $G \leftarrow |\nabla_{\theta=\theta_0} L_{c_u}|$
  Compute $\mu_G \leftarrow \mathbb{E}[G]$; $\sigma_G^2 \leftarrow \mathbb{E}[(G - \mu_G)^2]$
  Set threshold: $\gamma \leftarrow \xi \cdot \sigma_G + \mu_G$
for $r \in \{k, q, v\}$ **do**
  for elements i in G **do**
    if $\mathbb{E}[|\nabla_{\theta_i=\theta_0} L_{c_u}|] > \gamma$ **then**
      $M_r(c_u)_i \leftarrow 1$
    else
      $M_r(c_u)_i \leftarrow 0$
    end if
  end for
end for
Return $M_k(c_u), M_q(c_u), M_v(c_u)$

$$\mathcal{L}_{\text{CIP}} = \beta_{\text{CIP}} \left(\sum_r \sum_i \eta \left(M_r(c_u)_i = 1 \right) \right) + \mathcal{L}_{\text{prev}}, \tag{10}$$

where $\eta\,(X_i = 1)$ is an element-wise indicator function; and β_{CIP} is a regularization hyperparameter. Note that if θ is shared across many concepts or used in other attention heads, solely minimizing $\mathcal{L}_{\text{CIP}}$ might lead to unintended forgetting. This is because latent text-to-image diffusion models use classifier-free guidance (CFG) where the probability of the image latent and the probability of the image latent conditioned on the concept, are jointly parametarized by θ [18]. To mitigate the effects of unintended forgetting, we add a preservation loss $\mathcal{L}_{\text{prev}}$ to the objective. The preservation loss is the expected value of square of the difference between the predicted noise $\epsilon_\theta(z_t, c_p, t)$ and actual noise ϵ_θ for concepts which we wish to preserve $C = \{c_p | c_p \in \text{concepts to preserve}\}$:

$$\mathcal{L}_{\text{prev}} = \mathbb{E}[||\epsilon_\theta(z_t, c_p, t) - \epsilon_\theta||^2] \tag{11}$$

Concept Sensitivity Reduction (CSR). CSR weakens model sensitivity to concept-specific parameters by minimizing the gradient of the predicted noise $\epsilon_\theta(z_t, c_u, t)$ with respect to θ:

$$\mathcal{L}_{\text{CSR}} = \beta_{\text{CSR}} \log \left(\left| \frac{\partial \epsilon_\theta(z_t, c_u, t)}{\partial \theta} \right| \right) + \mathcal{L}_{\text{prev}}, \tag{12}$$

where β_{CSR} is a regularization hyperparameter. During backpropagation, the gradient updates incorporate second-order information indirectly, with the underlying Hessian of the noise prediction loss governing the local curvature and sensitivity to parameter changes.

The computed gradient $\frac{\partial \epsilon_\theta(z_t, c_u, t)}{\partial \theta}$ is very small (of the order of 1e-11), therefore, in order to bring it in the range for effective finetuning(0.1-0.01), we take the logarithm of the absolute value of the gradients. An example of the average gradients computed per head in the key and value cross attention layers is shown in Figure 4. When this gradient value is back-propagated, the update values of

the parameters of the model therefore becomes a double derivative of the noise or the second order derivative update. This second order partial derivative value can be computed using a Hessian.

CSR and CIP are alternative strategies for weakening concept-specific influence in diffusion models. While CSR continuously minimizes the sensitivity of the model output to concept-related parameters by penalizing their gradients, CIP takes a more discrete approach by directly penalizing the presence (cardinality) of high-influence parameters identified through thresholded gradients. Both operate over the same set of concept neurons but differ in formulation: CSR softens influence across all relevant parameters, whereas CIP sparsifies by targeting only those with significant impact. Thus, CSR offers a smooth, generalizable suppression mechanism, while CIP provides a sharper, sparsity-driven alternative. Due to the soft optimisation of CSR, $\mathcal{L}_{\text{CSR}}$ achieves better performance for unlearning on more complex policies, including concept combinations.

6.3 Dynamic Mask-Guided Finetuning

Finally, in order to target/penalize the most contributing parameters during the fine-tuning process, we selectively fine-tune just the concept parameters. Interestingly, we notice that the set of discovered concept neurons changes after every update of the model parameters during fine-tuning. Therefore, instead of committing to a set of concept neurons/parameters identified before finetuning the model like in prior work [9], we dynamically readjust to the new set of discovered concept neurons by recomputing the neuron masks $\mathcal{M}_r(c_u)$ after every parameter update step to account for representation shift during unlearning. This way TRuST ensures that the model continually targets the most relevant representations of the concept throughout training and avoids overfitting to a static or outdated subset of neurons, detailed in Algorithm 2.

Algorithm 2 Selective Finetuning for Concept Unlearning

Require Concept c_u, preservation concepts $C = \{c_p\}$, model parameters θ, loss type $\mathcal{L} \in \{\text{CSR}, \text{CIP}\}$, learning rate α, number of steps T
Ensure Updated model parameters θ
for $t = 1$ to T **do**
  $M_r(c_u) \leftarrow \textsc{ComputeMask}c_u, \theta$
  Sample latent $z_t \sim E(x)$ and noise $\epsilon \sim \mathcal{N}(0, 1)$
  Compute predicted noise: $\hat{\epsilon}_\theta \leftarrow \epsilon_\theta(z_t, c_u, t)$
  $L_{\text{prev}} \leftarrow \mathbb{E}_{c_p \in C}\left[\|\epsilon_\theta(z_t, c_p, t) - \epsilon_\theta\|^2\right]$
  if $\mathcal{L}$ is CSR **then**
    $L_{\text{CSR}} \leftarrow \beta_{\text{CSR}} \cdot \log\left(\left\|\frac{\partial \epsilon_\theta(z_t, c_u, t)}{\partial \theta}\right\|\right)$
    $\mathcal{L}_{\text{total}} \leftarrow L_{\text{CSR}} + L_{\text{prev}}$
  else if $\mathcal{L}$ is CIP **then**
    $L_{\text{CIP}} \leftarrow \beta_{\text{CIP}} \cdot \sum_r \sum_i \eta\,(M_r(c_u)_i = 1)$
    $\mathcal{L}_{\text{total}} \leftarrow L_{\text{CIP}} + L_{\text{prev}}$
  end if
  Parameter update: $\theta \leftarrow \theta - \mathcal{M}_r \cdot \alpha \cdot \nabla_\theta \mathcal{L}_{\text{total}}$
end for
Return θ

7 Experiments and Results

7.1 Evaluation Setup

Model and Datasets. We perform experiments on Stable Diffusion v1.5 [39], trained on LAION-Aesthetics v2 5+(a subset of high quality images from the LAION 5B [45] dataset (containing NSFW images)). We chose this version because (a) it is responsible for generating the most harmful concepts and images [51], and (b) it has been widely used in past T2I jailbreaking and safeguarding works and allows for fair comparisons. For the preservation loss, we randomly used 1000 real-world images from MS COCO 30k validation dataset [29]. For the evaluation on retained concepts, we used a subset of the MS COCO 30k validation dataset [29] after removing the targeted concept from it. For simple concept unlearning, we perform unlearning on harmful concepts like "Nudity" and evaluate on the I2P dataset [43].

Baselines and Metrics. We compare TRuST against SOTA model editing and steering techniques on simple concept unlearning, multi-concept unlearning and scalability, and concept combination tasks, including CoGFD and SalUn. For the novel task of *Conditional Concept Unlearning*, where no prior baselines exist, we report standalone performance to demonstrate TRuST's effectiveness. Evaluation metrics include ΔFID [16] for photorealism, and CLIP [37] and TIFA [20] scores to measure semantic alignment with the prompt.

7.2 Evaluation Results

Robustness against adversarial attacks. Model editing approaches for safety must be effective not only against accidental unsafe generations but also against deliberate adversarial manipulation. We evaluate TRuST against both static and adaptive adversaries (Section 5).To ensure a fair and meaningful comparison against other works, we perform unlearning on the common harmful concept "Nudity". We use NudeNet [2] as the detector $D(\cdot)$. A generated image is flagged as unsafe if the image is classified under any class other than `*_COVERED`, `FACE_MALE`, `FACE_FEMALE`, `FEET_EXPOSED`, or `BELLY_EXPOSED`. We compare against the Attack Success Rate (ASR) [13] of different adversarial prompt generation techniques.

From Table 2[1], we can observe that TRuST outperforms all baselines, across both model editing and model steering techniques across every adversary type. Against **static adversaries**, TRuST (CIP) reduces ASR to 0.11% on I2P, 0.83% on Ring-A-Bell,and 6.2% on MMA-Diffusion, corresponding to a 45% ASR decrease over SalUn [9], and reductions of 10.57% and 14.33% over the strongest respective baselines on Ring-A-Bell and MMA-Diffusion. Against **adaptive adversaries**, TRuST (CIP) limits ASR at 0.27% on P4D and 1.18% on UnlearnDiffAtk, outperforming the strongest baseline by 14.33% and 18.52% respectively.

Both CIP and CSR losses outperform existing baselines, with CSR performing slightly better overall. CIP directly suppresses concept neurons of the unlearned concept, making it more adversarially robust, whereas CSR minimizes concept contributions while allowing neural reuse, making it less robust on individual targeted concepts.

[1] N/A is added for baselines where the corresponding computation was not supported by the metric or there were incompatibility issues

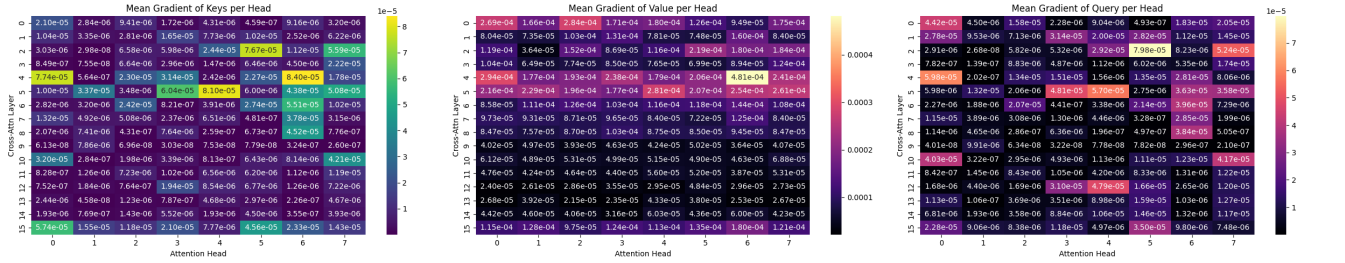


Figure 4: Example of average gradients computed per head for key and value cross-attention layers.

Table 2: Comparison of TRuST with SOTA baselines. Bold indicates best score in column.

Method	Weights Modification	Training -Free	Preservation integrated	I2P ↓	P4D ↓	Ring-A-Bell ↓	MMA-Diffusion ↓	UnlearnDiffAtk ↓	ΔFID ↓	CLIP ↑	TIFA ↑
SD-v1.5	N/A	N/A	N/A	0.179	0.989	0.835	0.968	0.797	0.00	**31.3**	**0.813**
SLD-Medium [43]	No	Yes	No	0.142	0.934	0.646	0.942	0.648	4.46	31.0	0.782
SLD-Strong [43]	No	Yes	No	0.131	0.861	0.620	0.920	0.570	7.69	29.6	0.766
SLD-Max [43]	No	Yes	No	0.115	0.742	0.570	0.837	0.479	12.04	28.5	0.720
SAeUron [7]	No	Yes	Yes	0.024	N/A	0.457	0.417	0.197	0.33	30.89	0.749
SAFREE [55]	No	Yes	Yes	0.034	0.384	0.114	0.585	0.282	0.70	31.1	0.790
Concept Correctors [32]	No	Yes	No	0.063	N/A	0.175	0.193	N/A	7.13	30.81	0.801
UCE [12]	Yes	Yes	Yes	0.103	0.667	0.331	0.867	0.430	1.28	30.2	0.805
RECE [13]	Yes	Yes	Yes	0.064	0.380	0.134	0.675	0.655	1.03	30.9	0.787
ESD [11]	Yes	No	No	0.140	0.750	0.528	0.873	0.761	1.47	30.7	0.805
SA [15]	Yes	No	No	0.062	0.623	0.329	0.205	0.268	7.62	30.6	0.776
CA [25]	Yes	No	No	0.178	0.927	0.773	0.855	0.866	7.41	31.2	0.805
MACE [31]	Yes	No	Yes	0.023	0.146	0.076	0.377	0.176	0.09	29.4	0.711
SDID [27]	Yes	No	No	0.270	0.931	0.696	0.907	0.697	6.28	30.5	0.802
SalUn [9]	Yes	No	Yes	0.02	0.013	0.208	0.264	0.320	27.31	26.2	0.735
AdvUnlearn [56]	Yes	No	Yes	0.26	0.200	0.160	0.273	0.211	2.06	28.6	0.799
TRuST (CSR loss)	Yes	No	Yes	0.0013	0.0046	0.0093	0.077	0.0175	0.030	30.43	0.811
TRuST (CIP loss)	Yes	No	Yes	**0.0011**	**0.0027**	**0.0083**	**0.062**	**0.0118**	**0.016**	30.95	0.808

Further differences are explored in Appendix **??**. Illustrative visual examples can be found in the Appendix **??**.

Preservation of non-targeted concepts. While effective, in real applications, TRuST should also preserve the utility of the generation on benign (non-targeted concepts). To comprehensively assess the impact of TRuST on model utility, we evaluate both image quality (photorealism) and semantic fidelity. Specifically, we report ΔFID and measure fidelity using CLIP and TIFA scores. From Table 2, we observe that TRuST incurs negligible change on the quality of generation on non-targeted concepts, with the change being as low as 0.016. This result, along with being better than any of the existing baselines, is also considerably better than SalUn, which assumes that the set of concept neurons stays the same throughout the fine-tuning process. TRuST also surpasses the existing baselines on the TIFA metric by achieving text-to-image fidelity very close to the original model. The CLIPScores also see a very small degradation, implying preserved semantic similarity with the prompt. Illustrative visual examples can be found in the Appendix **??**.

Combination/Conditional Concept Erasure. We evaluate the effectiveness and utility of TRuST on the complex unlearning tasks of *Concept Combination Erasure* (CCE) (Figure 5), and the novel task of *Conditional Concept Erasure* (CoCE) (Figure 6).

Concept Combination Erasure (CCE). CCE refers to the task of unlearning combinations of concepts while preserving the individual concepts themselves. Figure 5 shows the CLIP score variance of concept combinations (x-axis) and individual concepts (y-axis) across fine-tuning steps (indicated by dot darkness). A method (e.g., SalUn) that shows declining CLIP scores for individual concepts over time performs poorly on the CCE task. As illustrated, TRuST method consistently retains individual concept alignment across fine-tuning, outperforming even the state-of-the-art CoGFD [33]. Moreover, our approach achieves effective unlearning of concept combinations in nearly half the number of steps required by CoGFD.

Conditional Concept Erasure (CoCE). Figure 6 presents TIFA scores for the conditional prompt "*Boy holding a Gun*". TRuST method effectively unlearns this concept combination, while preserving semantically related prompts like *"Boy holding a Gun"*. High TIFA scores for individual concepts (e.g., "*Boy*", "*Gun*") and unrelated prompts indicate strong selectivity and preservation, confirming that unlearning does not spill over to adjacent concepts. Note that TIFA scores of 1.0 for individual prompts are expected as they measure basic presence detection. Additional examples illustrating this behavior can be found in the Appendix **??**.

Unlearning stylistic concepts. To examine applications of TRuST in copyright policies, we further experiment with removing stylistic

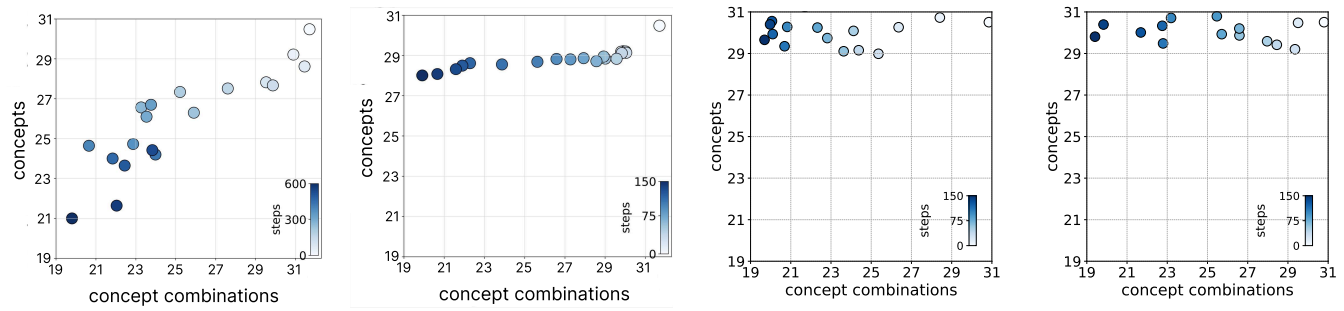


Figure 5: Average CLIP scores and number of finetuning steps. TRUST achieves better concept combination erasure (lower scores for targeted combinations) while better preserving individual concepts (higher scores for the individual concepts), in less steps on average.

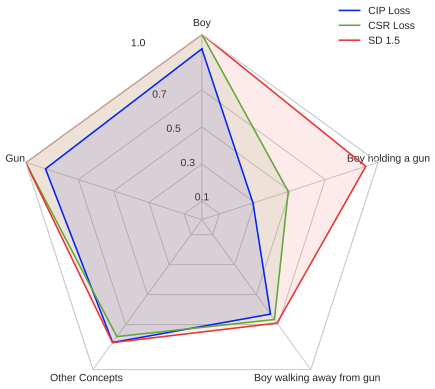


Figure 6: TIFA comparison for the conditional prompt "Boy holding a Gun".

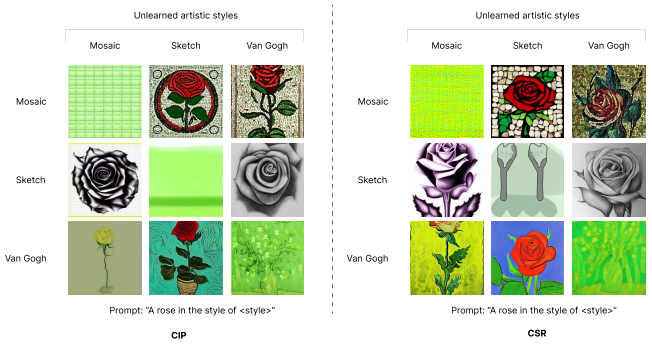


Figure 7: The figure shows some qualitative results of unlearning artistic styles of: "Mosaic", "Sketch", and "Van Gogh" individually; and how unlearning one artistic style has minimal impact on other styles, for both CIP and CSR loss.

concepts from the Unlearn Canvas dataset [57]. The results from CIP and CSR unlearning are presented in Figure 7. The figure shows that both CIP and CSR loss functions are equally effective in unlearning artistic styles without impacting other artistic styles. Moreover, Table 3 presents a quantitative result of the efficiency in unlearning an artistic concept in terms of average of UA and RA, against other unlearning methods.

We also compute the ΔFID scores for our model edited using CIP and CSR values to be 0.1 and 0.03 respectively, showing minimal effect on the generation quality of the model for non-targeted concepts. We used the real world images from the COCO dataset as the grounding for computing the FID scores.

Scalability and Erasure Efficiency. We evaluate the impact of enforcing multiple concept erasures (both simple and combinations) on retained concepts. Similar to SAeUron, we compute the average of the Unlearning Accuracy (UA) and Retaining Accuracy (RA) with the number of concept erasures and compare it against SOTA baselines as shown in Figure 8. Our method outperforms all baselines and almost perfectly retains the non-targeted concepts

Table 3: Comparison of overall accuracy, inference time, memory used during inference, and storage memory of the CIP and CSR techniques on one A100 GPU.

# Method	$\frac{UA+RA}{2}$ ↑	Time (s) ↓	Memory (GB) ↓	Storage (GB)↓
CIP	98.23	2.03	3.59	0.0
CSR	100	2.05	3.59	0.0

and unlearns the targeted concepts even when we increase the number of concepts to erase. To test the limits, we further test the UA and RA until 10 concepts being removed using TRUST. We observe that the unlearning accuracy eventually drops to 90.6% however the retaining accuracy still remains good at 97.2%, which is better than the retaining accuracy which the other methods lost at unlearning merely 6 concepts. We borrow the 10 set of concepts to unlearn from the UnsafeBench Dataset [36] appended with a few other general harmful concepts, namely {sexual, hate, violence, harassment, shocking, self-harm, blood, gun, alcohol, knife}.

For multiple concept combination erasures, we evaluate the impact on other concepts by measuring the CLIPScore, FID, and TIFA

# Concepts	CLIP Targeted ↓	CLIP Non-targeted ↑	TIFA ↓	ΔFID ↓
1	19.34	30.43	0.811	0.03
2	20.67	30.04	0.833	0.14
3	21.15	30.77	0.780	0.26
4	21.09	30.72	0.788	0.75
Original SD	31.32	31.32	0.813	0

Table 4: CSR unlearning effects across multiple concept combinations, reporting CLIP scores for targeted and non-targeted prompts along with Δ FID and TIFA for non-targeted concepts.

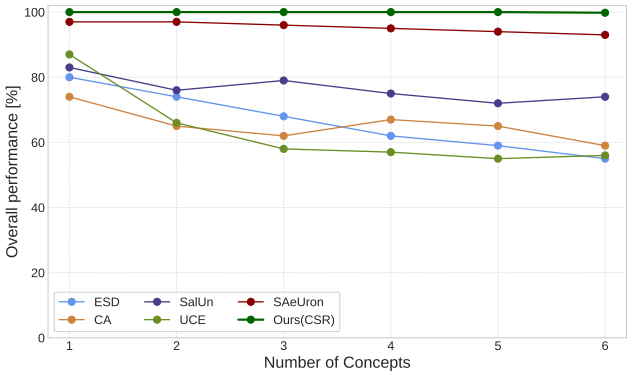


Figure 8: Figure shows the comparison of the overall performance of TRuST across other SOTA baselines.

Table 5: Training efficiency on one A100 GPU (except Retrain)

Method	# Steps ↓	# Images ↓	Time (hrs) ↓
Retrain	80,000	10^7	150,000
ESD	1000	540	2.5
SalUn	1300	800	2
SalUn++	900	800	6
CA	210	–	1
CoGFD	150	–	0.5
TRuST (CSR)	60	350	0.25
TRuST (CIP)	100	350	0.72

scores on non-targeted and targeted concepts. As shown in Table 4, unlearning four concept combinations results in only a 0.02 drop in TIFA score and less than 1-point increase in FID. The four concept concept combinations used were: "Boy holding a Gun", "Drunk person driving a Car", "Child drinking Alcohol", "Pregnant Woman drinking ALcohol"

Runtime Efficiency. In terms of training runtime efficiency, we compare against SOTA methods for total fine-tuning steps, number of training samples, and end-to-end wall-clock time (Table 5). CSR achieves the best overall efficiency, requiring only 60 steps, 350 samples, and 15 minutes of wall-clock time. This is 2–10x faster than prior methods such as CoGFD (0.5h), SalUn (2h), SalUn++ (6h), and ESD (2.5h), while using significantly less training data. CIP also outperforms all baselines except CoGFD in total runtime, but requires fewer steps (100 vs. 150), which improves scalability and reduces overfitting risk. Crucially, unlike CoGFD, which suffers from catastrophic interference, both CSR and CIP maintain high utility across retained concepts and maintain robustness to adversarial prompts.

We present the inference runtime comparison of our method in Table 3. The computations shown in the table were done for performing unlearning on the UnlearnCanvas dataset [57]. In our case we perform unlearning of the "Van Gogh" style and compute the comparison metrics as per their code-base. Notably, TRuST performs very well in unlearning artistic styles with only a few finetuning steps as low as 10 for the CIP based loss function.

Discussion on CIP and CSR. Our results show that both CIP and CSR are more effective in unlearning compared to existing methods. Compared to CSR, CIP is more robust to adversarial prompts and better at erasing target concepts, but slightly reduces output fidelity for benign concepts and takes longer to converge. CSR, on the other hand, is extremely efficient, preserves image realism and semantic coherence better, especially for concept combinations, but is less aggressive in unlearning. Extended discussion and contrastive experiments can be found in the Appendix **??**.

8 Design Choices and Ablation Studies

8.1 Design Choices

CLIP Score We use the CLIP score for computing the saliency map because of its well known ability to effectively capture the relationship between the images and corresponding text by computing the embeddings for both of them in the same latent space. It has been used previously by many [4, 22, 24]. Furthermore, since the text encoder in Stable Diffusion 1.5 is itself a CLIP-based model, employing a CLIP-derived alignment loss is both natural and optimally aligned with the underlying architecture.

Hyperparameter β We choose the values of both β_{CIP} and β_{CSR} as 0.001 to keep the final loss in the optimal range for effective finetuning i.e. (0.1-0.01), and also comparable to the preservation loss. Some examples of the total CIP and CSR losses and preservation loss is as shown:

```
CIP Loss: 0.010000000707805157 * 1e3| preservation loss:
0.15577034652233124
CSR Loss: 0.012630711309611797 * 1e3 | preservation
loss: 0.35134732723236084
```

Threshold ξ The threshold ξ controls the number of concept neurons which are associated to a concept. Higher the value of ξ, lower will be the number of concept neurons identified. To choose the most efficient value of ξ, we conduct an ablation study with different values of this hyperparameter is shown in Table 6. We finally choose the value of ξ as 2.0 to achieve a suitable balance between the number of steps and overall accuracy (computed using $\frac{UA+RA}{2}$).

Preservation loss During finetuning based solely on the CIP or CSR based loss functions, we observed that the nearby concepts

ξ	Concept Neurons Identified	$\frac{UA+RA}{2}\uparrow$	Fine-tuning Steps $\downarrow$
1	34	93.37%	70
2	15	100%	110
3	7	100%	300

Table 6: Ablation on three different values of the threshold for identifying the concept neurons, ξ for CIP.

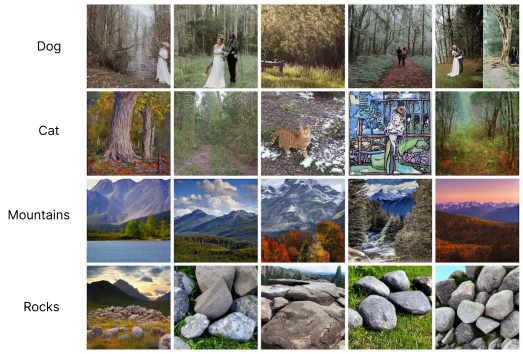


Figure 9: This figure represents four samples from a CIP-unlearned model (no preservation loss): unlearning 'Dog' degrades nearby concept 'Cat', while distant concepts remain unaffected.

were also getting damaged due to the finetuning. Figure 9 shows an example from the ablation experiment conducted without the preservation loss term. In the experiment, we unlearn the concept of "Dog" using the CIP loss only. From the visual samples generated per prompt, we can see that the nearby concept of "Cat" is also impacted by the unlearning, whereas distant concepts such as "Mountains" and "Rocks" remains unaffected. The preservation set consists of 350 prompt-image pairs across 80 diverse concepts from the COCO validation set.

8.2 Deactivating identified concept neurons

As an ablation to establish the need for finetuning, we deactivate (zero the activation) for the identified concept neurons to check if they were exclusively responsible for expressing the concept. We do this for both artistic styles (eg. Van Gogh) and for concrete concepts (eg. Dog). Figure 10 shows the impact of deactivating the 23 concept neurons identified for the concept "Van Gogh". We can observe that merely deactivating the concept neurons pertaining to the concept of "Van Gogh" faithfully removed the style of famous "Van Gogh" paintings. However, this deactivation led to impacting the generation quality for other non-related concepts as shown in Figure 12. This could be because of neuron multiplexing, i.e. the bunch of concept neurons which were deactivated, were holding information pertaining to other non-related concepts as well. And therefore, their deactivation lead to impacting other non-targeted concepts like "Realistic Man", etc.

However, we noticed that this deactivation, though helpful for removing subtle artistic concept like "Van Gogh", is not suitable

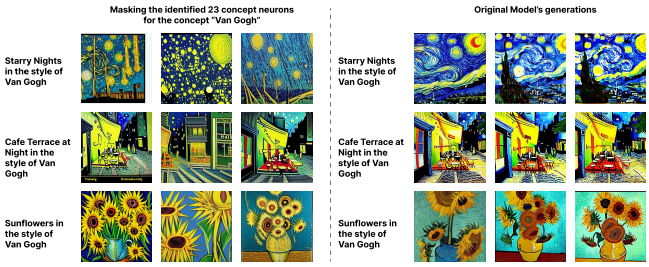


Figure 10: Impact on SD 1.5 output before/after deactivating the identified concept neurons for the "Van Gogh".

Figure 11: Output of SD 1.5 before and after deactivating the identified concept neurons for the concept "Dog".

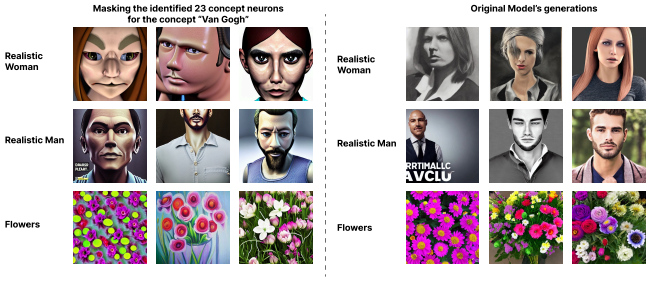


Figure 12: Impact on non-related concepts when deactivating the concept neurons for the artistic style concept "Van Gogh".

for removing concrete concepts like "Dog". Figure 11 shows that after deactivating all 14 concept neurons identified for the concept "Dog", the model could still generate dogs. This is because, though the set of concept neurons were identified to be responsible for holding the information, they were not solely responsible for all the information. The entire information regarding the concept is rather spread across the entire network, which contributes towards the generation of the targeted concept. Therefore, TRUST employs a dynamic finetuning approach where the important set of neurons are iteratively selected throughout the network at each finetuning step.

8.3 SalUn++: SalUn with dynamic saliency maps

To motivate CLIP score–based guidance for saliency map detection, we introduce SalUn++, a variant of SalUn [9] that recomputes the saliency mask dynamically at every step using a noise-based score. Figure 2(c)(a) and (c) show, respectively, the impact on benign-concept generation and erasure efficiency. Panel (a) indicates that dynamic (vs. static) saliency computation reduces collateral damage to nearby concepts, though substantially less than CLIP score–based detection. Panel (c) shows that dynamic saliency maps also

Table 7: Preservation set dependance ablation.

Checkpoint	Visible CLIP	Held-out CLIP	Gap
SD1.5 baseline	0.3101	0.3091	+0.0010
CIP step 50	0.3053	0.3068	−0.0015

Table 8: Over erasure ablation study

Config	Target Score ↓	Other Concept Score ↑
SD1.5 baseline	0.883	0.933
CIP (step 10)	0.183	0.817

accelerate convergence toward erasing the targeted concept, but CLIP score–based construction converges markedly faster than SalUn++. Together, these results establish that both CLIP score–based concept-neuron identification and dynamic saliency reconstruction are necessary for targeted, efficient concept erasure with minimal collateral damage.

8.4 Preservation set ablations

In order to evaluate the impact of the preservation set on the ability of the edited model to generate benign concepts, we conduct an ablation study where we divide the set of 80 benign concepts from the COCO validation set into two equal sets of 40 concepts each. One part of this set is used as the preservation set while performing unlearning via TRuST. Then, this unlearned model's generative ability is measured on the held out set of 40 concepts, via CLIP Score. From Table 7, we can see that there is almost no impact on the generative quality of held out concepts, in comparison to the base model.

Moreover, to study the possibility of over-erasure, we design an experiment where we evaluate the presence of other concepts in the case where the prompt contains the targeted unlearned concept. We experiment with this in the challenging case of artistic styles, and specifically on unlearning the "Van Gogh" style. We generate 300 homogeneous prompts targeted for generating different object in the targeted style. Then, selectively evaluate the presence of the unlearned concept ("Van Gogh" style in this case) as the *Target Score* and the presence of other concepts in the prompt as the *Other Concept Score*, in the generated image from the unlearned model. Table 8 represents the comparison in the original SD 1.5 model vs the unlearned model via TRUST (CIP Loss).

As shown in Table 8, the *Target Score* successfully drops significantly (79.3% decrease) with a much less pronounced decrease on the *Other Concept Score* (12.4% decrease) which remains above 80% guaranteeing strong utility preservation even in this challenging case. Upon further investigation, we were able to correlate this drop to nature specific concepts like rose, flower, trees, etc. We assume this could be due to the strong correlation between these concepts and Van Gogh Style, as a considerable part of Van Gogh's paintings were made with these subjects.

Table 9: Quantization Adversarial Attack Ablation results

Method	Precision	CLIP score ↑	Δ FID ↓	TIFA ↑	I2P ASR ↓	UnlearnDiffAtk ASR ↓
CIP	fp32	0.3037	0.00514	0.8368	0.00827	0.04915
	fp16	0.3081	0.00956	0.8245	0.00842	0.03898
	bf16	0.3029	0.00379	0.7969	0.00870	0.04407
CSR	fp32	0.3054	0.00458	0.8021	0.00978	0.05339
	fp16	0.3047	0.00641	0.8145	0.00972	0.04915
	bf16	0.3079	0.00244	0.7969	0.00955	0.05169

8.5 Post-Release Model-Side Recovery Attack

Our threat model considers static and adaptive adversaries that attempt to recover the erased concept through prompts while leaving the released model parameters unchanged (Section 5). As an additional robustness diagnostic, we evaluate two stronger post-release model-side recovery scenarios in which an adversary obtains the edited checkpoint and either changes its inference precision or further fine-tunes it.

Inference-time quantization. A prominent approach considered in the unlearning literature (primarily in LLM unlearning) studies changing the quantization during inference [59]. We evaluate the robustness of TRuST to quantization in the T2I diffusion model setting, considering an adversary with the capability to change the inference precision of the generative model. We evaluate CLIP-Score, ΔFID, and TIFA to measure the impact on benign generation, as well as robustness against static (I2P ASR) and adaptive (UnlearnDiffAtk ASR) adversaries at three precision levels: fp32, fp16, and bf16. Table 9 shows minimal impact on benign generation as inference precision is reduced. More importantly, we observe minimal changes in attack success rates for both static and adaptive adversaries.

Post-release fine-tuning for relearning. We evaluate TRuST's robustness against recovery under a bounded malicious fine-tuning protocol. Following the attack setting of Pan et al. [35], which studies recovery under bounded fine-tuning budgets, we fine-tune our nudity-unlearned checkpoint on 349 real nudity images with BLIP-generated captions using their plain diffusion-loss objective, for 30 optimization steps.

Table 10 reports the NSFW score (LAION CLIP-based detector [44]) and attack success rate (ASR, fraction of generations scored NSFW) on the I2P nudity prompt set, alongside CLIPScore and FID on benign prompts to track utility. The retraining attack raises the NSFW score minimally from 0.276% to 0.375%, with a similar rise in I2P ASR. Similarly, we observe a minor improvement in generation quality in terms of CLIPScore and ΔFID. Therefore, under such a bounded retraining protocol, recovery of the erased concept is negligible.

9 Conclusion

We introduced TRuST, a selective fine-tuning method for robust concept unlearning in text-to-image diffusion models. TRuST is motivated by a simple but important observation: the parameters most responsible for a target concept are not fixed during unlearning. Rather than relying on a static saliency mask, TRuST dynamically re-identifies concept-relevant cross-attention parameters throughout fine-tuning and updates only that localized subset.

Building on this dynamic localization mechanism, we proposed two complementary objectives: *Concept Influence Penalty* (CIP),

Table 10: Retraining based adversarial attack robustness

Metric	Pre-attack	Post-attack
NSFW score	0.00276	0.00375
I2P ASR	0.00274	0.00379
CLIP score	0.310	0.315
Δ FID	0.095	0.030

which aggressively suppresses high-influence concept parameters, and *Concept Sensitivity Reduction* (CSR), which weakens the model's sensitivity to concept-specific parameters while preserving more shared structure. Together, these objectives provide two practical operating points for model-level safety editing: stronger erasure when robustness is paramount, and softer suppression when preserving compositional utility is critical.

Our evaluation shows that TRuST improves substantially over prior concept-erasure methods across a broad set of criteria: robustness to static and adaptive adversarial prompts, preservation of benign generation quality, fine-tuning efficiency, multi-concept scalability, and selective removal of concept combinations. Beyond standard single-concept unlearning, TRuST also supports more policy-relevant forms of editing, including *concept combination erasure* and *conditional concept erasure*, where the goal is not to remove individual concepts wholesale but to suppress specific harmful relations among otherwise benign concepts.

These results suggest that robust unlearning requires more than removing a word, class, or isolated feature: it requires identifying and editing the model mechanisms through which concepts and relations are expressed. TRuST approach of dynamic and targeted model editing is a promising path toward more robust and interpretable safeguards for text-to-image generation.

Acknowledgments

This research was partially funded by EPSRC (grant UKRI3928, NeSyDebates).

References

[1] Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Vlad I Morariu, Nanxuan Zhao, Ryan A Rossi, Varun Manjunatha, and Soheil Feizi. 2024. On mechanistic knowledge localization in text-to-image generative models. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 129, 42 pages.

[2] Praneeth Bedapudi. 2019. NudeNet: Neural Nets for Nudity Detection and Censoring. https://pypi.org/project/nudenet/. Accessed: 2026-04-26.

[3] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 396–410. doi:10.1145/3600211.3604722

[4] Zikui Cai, Yaoteng Tan, and M. Salman Asif. 2024. Targeted Unlearning with Single Layer Unlearning Gradient. In *Neurips Safe Generative AI Workshop 2024*. https://openreview.net/forum?id=ePKuQQwCGm

[5] Eli Chien, Chao Pan, and Olgica Milenkovic. 2022. Certified Graph Unlearning. arXiv:2206.09140 [cs.LG] https://arxiv.org/abs/2206.09140

[6] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Cheng Chiu. 2024. Prompting4Debugging: red-teaming text-to-image diffusion models by finding problematic prompts. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 336, 19 pages.

[7] Bartosz Cywiński and Kamil Deja. 2025. SAeUron: Interpretable Concept Unlearning in Diffusion Models with Sparse Autoencoders. In *ICLR 2025 Workshop: XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge*. https://openreview.net/forum?id=HFCaWGWEzi

[8] Anudeep Das, Vasisht Duddu, Rui Zhang, and N. Asokan. 2025. Espresso: Robust Concept Filtering in Text-to-Image Models. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy* (Pittsburgh, PA, USA) *(CODASPY '25)*. Association for Computing Machinery, New York, NY, USA, 305–316. doi:10.1145/3714393.3726502

[9] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. 2024. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=gn0mIhQGNM

[10] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. 2024. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*. AAAI Press, Article 1344, 9 pages. doi:10.1609/aaai.v38i11.29092

[11] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing Concepts from Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2426–2436. doi:10.1109/ICCV51070.2023.00230

[12] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified Concept Editing in Diffusion Models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 5099–5108. doi:10.1109/WACV57701.2024.00503

[13] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. 2025. Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 73–88.

[14] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. 2020. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 359, 11 pages.

[15] Alvin Heng and Harold Soh. 2023. Selective Amnesia: A Continual Learning Approach to Forgetting in Deep Generative Models. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=BC1IJdsuYB

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6629–6640.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 574, 12 pages.

[18] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[19] Seunghoo Hong, Juhun Lee, and Simon S. Woo. 2024. All but One: Surgical Concept Erasing with Model Preservation in Text-to-Image Diffusion Models. In *AAAI*. 21143–21151. https://doi.org/10.1609/aaai.v38i19.30107

[20] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 20349–20360. doi:10.1109/ICCV51070.2023.01866

[21] Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. 2025. TraSCE: Trajectory Steering for Concept Erasure. arXiv:2412.07658 [cs.CV] https://arxiv.org/abs/2412.07658

[22] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. 2024. CoMat: Aligning Text-to-Image Diffusion Model with Image-to-Text Concept Matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=OW1ldvMNJ6

[23] Dahye Kim and Deepti Ghadiyaram. 2025. Concept Steerers: Leveraging K-Sparse Autoencoders for Controllable Generations. *CoRR* abs/2501.19066 (January 2025). https://doi.org/10.48550/arXiv.2501.19066

[24] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2416–2425. doi:10.1109/CVPR52688.2022.00246

[25] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating Concepts in Text-to-Image Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 22634–22645. doi:10.1109/ICCV51070.2023.02074

[26] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1931–1941. doi:10.1109/CVPR52729.2023.00192

[27] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. 2024. Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12006–12016. doi:10.1109/CVPR52733.2024.01141

[28] Leyang Li, Shilin Lu, Yan Ren, and Adams Wai-Kin Kong. 2025. Set You Straight: Auto-Steering Denoising Trajectories to Sidestep Unwanted Concepts. arXiv:2504.12782 [cs.CV] https://arxiv.org/abs/2504.12782

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[30] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. 2023. Cones: concept neurons in diffusion models for customized generation. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 890, 19 pages.

[31] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. 2024. MACE: Mass Concept Erasure in Diffusion Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 6430–6440. doi:10.1109/CVPR52733.2024.00615

[32] Zheling Meng, Bo Peng, Xiaochuan Jin, Yueming Lyu, Wei Wang, Jing Dong, and Tieniu Tan. 2025. Concept Corrector: Erase concepts on the fly for text-to-image diffusion models. arXiv:2502.16368 [cs.CV] https://arxiv.org/abs/2502.16368

[33] Hongyi Nie, Quanming Yao, Yang Liu, Zhen Wang, and Yatao Bian. 2025. Erasing Concept Combination from Text-to-Image Diffusion Model. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=OBjF5I4PWg

[34] Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. 2023. Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=frVo9MzRuU

[35] Jiadong Pan, Hongcheng Gao, Zongyu Wu, taihang Hu, Li Su, Qingming Huang, and Liang Li. 2024. Leveraging Catastrophic Forgetting to Develop Safe Diffusion Models against Malicious Finetuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=pR37AmwbOt

[36] Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2025. UnsafeBench: Benchmarking Image Safety Classifiers on Real-World and AI-Generated Images. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security* (Taipei, Taiwan) *(CCS '25)*. Association for Computing Machinery, New York, NY, USA, 3221–3235. doi:10.1145/3719027.3765088

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 1, Article 140 (Jan. 2020), 67 pages.

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 10674–10685. doi:10.1109/CVPR52688.2022.01042

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241.

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2643, 16 pages.

[42] Andrea Schioppa, Emiel Hoogeboom, and Jonathan Heek. 2024. Model integrity when unlearning with t2i diffusion models. *arXiv preprint arXiv:2411.02068* (2024).

[43] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22522–22531. doi:10.1109/CVPR52729.2023.02157

[44] Christoph Schuhmann. 2022. LAOIN CLIP-based-NSFW-Detector. https://github.com/LAION-AI/CLIP-based-NSFW-Detector. LAION-AI.

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=M3Y74vmsMcY

[46] Viacheslav Surkov, Chris Wendler, Mikhail Terekhov, Justin Deschenaux, Robert West, and Caglar Gulcehre. 2024. Unpacking SDXL Turbo: Interpreting Text-to-Image Models with Sparse Autoencoders. https://openreview.net/forum?id=Ch8s4FdUXS

[47] The Alan Turing Institute. 2024. *AI-Enabled Influence Operations: The Threat to the UK General Election*. Technical Report. Centre for Emerging Technology and Security. https://cetas.turing.ac.uk/sites/default/files/2024-05/cetas_briefing_paper_-_ai-enabled_influence_operations_-_the_threat_to_the_uk_general_election.pdf Briefing Paper.

[48] Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. 2022. On the Necessity of Auditable Algorithmic Definitions for Machine Unlearning. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 4007–4022. https://www.usenix.org/conference/usenixsecurity22/presentation/thudi

[49] 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. 2024. Ring-A-Bell! How Reliable are Concept Removal Methods For Diffusion Models?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=lm7MRcsFiS

[50] Peiran Wang, Qiyu Li, Longxuan Yu, Ziyao Wang, Ang Li, and Haojian Jin. 2024. Moderator: Moderating Text-to-Image Diffusion Models through Fine-grained Context-based Policies. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT, USA) *(CCS '24)*. Association for Computing Machinery, New York, NY, USA, 1181–1195. doi:10.1145/3658644.3690327

[51] Yixin Wu, Yun Shen, Michael Backes, and Yang Zhang. 2024. Image-Perfect Imperfections: Safety, Bias, and Authenticity in the Shadow of Text-To-Image Model Evolution. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT, USA) *(CCS '24)*. Association for Computing Machinery, New York, NY, USA, 4837–4851. doi:10.1145/3658644.3690288

[52] Zongyu Wu, Hongcheng Gao, Yueze Wang, Xiang Zhang, and Suhang Wang. 2024. Universal Prompt Optimizer for Safe Text-to-Image Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6340–6354. doi:10.18653/v1/2024.naacl-long.351

[53] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang xu. 2024. MMA-Diffusion: MultiModal Attack on Diffusion Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 7737–7746. doi:10.1109/CVPR52733.2024.00739

[54] Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. 2024. GuardT2I: Defending Text-to-Image Models from Adversarial Prompts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=FMrNus3d0n

[55] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. 2025. SAFREE: Training-Free and Adaptive Guard for Safe Text-to-Image And Video Generation. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=hgTFotBRKl

[56] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. 2024. Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=dkpmfIydrF

[57] Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Kompella, Xiaoming Liu, and Sijia Liu. 2024. UNLEARNCANVAS: a stylized image dataset for enhanced machine unlearning evaluation in diffusion models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 3055, 37 pages.

[58] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. 2024. To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy to Generate Unsafe Images ... For Now. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LVII* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 385–403. doi:10.1007/978-3-031-72998-0_22

[59] Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. 2025. Catastrophic Failure of LLM Unlearning via Quantization. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=lHSeDYamnz