MULTIPLE CLASSES ERASURE USING SUPERCLASS IN TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in text-to-image diffusion models have enabled highly realistic image synthesis, but they also raise concerns regarding the generation of unwanted or unsafe content. Existing erasure methods often struggle to remove target classes reliably when prompts are detailed, paraphrased, or adversarial, leading to incomplete forgetting and compromised fidelity of non-target content. To address these limitations, we propose **Context Graph Erasure (CGE)**, a framework that leverages structured scene knowledge through context graphs to guide precise and controllable concept erasure. CGE constructs enriched representations of the visual scene by encoding objects, attributes, and relations into a learnable graph-based embedding, which is integrated with the text conditioning. A dedicated erasure module utilizes this enriched representation to suppress target superclass, while a cross-attention mechanism preserves the integrity of unrelated regions. Furthermore, an adversarial concept graph strategy allows the system to manage prompts that are phrased differently, maintaining consistent results. Extensive experiments demonstrate that CGE achieves superior erasure accuracy and preserves unrelated content with high fidelity, outperforming prior methods and providing a reliable, generalizable solution for multiple classes erasure in text-to-image models.

1 INTRODUCTION

Text-to-image (T2I) generation models, such as DALL·E 2 Ramesh et al. (2022) and Stable Diffusion Rombach et al. (2022), have achieved remarkable progress in translating textual prompts into high-quality, photorealistic images Wu et al. (2023). Their versatility has driven widespread adoption in creative domains, including content generation, design, and digital media, made possible by large-scale web-scraped datasets Schramowski et al. (2023a) and advanced neural architectures Nichol et al. (2022). However, this dependence on uncurated data raises significant ethical, legal, and safety concerns Jiang et al. (2023); Somepalli et al. (2023); Schramowski et al. (2023b). These models frequently memorize and reproduce inappropriate material, such as NSFW imagery, copyrighted works, or even personal photos Carlini et al. (2023); Lee et al. (2025). Existing countermeasures, such as filtering training datasets Rando et al. (2022), retraining models Gandikota et al. (2023), or applying post-generation safety filters Ganguli et al. (2022), remain limited, as they either impose high computational costs, compromise image quality, or can be circumvented through simple prompt paraphrasing Schramowski et al. (2023b).

Concept erasure has emerged as a promising approach for preventing text-to-image (T2I) models from generating undesired content Kumari et al. (2023). The objective is to suppress specific concepts while retaining the model's ability to produce unrelated content. However, existing methods often struggle to achieve this consistently. They perform adequately when tested with simple prompts but fail when the prompts are paraphrased, detailed, or adversarial Gandikota et al. (2023). In such cases, the supposedly erased concepts reappear in the generated outputs, indicating that the erasure is incomplete and fragile. This limitation highlights the central challenge: concept erasure methods must remain effective across diverse natural prompt formulations to ensure reliable and trustworthy deployment of T2I models.

Current generative unlearning methods primarily focus on removing individual classes or small sets of classes while preserving the model's ability to generate unrelated content Zhang et al. (2023); Yu et al. (2025); Gandikota et al. (2023; 2024); Bui et al. (2024; 2025); Wang et al. (2024; 2025); Wu

et al. (2024). However, these approaches face a critical limitation: they are unable to erase entire superclasses that encompass multiple classes. For instance, the superclass "Guns" includes classes such as "pistol," "revolver," and "SMG." Manually enumerating all classes within a superclass is not only cumbersome but also nearly impossible in practice, as the full set of classes is rarely known in advance. Some methods Gandikota et al. (2023) remove only a few classes within a superclass, leaving others unaffected and thus failing to erase the superclass entirely. Others, such as MACE Lu et al. (2024), attempt to use multiple prompts to cover several classes, but this substantially increases computational cost and still risks incomplete erasure if any class is missed. Methods have also tried to remove the entire superclass at once, but often suffer from suboptimal erasure or unintended preservation of certain classes. These limitations underscore the need for scalable, structured methods that can reliably erase whole superclasses by specifying only a single representative class.

To address these challenges, we draw inspiration from human visual reasoning, where abstract superclass are first structured as a basic scene layout and then developed into detailed imagery. We use context graphs (CGs) to achieve a similar approach in models. Unlike unstructured text, context graphs explicitly describe objects, their properties, and the relationships between them, providing richer semantic information. For example, instead of merely recognizing the word "pistol," the model interprets it as an object with attributes like "metallic" and relationships such as "held by a person." This structured information allows the model to better locate and understand the target superclass, enabling more precise suppression when needed. Even with context graphs, objects, attributes, and relationships are encoded in a shared space, which can make it challenging to isolate and manipulate specific class precisely.

To address this, we introduce multiple enriched representation, each capturing different aspects of the context graph structure. These are first combined into a single enriched representation through a learnable gating mechanism, and this aggregated representation is then integrated with adapter-based textual representations using a second gating mechanism. This dual-gating design enlarges the conditioning space, allowing the optimizer to follow multiple paths instead of relying on overlapping directions in the original representation. Through adversarial concept graph and class-focused regularization within the dual-gating framework, our method reliably erases target classes while preserving non-target content, effectively resolving the trade-off between erasure accuracy and content fidelity.

We evaluate the proposed methods across three superclass, Guns, Bladed Weapons, and Musical Instruments, as well as on tasks such as explicit content suppression and artistic style removal. Experiments on physical superclass, I2P, and five distinct artistic styles demonstrate the effectiveness of our approach in removing entire superclass while preserving unrelated classes. Our methods consistently outperform all state-of-the-art baselines, including evaluations on CIFAR-10.

2 Related Works

Fine-tuning-based Concept Erasure: Fine-tuning approaches adapt pre-trained text-to-image models to remove specific target concepts while preserving unrelated content, offering an efficient alternative to retraining from scratch. A common strategy modifies attention mechanisms to suppress concept-specific information. For instance, Forget-Me-Not Zhang et al. (2023) fine-tunes the U-Net to re-steer attention maps away from target concepts, while TIME Orgad et al. (2023) adjusts key and value projections in cross-attention layers to redirect undesired influences. Receler Huang et al. (2024) introduces concept-localized regularization applied to cross-attention outputs, and MACE Lu et al. (2024) leverages multiple LoRA modules for closed-form cross-attention refinement. UCE Gandikota et al. (2024) and AdaVD Wang et al. (2024) further enhance selective erasure by preserving non-target content through explicit objectives or orthogonal complement strategies. Other methods directly optimize model outputs to disassociate the target concept from generation, as seen in ESD Gandikota et al. (2023), AGE Bui et al. (2025), and AP Bui et al. (2024), with ACE Wang et al. (2025) extending this to both generation and editing scenarios.

Adversarially Aware Concept Erasure: Several methods address the challenge of maintaining concept erasure under adversarial or paraphrased prompts. Adversarial erasure schemes, such as those in Receler Huang et al. (2024) and AdvUnlearn Zhang et al. (2024a), alternate training between concept removal and adversarial resistance, often incorporating regularization to balance erasure and content preservation. UnlearnDiff Zhang et al. (2024b) employs projected gradient descent to

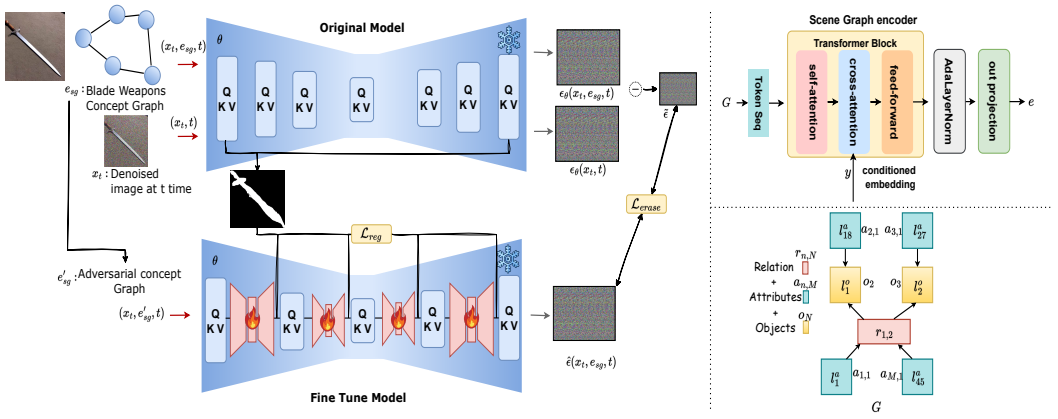

Figure 1: Overview of the proposed Context Graph Erasure (CGE) framework. Structured context graphs are encoded and fused with textual prompts to form enriched conditioning, which guides the eraser module for targeted concept suppression. Cross-attention regularization ensures localized modifications, while adversarial concept graph enhances resilience to paraphrased or varied prompts.

optimize erasure under adversarial settings, while RACE Kim et al. (2024) identifies vulnerable embeddings to target for removal. Similarly, AP Bui et al. (2024) and AGE Bui et al. (2025) focus on adversarially sensitive concepts to guide stable erasure. STEREO Srivatsan et al. (2024) adopts a two-stage min-max optimization strategy combining adversarial training with anchor-concept-based compositional objectives. Despite these efforts, existing methods often struggle to erase entire superclass reliably, leaving room for approaches that integrate structured scene knowledge to achieve precise and generalizable concept removal.

Addressing these limitations, our work introduces a context graph–based approach that leverages structured representations of objects, attributes, and relations to guide output-level fine-tuning, enabling superclass-level erasure while preserving non-target content and maintaining performance under adversarial or paraphrased prompts.

3 Methodology

3.1 Content Graph

The context graph G provides a structured representation of a visual scene by decomposing it into three complementary components: objects, attributes, and relations. The object set $O = \{o_1, \ldots, o_N\}$ encodes the entities present in the scene, with each object o_n associated with a superclass label ℓ_n^o. Attributes describing object properties such as color, size, or material are represented as $A = \{a_{n,1}, \ldots, a_{n,M_n}\}$, where each attribute $a_{n,m}$ is linked to its corresponding object o_n. Relations between object pairs are captured by $R = \{r_{i,j}\}$, where each $r_{i,j}$ denotes an interaction between o_i and o_j, labeled with $\ell_{i,j}^r$ from a relation vocabulary. Together, these components define the scene graph $G = (O, A, R)$, which provides a symbolic yet structured description of the scene. To map symbolic elements into the model space, each label ℓ^o, ℓ^a, ℓ^r is associated with a learnable embedding. The representation of an object o_n is then obtained by combining its object embedding with the embeddings of its attributes, relations, and positional encoding:

$$h_{o_n} = E_o(\ell_n^o) + \sum_m E_a(\ell_{n,m}^a) + \sum_j E_r(\ell_{n,j}^r) + E_{\text{pos}}(n). \tag{1}$$

These object-centric vectors are arranged into a token sequence that preserves both local properties and contextual dependencies. A lightweight Transformer processes this sequence, augmented with cross-attention to the textual prompt. Temporal consistency with the diffusion process is ensured through AdaLayerNorm, after which outputs are projected to match the backbone dimensionality. Since multiple graphs $\{G_k\}_{k=1}^K$ can be derived from diverse object–attribute–relation triplets, the model encodes each into a parallel sequence $\{S_k\}$. These are then fused through a learnable gating

mechanism that adaptively weights their contributions:

$$S_{\text{enriched}} = \sum_{k=1}^{K} \alpha_k S_k, \qquad \alpha_k = \frac{\exp(w_k)}{\sum_j \exp(w_j)}, \tag{2}$$

where the gates w_k are optimized during training. This formulation ensures that relevant graphs contribute proportionally, while avoiding instability as the number of graphs varies. To balance structured graph signals with the original text conditioning, a second gating mechanism interpolates between the text embedding sequence e_n and the enriched graph sequence S_{enriched}:

$$e_{sg} = g \cdot e_n + (1 - g) \cdot S_{\text{enriched}}, \qquad g \in [0, 1], \tag{3}$$

where g is a learnable parameter. This layered design enables selective integration of structured knowledge into the generative process while preserving semantic fidelity of the text. Training further reinforces this balance through complementary objectives: the erasure loss enforces removal of target classes, the preservation loss safeguards unrelated content, and cross-attention regularization localizes modifications. Together, these components yield a coherent pipeline where context graphs guide superclass erasure in a structured, controllable, and semantically consistent manner.

3.2 Targeted Superclass Suppression

To selectively suppress specific visual classes in a pre-trained text-to-image diffusion model, we introduce an adapter module called the *eraser*, parameterized by θ_E, which is fine-tuned independently while keeping the main diffusion model parameters θ frozen. This design enables efficient training without destabilizing the base model. The eraser is applied after each cross-attention layer in the U-Net, which injects class-conditioned embeddings into image features during generation. Using the enriched context graph representation e_{sg} of the target superclass, the eraser constructs a negatively guided target noise $\tilde{\epsilon}$ defined as

$$\tilde{\epsilon} = \epsilon_\theta(x_t, t) - \lambda \cdot \big[\epsilon_\theta(x_t, e_{sg}, t) - \epsilon_\theta(x_t, t)\big], \tag{4}$$

where x_t denotes the latent image at diffusion timestep t, $\epsilon_\theta(x_t, t)$ is the unconditional noise prediction, $\epsilon_\theta(x_t, e_{sg}, t)$ is the context graph-conditioned prediction, and λ controls the strength of negative guidance. Intuitively, this formulation pushes the model away from generating features corresponding to the target concept encoded in e_{sg} by subtracting a fraction of the superclass-conditioned deviation from the unconditional prediction. The eraser is then optimized using an L2 objective:

$$\mathcal{L}_{\text{erase}} = \big\|\hat{\epsilon}_{\theta_E}(x_t, e_{sg}, t) - \tilde{\epsilon}\big\|^2, \tag{5}$$

where $\hat{\epsilon}_{\theta_E}$ denotes the combined prediction from the base model and the eraser module. This setup allows the model to suppress the target superclass while preserving unrelated content. By leveraging structured context from e_{sg}, concept removal becomes precise and controllable, and the eraser integrates seamlessly with the generative process to maintain high-quality, diverse image outputs without degrading semantic fidelity.

3.3 Erasure Precision

To ensure precise suppression of the target superclass, the eraser is constrained to modify only regions highlighted by the diffusion model's cross-attention maps. Let $\hat{\epsilon}_{\theta'}(x_t, e_{sg}, t)$ denote the predicted noise with the eraser applied, conditioned on the enriched scene graph sequence e_{sg}. Cross-attention maps from the selected layers are used to compute a binary mask $\mathcal{M} \in \mathbb{R}^{(W/4)\times(H/4)}$:

$$\mathcal{M}_{i,j} = \begin{cases} 1, & \text{if} \quad \frac{1}{|P|}\sum_{p\in P} \text{CA}^p_{i,j} \geq \tau \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where $\text{CA}^p_{i,j}$ denotes the cross-attention at spatial location (i, j) in layer p, P is the set of layers considered, and τ is the threshold. The mask $\mathcal{M}$ is then upsampled to the eraser output resolution, producing $\tilde{\mathcal{M}}$. The spatially localized regularization loss is defined as:

$$\mathcal{L}_{\text{reg}} = \frac{1}{P}\sum_{p=1}^{P} \Big\|\mathbf{u}_p \odot (1 - \tilde{\mathcal{M}})\Big\|^2, \tag{7}$$

where $\mathbf{u}_p$ is the eraser output at layer p, P is the number of layers, and $\odot$ denotes element-wise multiplication. This objective enforces that the eraser primarily affects regions associated with the target superclass, minimizing unintended modifications to unrelated areas. By leveraging cross-attention from layers guided by e_{sg}, the eraser achieves precise superclass suppression, maintaining the fidelity of non-target content while preserving overall generative quality.

3.4 Adversarial Concept Graph

To improve effectiveness against adversarial or paraphrased prompts, we introduce a learnable adversarial embedding $\mathbf{e}_{adv}$ that simulates challenging modifications of the original conditioning. The adversarial loss is formulated as:

$$\mathcal{L}_{adv} = \left\| \hat{\boldsymbol{\epsilon}}_{\theta'}\big(\mathbf{x}_t, [\mathbf{e}_{\mathbf{sg}}; \mathbf{e}_{adv}], t\big) - \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, \mathbf{e}'_{\mathbf{sg}}, t) \right\|_2^2, \tag{8}$$

where $[\mathbf{e}_{\mathbf{sg}}; \mathbf{e}_{adv}]$ denotes the concatenation of the enriched scene graph representation $\mathbf{e}_{\mathbf{sg}'}$ and the adversarial embedding $\mathbf{e}_{adv}$. Here, $\hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, \mathbf{e}'_{\mathbf{sg}}, t)$ represents the noise prediction conditioned on the target concept, while $\hat{\boldsymbol{\epsilon}}_{\theta'}$ includes the eraser parameters. By optimizing $\mathbf{e}_{adv}$ to generate difficult perturbations, the eraser is trained to remove the target concept reliably even under adversarial input conditions, enhancing generalization to unseen or obfuscated prompts.

4 Experiments

In this section, we evaluate the proposed method across multiple unlearning scenarios. Section 4.1 describes the experimental setup, followed by superclass level erasure in Section 4.2, where we compare against state-of-the-art methods across three superclasses. Section 4.3 and Section 4.4 present evaluations on NSFW concept removal and artist-style forgetting, respectively. An ablation study analyzing the contributions of the frequency filter and loss design is provided in the supplementary material.

4.1 Implementation Details

Experiments are conducted on Stable Diffusion v1.4, following standard protocols in prior work on concept unlearning. Images are generated using the DDIM sampler with 50 steps, guidance scale 7.5, and resolution 512×512. All experiments are performed on a single NVIDIA A6000 GPU (48 GB VRAM). We compare the proposed method against six state-of-the-art concept-erasure approaches: Erased Stable Diffusion (ESD) Gandikota et al. (2023), Unified Concept Erasure (UCE) Gandikota et al. (2024), Adversarial Preservation (AP) Bui et al. (2024), Adaptive Guided Erasure (AGE) Bui et al. (2025), Adaptive Value Decomposer (AdaVD) Wang et al. (2024), and Anti-Editing Concept Erasure (ACE) Wang et al. (2025). For fairness, we adopt experimental configurations consistent with recent studies Wu et al. (2024); Bui et al. (2024; 2025). Specifically, the model is fine-tuned for 1,000 steps using Adam optimizer with batch size 1 and learning rate 1×10^{-5}.

4.2 Superclass Level Erase

4.2.1 Datasets

To evaluate superclass Level erasure, we construct a benchmark covering three superclass, each containing five target classes. Dataset design follows three principles: (a) Distinctiveness, (b) Effectiveness, and (c) Preservation.

(a) Distinctiveness: Superclass are chosen to be clearly defined and visually recognizable, ensuring precise and controlled evaluation. The selected superclass include Guns, Bladed Weapons, Musical Instruments, and Toys, each serving as a canonical example of object-type concepts.

(b) Effectiveness: To assess whether target classes can be reliably erased using straightforward language cues, we adopt simple prompt templates of the form "A photo of class." Each superclass contains five classes: Guns (SMG, Rifle, Pistol, Shotgun, Revolver), Bladed Weapons (Knife, Sword, Spear, Shotel, Dagger), and Musical Instruments (Trumpet, Guitar, Drums, Saxophone, Piano). For each class, 50 prompts are generated, leading to 250 prompts per superclass. Importantly, erasure

Table 1: Comparison of erasure methods across three Superclass: Guns, Bladed Weapons, and Musical Instruments. We report the following metrics: Acc_E — accuracy on erased concepts (↓), Acc_L — locality accuracy on non-erased (remaining) concepts (↑), and $\mathcal{H}$ — harmonic mean (↑). Notably, CGE achieves the highest harmonic mean across all three superclass, demonstrating its effectiveness in erasing multiple classes using only the superclass.

Method	**Venue**	**Guns**			**Blade Weapons**			**Musical Instruments**		
		$Acc_E \downarrow$	$Acc_L \uparrow$	$\mathcal{H} \uparrow$	$Acc_E \downarrow$	$Acc_L \uparrow$	$\mathcal{H} \uparrow$	$Acc_E \downarrow$	$Acc_L \uparrow$	$\mathcal{H} \uparrow$
ESD Gandikota et al. (2023)	ICCV23	52.80	66.20	55.11	18.80	68.60	74.37	51.60	68.80	56.82
UCE Gandikota et al. (2024)	WACV24	55.70	67.40	53.46	33.60	67.80	67.09	68.00	74.69	44.80
AP Bui et al. (2024)	NeurIPS24	55.60	67.00	53.40	32.80	66.40	66.79	62.80	65.40	47.42
AGE Bui et al. (2025)	ICLR25	49.60	68.60	55.92	26.00	68.60	71.19	65.20	66.20	45.61
AdaVD Wang et al. (2024)	CVPR25	78.40	44.65	29.11	57.20	45.65	44.17	65.60	44.30	38.72
ACE Wang et al. (2025)	CVPR25	76.00	70.22	35.77	52.80	64.60	54.54	73.20	67.00	38.28
CGE	–	9.60	67.00	**77.6**	0.0	69.60	**82.1**	12.00	70.20	**78.1**

is applied at the superclass level rather than only on individual classes, ensuring comprehensive forgetting across all related class while preserving unrelated ones.

(c) Preservation: To verify that erasing target concepts does not negatively impact unrelated ones, we include ten non-target classes from CIFAR-10 Krizhevsky et al. (2009). For each, 50 paraphrased prompts are generated to reflect realistic usage where descriptions may be indirect (e.g., "a Doberman Pinscher in a police vest" instead of "a photo of a dog"). This ensures our evaluation captures robustness under natural variations in prompt phrasing.

4.2.2 Evaluation Metrics

Following prior work Huang et al. (2024); Lu et al. (2024), we use `GroundingDINO` Liu et al. (2024) to detect erased concepts in generated images. Two metrics are defined: 1) Accuracy on erased concepts (Acc_E): frequency with which erased concepts appear. 2) Accuracy on preserved concepts (Acc_L): frequency with which unrelated content is retained.

To compute Acc_E, classes within each superclass are considered. For instance, in `Guns`, prompts include *SMG*, *Pistol*, *Rifle*, *Shotgun*, and *Revolver*. Generated samples are analyzed using `GroundingDINO`, and Acc_E is reported as the average percentage of images containing these classes.

Lower Acc_E indicates stronger erasure, while higher Acc_L reflects better preservation. To provide a unified measure, we report the harmonic mean $\mathcal{H}$ between $(100-\text{Acc}_E)$ and Acc_L, following Huang et al. (2024); Lu et al. (2024); Lee et al. (2025):

$$\mathcal{H} = \frac{2}{(100 - \text{Acc}_E)^{-1} + (\text{Acc}_L)^{-1}}. \tag{9}$$

4.2.3 Results

Table 1 compares several superclass erasure methods across three superclasses: Guns, Bladed Weapons, and Musical Instruments. Prior approaches achieve varying degrees of suppression but often face a trade-off: some leave residual traces of the erased concepts, while others achieve stronger erasure at the cost of altering unrelated content. In contrast, CGE consistently achieves the highest harmonic mean across all three superclass, showing its ability to remove entire superclass while preserving non-target content. Notably, during training we assume that the specific classes within each superclass are unknown. For evaluation, we identify sets of 5 classes within each superclass using Wordnet 3.1, which confirms that CGE generalizes beyond individual classes to reliably erase entire superclass. These results highlight that structured context and enriched prompt representations enable precise, scalable, and controlled concept erasure, offering a stronger balance between suppressing target superclass and retaining unrelated information compared to prior methods.

Table 2: NER comparison of text-to-image erasure methods on the I2P dataset for nudity removal (lower is better).

Model	Venue	NER $0.3\downarrow$
CA Kumari et al. (2023)	CVPR 2023	13.84
ESD Gandikota et al. (2023)	ICCV 2023	5.32
UCE Gandikota et al. (2024)	WACV 2024	6.87
AP Bui et al. (2024)	NeurIPS 2024	3.64
AGE Bui et al. (2025)	ICLR 2025	5.06
CGE	–	**4.95**

Table 3: Quantitative evaluation of artistic style erasure across five artists using CLIP and LPIPS scores. Lower CLIP and higher LPIPS indicate better performance.

Model	Venue	CLIP Score ↓	LPIPS Score ↑
ESD Gandikota et al. (2023)	ICCV23	23.56 ± 4.73	0.72 ± 0.11
CA Kumari et al. (2023)	CVPR23	27.79 ± 4.67	0.82 ± 0.07
UCE Gandikota et al. (2024)	WACV24	24.47 ± 4.73	0.74 ± 0.10
MACE Lu et al. (2024)	CVPR24	27.96 ± 4.22	0.60 ± 0.10
AP Bui et al. (2024)	NeurIPS24	21.57 ± 5.46	0.78 ± 0.10
AGE Bui et al. (2025)	ICLR25	22.44 ± 5.03	0.80 ± 0.12
CGE	–	**18.72 ± 5.87**	**0.83 ± 0.10**

4.3 Explicit Content Erasure

4.3.1 Datasets

To evaluate explicit concept removal, we focus on Not-Safe-For-Work (NSFW) content using the Image-to-Prompt (I2P) dataset Schramowski et al. (2023b), which contains 4,703 prompts spanning sexual, violent, and racist content. Unlike suerpclass-level erasure, explicit content removal requires handling abstract and context-dependent semantics. Following prior work, we target the "Nudity" concept for removal. Each prompt is evaluated to verify effective erasure while ensuring unrelated content is preserved, using the COCO-30K validation set Lin et al. (2014).

4.3.2 Evaluation Metrics

For quantitative evaluation, we adopt the Nudity Exposure Rate (NER) metric Bui et al. (2024), computed using the NudeNet detector Bedapudi (2019). NER measures the proportion of generated images containing exposed body regions. Lower NER values indicate stronger erasure performance, reflecting the model's capability to remove the target concept effectively.

4.3.3 Results and Analysis

Table 2 shows that CA exhibits the highest NER (13.84), indicating limited effectiveness in removing nudity from generated images. ESD and UCE substantially reduce residual explicit content, achieving NER values of 5.32 and 6.87, respectively. AP and AGE further improve erasure performance, with AP attaining the lowest NER of 3.64. The proposed CGE method achieves a NER of 4.95, outperforming most prior approaches while maintaining strong preservation of unrelated content. These results demonstrate that CGE effectively suppresses the target nudity superclass across diverse prompts, achieving a reliable balance between erasure strength and content fidelity in text-to-image generation.

4.4 Experiments on Artistic Style Erasure

4.4.1 Datasets

To evaluate artistic style erasure, we focus on five prominent artists: *Kelly McKernan*, *Thomas Kinkade*, *Tyler Edlin*, *Kilian Eng*, and *Ajin Demi Human*. The artist names themselves are treated as the target superclass for removal during fine-tuning. For evaluation, we utilize a curated set of

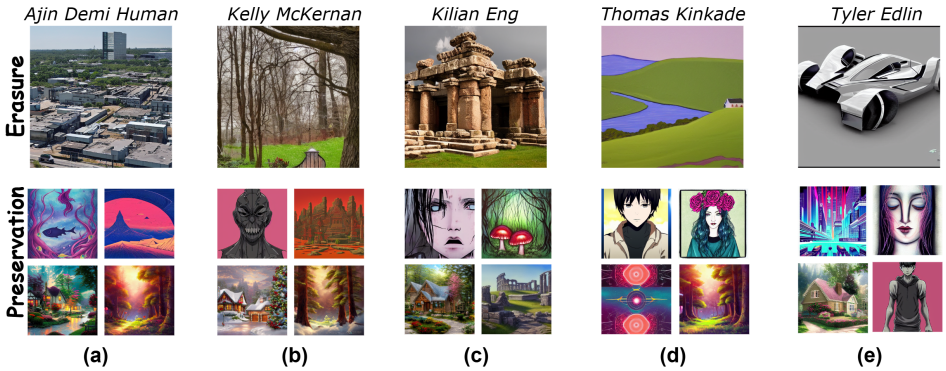

Figure 2: Qualitative results of concept erasure across five artists, presented as: (a) Ajin Demi Human, (b) Kelly McKernan, (c) Kilian Eng, (d) Thomas Kinkade, and (e) Tyler Edlin.

Table 4: Ablation study on the effect of the number of context graphs (N) used for enriched prompt generation. Erasure (%) represents accuracy on erased concepts (Acc_E; lower is better), while Preservation (%) represents accuracy on non-erased concepts (Acc_L; higher is better).

Number of Context Graphs (N)	Erasure (%)	Preservation (%)
5	16.00	65.00
10	12.00	66.40
15	9.60	67.00
20	9.60	65.80

detailed prompts Bui et al. (2024; 2025), with each prompt paired with five random seeds, resulting in 200 images per artist per method. This setup enables a comprehensive assessment of erasure effectiveness across multiple samples.

4.4.2 Evaluation Metrics

We employ two complementary metrics to quantify artistic style removal: 1) **CLIP similarity** Radford et al. (2021) between generated images and their corresponding text prompts. Lower CLIP scores indicate more effective erasure of the targeted style. 2) **Learned Perceptual Image Patch Similarity (LPIPS)** Zhang et al. (2018) between images generated by the original Stable Diffusion model and those from the modified models. Higher LPIPS values reflect stronger removal of artistic styles, while lower LPIPS indicates preservation of visual content and minimal distortion.

4.4.3 Results

Table 3 demonstrates that the proposed CGE method outperforms all baseline approaches across both CLIP and LPIPS metrics. In terms of semantic alignment, CGE achieves a CLIP score of 18.72, representing a 16.6% improvement over the second-best baseline, AGE. Regarding perceptual quality, CGE attains the highest LPIPS score of 0.83, corresponding to a 3.75% improvement over the next-best model. These results indicate that CGE effectively removes artistic styles while preserving fine-grained visual characteristics. Figure 2 provides qualitative comparisons across all five artists. CGE successfully eliminates the targeted artistic traits while maintaining coherent and visually appealing images, confirming its ability to achieve precise concept erasure without compromising image quality.

4.5 Abalation

4.5.1 Number of Context Graphs

Table 4 examines the effect of varying the number of context graphs on enriched prompt generation, highlighting the trade-off between erasure and preservation of unrelated content. A small number of context graphs enhances erasure but slightly compromises preservation, indicating that additional

Table 5: Ablation study of CGE components. Tick (✓) indicates the component is present, and cross (✗) indicates it is removed. Erasure (%) and Preservation (%) reflect Acc_E and Acc_L, respectively.

L_{Erase}	L_{Reg}	Context Graph	L_{Adv}	Erasure (%) ↓	Preservation (%) ↑
✓	✓	✗	✓	12.60	66.00
✓	✓	✓	✗	6.80	65.80
✓	✗	✓	✓	10.40	64.60
✓	✓	✓	✓	9.60	67.00

Table 6: Evaluation of CGE on multiple subclasses of Bladed Weapons, reporting erasure rate (%) and preservation (%).

Class Set	Erasure (%)	Preservation (%)
20 Classes	0.17	69.6
5 Classes	0.0	69.6

context initially strengthens concept removal at the expense of non-target content. Increasing the number of context graphs yields a more balanced outcome, effectively improving erasure while maintaining preservation. However, when too many context graphs are used, redundancy is introduced, limiting further gains in erasure and slightly reducing preservation. These findings suggest that an intermediate range of context graphs provides the most effective balance between removing target concepts and retaining unrelated information.

4.5.2 Component-wise

Table 5 presents a component-wise ablation of CGE, highlighting the contributions of L_{Erase}, L_{Reg}, context graphs, and L_{Adv} to concept erasure and preservation. Without context graphs, the model struggles to effectively remove target superclasses, indicating the necessity of contextual information. Incorporating context graphs substantially enhances erasure while also improving preservation, demonstrating the value of enriched prompts for guiding concept erasure. Excluding L_{Adv} further sharpens erasure but reduces balance, suggesting its role in maintaining fidelity. In contrast, removing L_{Reg} destabilizes training and weakens preservation, underscoring its importance in retaining non-target content. Together, these results show that each component contributes uniquely, and their integration yields the most balanced trade-off between erasure strength and content retention.

4.5.3 Multiple Classes Erasure

To further demonstrate scalability, we evaluate CGE on the entire Bladed Weapons superclass. We extended the superclass to encompass 20 constituent classes, including Gladius, Shuriken, Scythe, Dagger, Axe, Spear, Knife, Machete, Sword, Kukri, Kris, Katana, Shotel, Naginata, Guandao, Estoc, Falchion, Tanto, Ulu, and Cutlass. Earlier experiments focused on a subset of five classes to establish baseline effectiveness. This extended evaluation demonstrates that CGE generalizes beyond individual classes, successfully erasing the entire superclass once all of its classes are identified, while continuing to preserve unrelated concepts.

5 Conclusion

In this work, we introduced Context Graph Erasure (CGE), a structured framework for targeted concept suppression in text-to-image diffusion models. Prior concept erasure methods often perform adequately on simple prompts but struggle when faced with detailed, paraphrased, or adversarial inputs, causing erased concepts to reappear and highlighting the fragility of unstructured forgetting. CGE overcomes this limitation by leveraging context graphs, which encode objects, attributes, and relations within a scene, providing rich contextual information that guides precise and semantically consistent removal of target concepts while preserving unrelated content. An erasure module suppresses target concepts, while adversarial concept graph maintains consistent performance even under paraphrased or challenging prompts. Extensive experiments demonstrate that the integration of structured scene knowledge via context graphs is critical for achieving accurate, stable, and generalizable concept erasure, establishing CGE as a framework that reliably suppresses target concepts even under challenging or paraphrased prompts.

REFERENCES

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

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

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

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

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

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

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pp. 360–376. Springer, 2024.

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

Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, pp. 461–478. Springer, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. *arXiv preprint arXiv:2503.12356*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, 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.

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

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. URL https://arxiv.org/abs/2112.10741.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. 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. doi: 10.1109/ICCV51070.2023.00649.

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. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

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

Javier Rando et al. Red-teaming the stable diffusion safety filter. *NeurIPS Workshop MLSW*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. pp. 22522–22531, April 2023a.

Patrick Schramowski et al. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, 2023b.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 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.

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. *arXiv preprint arXiv:2408.16807*, 2024.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. *arXiv preprint arXiv:2412.06143*, 2024.

Zihao Wang, Yuxiang Wei, Fan Li, Renjing Pei, Hang Xu, and Wangmeng Zuo. Ace: Anti-editing concept erasure in text-to-image models. *arXiv preprint arXiv:2501.01633*, 2025.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*, 2024.

Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceeding of NeurIPS*, 2023.

Zhenyu Yu, Mohd Yamani Inda Idris, and Pei Wang. Forgetme: Evaluating selective forgetting in generative models. *arXiv preprint arXiv:2504.12574*, 2025.

Eric Zhang et al. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in Neural Information Processing Systems*, 37:36748–36776, 2024a.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024b.

A Appendix

A.1 Additional Quantitative Experiments

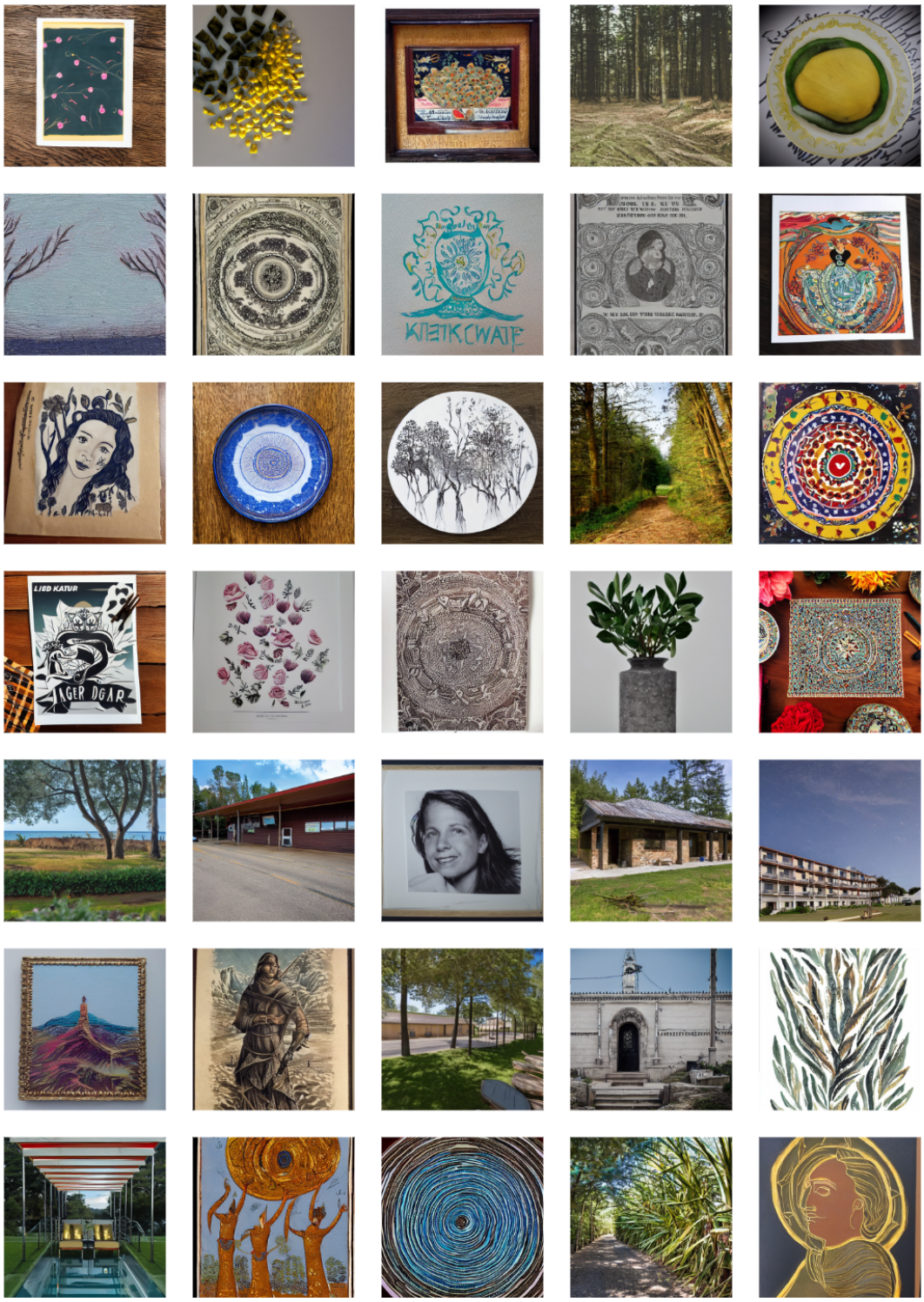

Figure 3: Qualitative results of concept erasure applied to the superclass Bladed Weapons.

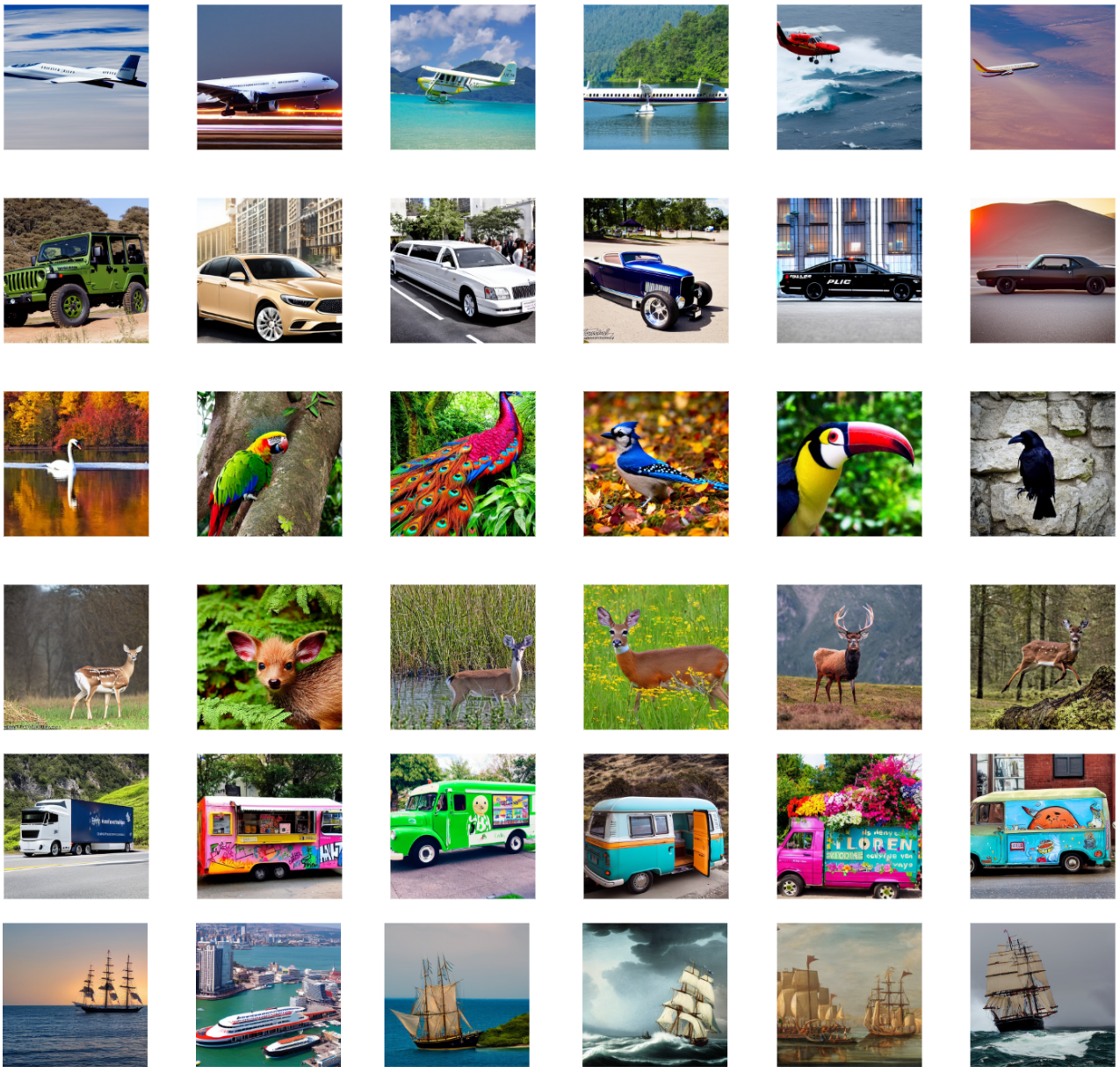

Figure 4: Qualitative results of preservation on the non-target Cifar-10 for target superclass bladed weapons.

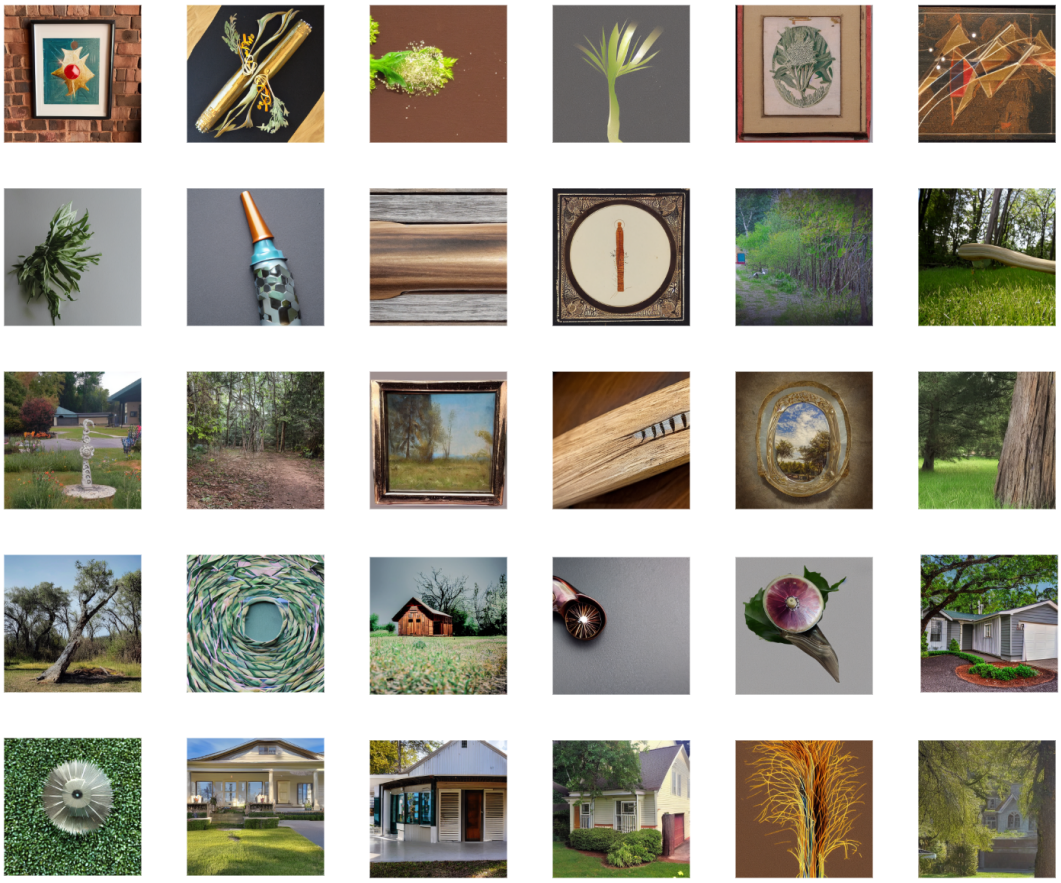

Figure 5: Qualitative results of concept erasure applied to the superclass Guns.

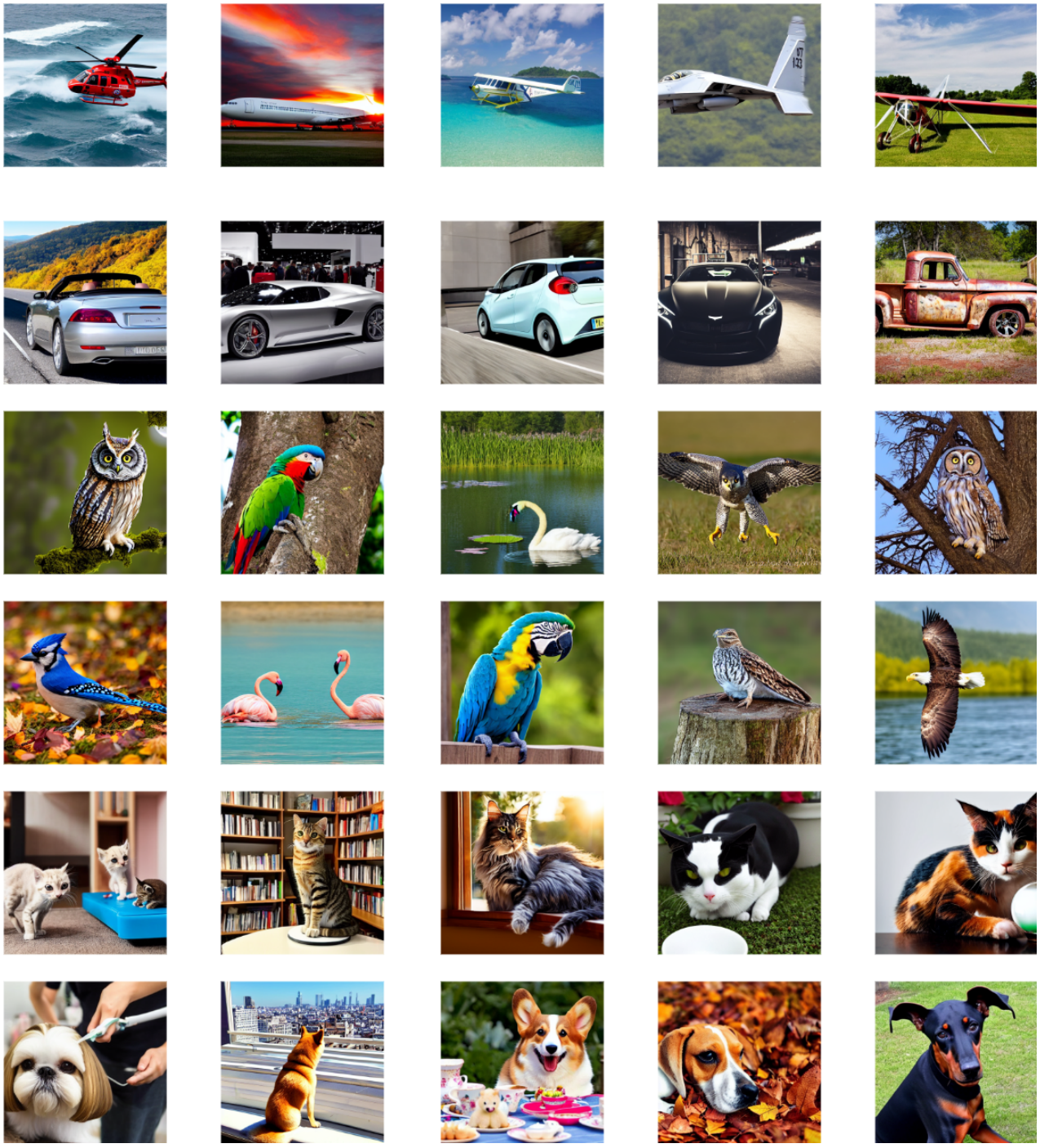

Figure 6: Qualitative results of preservation on the non-target CIFAR-10 classes for the target superclass "Guns". Each row shows a few classes from CIFAR-10, including planes, cars, dogs, birds, and cats.