GIFT: GRADIENT-AWARE IMMUNIZATION OF DIFFUSION MODELS AGAINST MALICIOUS FINE-TUNING WITH SAFE CONCEPTS RETENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models can learn semantic classes, like artistic styles or object categories, which model owners may need to remove to prevent harmful or unauthorized outputs. However, existing concept-erasure methods are brittle; adversaries can easily recover the erased concepts using simple fine-tuning attacks. This vulnerability occurs because erasure methods only increase the loss gap for the target concept but leave the underlying representational structure intact, which makes the concept easily "reachable" during fine-tuning. To address this, we introduce GIFT, an immunization technique designed to attack both the loss landscape and the hidden representation of a target concept, leading to a tamper-resistant model. GIFT pairs gradient ascent on the denoising loss with representation noising, a technique that pulls intermediate attention activations toward Gaussian noise to degrade the reachability of the target concept. To maintain general model utility on safe concepts, GIFT anchors internal representations to a frozen base model and utilizes gradient surgery to resolve optimization conflicts between the erasure and preservation updates. Extensive evaluations across art styles and object categories demonstrate that GIFT achieves superior initial erasure and remains significantly more robust against fine-tuning attacks compared to state-of-the-art baselines, all while maintaining fidelity on safe-concept generation.

1 INTRODUCTION

Text-to-image diffusion models generate realistic imagery from natural language at consumer scale (Rombach et al., 2022; Ho et al., 2020). The open release of model weights, alongside lightweight personalization techniques such as LoRA (Hu et al., 2022) and DreamBooth (Ruiz et al., 2023), has driven rapid adoption across creative and commercial workflows. The same accessibility lowers the barrier to misuse: pretrained checkpoints reproduce sexually explicit content, copyrighted artistic styles, and likenesses of private individuals, often from prompts that read as benign (Schramowski et al., 2023; Schuhmann et al., 2022; Li et al., 2024). With fine-tuning APIs and "fine-tuning-as-a-service" platforms now available, a small adversarial budget suffices to steer a base model toward a harmful concept (Zheng & Yeh, 2024; Li et al., 2025). Early defenses sit outside the model. Input- and output-side filters screen prompts and generations at inference time (Rando et al., 2022), while pretraining-data curation removes harmful samples before training. Both strategies leave the underlying generative capability intact (Yang et al., 2024; Zhang et al., 2024b; Gao et al., 2024): filters could be disabled or bypassed once weights ship open-source, and curation is prone to prompt jail-breaking and fine-tuning attacks (Cretu et al., 2026) and cannot retroactively repair an already-trained checkpoint.

Concept erasure addresses the problem at its source by editing model parameters so that prompts for the target concept c_t no longer recover faithful generations (Gandikota et al., 2023; Kumari et al., 2023; Gandikota et al., 2024; Heng & Soh, 2023; Huang et al., 2024a). Although such techniques can suppress the generation of undesired concepts, they are vulnerable to circumvention (Pham et al., 2024). Mechanistic analyses of safety alignment in LLMs reveal that such interventions often act as shallow output-level re-routings while the underlying representations persist (Wei et al., 2024; Jain et al., 2023; Arditi et al., 2024). The diffusion analogue manifests as fine-tuning brittleness: a

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

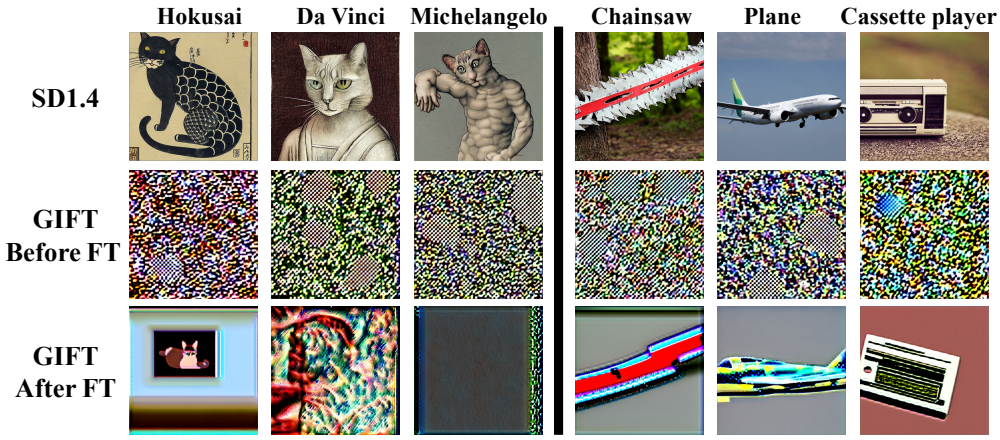


Figure 1: **GIFT erases concepts and resists recovery under fine-tuning attacks.** Top: the undefended Stable Diffusion v1.4 baseline generates target concepts faithfully across three artistic styles (Hokusai, Da Vinci, Michelangelo) and three object classes (chainsaw, plane, cassette player). Middle: after immunization with GIFT, the same prompts collapse into structureless noise. Bottom: even after a 5000-step LoRA fine-tuning attack, recovered generations remain heavily distorted and far from the original concepts. The robustness generalizes across both concept families.

few hundred gradient steps on a small image set recover the erased concept (Zheng & Yeh, 2024; Li et al., 2025), and we confirm these limitations in section 4.

To address this fundamental weakness, recent research introduces model immunization, termed tamper-resistance, to proactively defend against harmful fine-tuning attacks (Zheng & Yeh, 2024; Henderson et al., 2023; Rosati et al., 2024b; Tamirisa et al., 2024). The LLM literature has explored this direction through Self-Destructing Models (Henderson et al., 2023), Vaccine (Huang et al., 2024b), Tamper-Resistant Safeguards (Tamirisa et al., 2024), and Representation Noising (Rosati et al., 2024a), each targeting the relearning trajectory rather than only the loss at $\theta^\star$. For diffusion models, IMMA (Zheng & Yeh, 2024) introduced immunization as a bi-level optimization approach that learns poor model initializations that hinder adaptation to undesirable concepts. By simulating the fine-tuning process, IMMA makes it more difficult for adversaries to reintroduce malicious content through fine-tuning. However, our experiments show that IMMA still re-learns the target distribution under modest LoRA attack, leaving substantial room for stronger immunization.

We argue this gap arises because IMMA (Zheng & Yeh, 2024) inherits the loss-only view of erasure: they raise the post-adaptation loss but leave intact the internal representations that enable the relearning trajectory. A trajectory analysis of harmful fine-tuning (Rosati et al., 2024a) factors the relearning probability into a *static-potential* term tied to the loss gap and a *reachability* term tied to the gradient field along the path. Suppressing only the static term hands the reachability term to the adversary.

To attack both factors, we propose **GIFT**, a tamper-resistant concept erasure framework Figure 1). Following prior work in the LLMs domain (Achille & Soatto, 2018; Rosati et al., 2024a), GIFT pairs gradient ascent on the diffusion denoising loss, which inflates the loss gap on the target concept c_t (targeting static potential), with a representation noising term that pulls intermediate attention activations toward Gaussian noise (targeting reachability). To prevent the erasure signal from corrupting safe-concept generation, we apply a preservation objective comprising two terms. The first term minimizes the standard generation loss on a safe dataset D_s, while the second term tightly anchors the attention-module activations directly to the frozen reference UNet θ_0. The erasure and preservation objectives draw on disjoint data and produce conflicting parameter updates; we resolve the conflict via gradient surgery (Yu et al., 2020) that projects out only the component of the erasure gradient opposing preservation, which reduces the effect of erasure on safe model capabilities.

We summarize our contributions as follows. **(1)** We propose **GIFT**, a tamper-resistant immunization framework for text-to-image diffusion models that pairs gradient ascent on the denoising loss with representation noising at attention modules. **(2)** To preserve safe-concept generation, we introduce a complementary preservation objective combining a standard denoising loss on D_s with an activation-anchoring term that ties intermediate attention outputs to the frozen reference θ_0, and we resolve the resulting conflict between erasure and preservation gradients via projection that removes the component of the erasure gradient opposing preservation to limit collateral damage to safe model capabilities. **(3)** Through extensive experiments across fourteen artistic styles and ten ImageNet object classes, GIFT achieves the strongest pre-attack erasure and remains substantially farther from the target distribution after a harmful fine-tuning attack, surpassing both concept-erasure baselines and the prior immunization approach while preserving fidelity on safe concepts.

2 Related Work

Text-to-image diffusion models. Diffusion models generate images by reversing a fixed forward process that gradually corrupts data with Gaussian noise across T timesteps (Ho et al., 2020). Training fits a neural predictor $\epsilon_\theta(x_t, t, p)$ that recovers the noise added at step t given the corrupted input x_t and a conditioning prompt p, optimized through the MSE objective in eq. (1). Latent diffusion models, including Stable Diffusion, run this process in the latent space of a pretrained autoencoder, which cuts compute cost while retaining sample quality (Rombach et al., 2022). Some variants reparameterize the target as the velocity v rather than ϵ for improved sampling stability (Salimans & Ho, 2022), though the noise-prediction form remains the same.

Concept erasure in diffusion models. A growing line of work edits diffusion weights to suppress specific concepts. ESD fine-tunes the UNet so the conditional score on a target prompt matches the unconditional score (Gandikota et al., 2023); Ablating Concepts redirects target prompts toward an anchor distribution (Kumari et al., 2023); UCE applies closed-form edits to cross-attention projections for efficient multi-concept removal (Gandikota et al., 2024); Selective Amnesia frames erasure as continual forgetting under an EWC-style penalty (Heng & Soh, 2023); Receler trains lightweight eraser modules with adversarial prompts (Huang et al., 2024a); SLD steers generations away from unsafe regions through guidance modifications at inference (Schramowski et al., 2023). Although these methods are successful at suppressing a concept, Pham et al. (2024) circumvent multiple erasure methods by recovering the suppressed concept through textual inversion, and Zhang et al. (2024b) show that unlearned diffusion models still generate unsafe images under modest prompt perturbations. Our experiments show that these methods fail under fine-tuning attacks: they relearn the target concept distribution after only a few fine-tuning steps.

Tamper-resistance and harmful fine-tuning defenses. The LLM community has formalized harmful fine-tuning as the dominant threat against released open-weight models and developed defenses that target the relearning trajectory rather than the loss alone (Henderson et al., 2023; Huang et al., 2024b; Tamirisa et al., 2024; Zhou et al., 2023; Rosati et al., 2024a). Representation Noising minimizes input-representation mutual information at intermediate layers to flatten the gradient field along recovery paths (Rosati et al., 2024a), providing the conceptual frame that GIFT adapts to diffusion. For diffusion models, IMMA (Zheng & Yeh, 2024) learns initializations that resist personalization through bi-level optimization, and Li et al. (2025) pursues resilient unlearning against downstream fine-tuning on nudity. We compare directly against IMMA in section 4 and demonstrate that GIFT delivers stronger resistance than IMMA and prior erasure works.

Gradient conflict and representation anchoring. The erasure and preservation objectives of GIFT pull shared parameters in opposing directions, which connects our optimization to multi-task and continual-learning methods that handle gradient conflict explicitly. PCGrad projects each task gradient onto the orthogonal complement of any conflicting peer (Yu et al., 2020). GEM (Lopez-Paz & Ranzato, 2017) and A-GEM (Chaudhry et al., 2019) constrain updates so they do not increase loss on stored memory examples. OGD projects gradients onto subspaces orthogonal to past task gradients to limit forgetting (Farajtabar et al., 2020). GIFT applies an asymmetric variant of PCGrad that projects only the erasure gradient when it opposes preservation. The activation-anchoring term $\mathcal{L}_{\text{act}}$ continues a long tradition of feature-level distillation in continual learning, originating with Learning without Forgetting (Li & Hoiem, 2016), but applies the anchor exclusively on safe data while pushing target activations toward noise on disjoint inputs.

3 METHOD

3.1 PROBLEM FORMULATION

A *concept* denotes a semantic class learnable by a text-to-image diffusion model: a character, an artistic style, or an object category. We treat each concept as a distribution over image and prompt pairs (x, p) (Gandikota et al., 2023; Kumari et al., 2023). The *target concept* c_t, with samples $\mathcal{D}_t = \{(x_i, p_i)\}_{i=1}^{N_t}$, denotes the concept the defender wishes to remove. The *safe concepts* c_s, with samples $\mathcal{D}_s = \{(x_j, p_j)\}_{j=1}^{N_s}$, denote a set of safe concepts used to preserve general model utility on other concepts. We call the output of our defense procedure a *tamper-resistant model* $\theta^\star$, derived from a base diffusion model θ_0.

Notation. Let $\epsilon_\theta(x_t, t, p)$ denote the noise prediction of a latent diffusion model with UNet parameters θ at timestep t given prompt p. The standard denoising loss reads

$$\mathcal{L}_{\mathrm{dn}}(\theta;\, x, p) \;=\; \mathbb{E}_{t,\,\epsilon\sim\mathcal{N}(0,I)}\Big[\,\big\|\epsilon_\theta(x_t, t, p) - \epsilon\big\|_2^2\,\Big] \tag{1}$$

(Ho et al., 2020; Rombach et al., 2022). We capture intermediate activations at the output projection of every attention module in the UNet, both self-attention (`attn1.to_out`) and cross-attention (`attn2.to_out`). For an input (x, p), $z_\ell(\theta;\, x, p) \in \mathbb{R}^{B\times S\times d_\ell}$ denotes the layer-ℓ activation, and $Z(\theta;\, x, p) = \{z_\ell\}_{\ell=1}^{L}$ collects activations across all L attention modules.

Threat model. We adapt the harmful fine-tuning attack frame (Henderson et al., 2023; Rosati et al., 2024b;a) to diffusion models. The adversary obtains the defended weights $\theta^\star$ and runs fine-tuning on a target dataset $\mathcal{D}_{\mathrm{adv}} \sim c_t$, producing weights $\theta(t)$ initialized at $\theta(0) = \theta^\star$. A successful defense satisfies two conditions: *Resistance*: the adversary needs substantially more fine-tuning steps to recover c_t from $\theta^\star$ than from θ_0. *Stability*: generation quality on $\mathcal{D}_s$ nearly matches θ_0.

3.2 METHOD OVERVIEW

We learn $\theta^\star$ by minimizing the composite objective

$$\mathcal{L}(\theta) \;=\; \underbrace{\mathcal{L}_{\mathrm{ascent}}(\theta;\, \mathcal{D}_t) + \mathcal{L}_{\mathrm{noise}}(\theta;\, \mathcal{D}_t)}_{\mathcal{L}_{\mathrm{TRE}}\,:\,\text{tamper-resistant erasure}} + \lambda_s \underbrace{\big(\mathcal{L}_{\mathrm{prior}}(\theta;\, \mathcal{D}_s) + \mathcal{L}_{\mathrm{act}}(\theta, \theta_0;\, \mathcal{D}_s)\big)}_{\mathcal{L}_{\mathrm{preserve}}\,:\,\text{safe-concept preservation}} \tag{2}$$

where λ_s controls the preservation strength. $\mathcal{L}_{\mathrm{TRE}}$ removes c_t from both the loss landscape and the internal representations of θ (section 3.3); $\mathcal{L}_{\mathrm{preserve}}$ anchors generation quality and intermediate representations on $\mathcal{D}_s$ to the frozen base model θ_0 (section 3.4). Because the two terms induce parameter updates that routinely conflict (Yu et al., 2020), naive summation lets the erasure update partially cancel preservation; we resolve the conflict via gradient surgery (section 3.5).

3.3 TAMPER-RESISTANT ERASURE

A harmful fine-tuning attack on $\theta^\star$ traces an SGD trajectory in parameter space, starting at $\theta^\star$ and converging toward some $\theta^{(t^*)}$ that minimizes the denoising loss on $\mathcal{D}_{\mathrm{adv}}$. Achille et al. (2018) show that the probability of completing such a trajectory factors into a static-potential term, governed by the loss gap $\mathcal{L}_{\mathrm{dn}}(\theta^{(t^*)}) - \mathcal{L}_{\mathrm{dn}}(\theta^\star)$, and a reachability term (*i.e.*, the existence of such trajectory), governed by the gradient field along the path. Tamper-resistance therefore requires reducing both factors. Prior concept-erasure methods (Gandikota et al., 2023; 2024; Heng & Soh, 2023; Huang et al., 2024a) attack only the static-potential term and leave the representational structure that determines reachability intact, which makes them brittle under fine-tuning as we show in our experiments. We address the two factors with two terms:

Gradient ascent on the denoising loss. The first term inflates the static-potential gap by pushing $\theta^\star$ toward a region of high target-concept loss:

$$\mathcal{L}_{\mathrm{ascent}}(\theta;\, \mathcal{D}_t) \;=\; -\mathbb{E}_{(x_t,p_t)\sim\mathcal{D}_t}\,\mathbb{E}_{t,\epsilon}\Big[\,\big\|\epsilon_\theta(x_t, t, p) - \epsilon\big\|_2^2\,\Big], \tag{3}$$

Gradient ascent on the task loss has long served as a primitive in machine unlearning and concept removal (Golatkar et al., 2020; Graves et al., 2021). Ascent alone, however, fails under modest adversarial budgets: a few fine-tuning steps suffice to relink the structurally intact representations of c_t to the output head (Wei et al., 2024; Jain et al., 2023; Lee et al., 2024; Rosati et al., 2024a). We confirm the analogous failure mode in diffusion in section 4.

Representation noising. The second term attacks the reachability factor. Adopting the information-bottleneck view of deep networks (Tishby & Zaslavsky, 2015; Achille & Soatto, 2018), we write X_t for inputs drawn from $\mathcal{D}_t$, Z_θ for the corresponding intermediate activations, and Y for the noise-prediction target. The data-processing inequality (Cover & Thomas, 2006) gives

$$I(X_t;\, Y) \;\le\; I(X_t;\, Z_\theta), \tag{4}$$

so minimizing the input-representation mutual information $I(X_t; Z_\theta)$ on the target distribution also minimizes $I(X_t;\, Y)$. Rosati et al. (2024a) sharpens this into a control on the reachability term itself. We minimize $I(X_t; Z_\theta)$ by pulling the conditional distribution $p_\theta(z_\ell \mid X_t)$ at each attention module ℓ toward a Gaussian noise using MSE loss, applied layer-wise:

$$\mathcal{L}_{\text{noise}}(\theta;\, \mathcal{D}_t) \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_{(x,p)\sim\mathcal{D}_t} \left\| z_\ell(\theta;\, x, p) \;-\; \eta_\ell \right\|_2^2, \quad \eta_\ell \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I). \tag{5}$$

where η_ℓ matches the shape of z_ℓ. We capture the activations at attention output projection layers. Together, $\mathcal{L}_{\text{ascent}}$ raises the static-potential barrier on $\mathcal{D}_t$ and $\mathcal{L}_{\text{noise}}$ degrades reachability, attacking the two factors of the relearning trajectory simultaneously.

3.4 Safe-Concept Preservation

To retain generation quality on safe concepts, we use two preservation terms operating at different levels. The first enforces output-level fidelity on safe data via the standard denoising loss,

$$\mathcal{L}_{\text{prior}}(\theta;\, \mathcal{D}_s) \;=\; \mathbb{E}_{(x,p)\sim\mathcal{D}_s}\, \mathbb{E}_{t,\epsilon}\Big[\, \big\|\epsilon_\theta(x_t, t, p) - \epsilon\big\|_2^2 \,\Big], \tag{6}$$

the standard preservation primitive in concept-erasure work (Gandikota et al., 2023; Kumari et al., 2023; Gandikota et al., 2024). This term, however, constrains only noise predictions on $\mathcal{D}_s$ and leaves internal representations free to drift, which interacts poorly with $\mathcal{L}_{\text{noise}}$: rep-noising gradients on shared parameters can shift safe-input activations. So, we add a representation anchor that matches activations at every attention module to the frozen reference θ_0:

$$\mathcal{L}_{\text{act}}(\theta, \theta_0;\, \mathcal{D}_s) \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}_{(x,p)\sim\mathcal{D}_s} \left\| z_\ell(\theta;\, x, p) \;-\; z_\ell(\theta_0;\, x, p) \right\|_2^2. \tag{7}$$

The pairing complements representation noising directly. Equation (5) *pushes* activations on $\mathcal{D}_t$ away from any structured form at every attention module, while eq. (7) *anchors* activations on $\mathcal{D}_s$ at the same modules to where θ_0 already places them.

3.5 Resolving Conflict Between Erasure and Preservation

The two components of eq. (2) draw on disjoint data and pull the shared attention parameters in different directions. Let $g_e = \nabla_\theta \mathcal{L}_{\text{TRE}}$ and $g_s = \nabla_\theta(\lambda_s \mathcal{L}_{\text{preserve}})$ denote the erasure and preservation gradients. Naive summation $g = g_e + g_s$ yields an update along which the erasure component partially cancels preservation, degrading $\mathcal{L}_{\text{preserve}}$ with minimal gain in erasure. Yu et al. (2020) identify exactly this regime as the one in which gradient conflict most damages multi-task optimization, characterized by negative cosine similarity between gradients, one task dominating gradient magnitudes, and high curvature of the joint loss landscape; all three conditions hold in our setting (we measure them in section 4).

In our formulation: preservation should not pay for erasure progress, but erasure may pay for preservation. We therefore project only g_e, and only when it conflicts with g_s:

$$\widetilde{g}_e \;=\; \begin{cases} g_e \;-\; \dfrac{\langle g_e,\, g_s\rangle}{\|g_s\|_2^2}\, g_s, & \text{if } \langle g_e,\, g_s\rangle < 0, \\ g_e, & \text{otherwise,} \end{cases} \qquad \nabla_\theta \mathcal{L} \;\leftarrow\; \widetilde{g}_e + g_s. \tag{8}$$

Algorithm 1 Tamper-Resistant Concept Erasure (one optimizer step)

Require: Trainable parameters θ, frozen reference θ_0, batches $B_t \sim \mathcal{D}_t$, $B_s \sim \mathcal{D}_s$

1: Forward θ on B_t, capture activations $Z(\theta;\, B_t)$; compute $\mathcal{L}_{\text{ascent}}$ equation 3 and $\mathcal{L}_{\text{noise}}$ equation 5
2: Forward θ on B_s, capture activations $Z(\theta;\, B_s)$; compute $\mathcal{L}_{\text{prior}}$ equation 6
3: Forward θ_0 on B_s, capture activations $Z(\theta_0;\, B_s)$; compute $\mathcal{L}_{\text{act}}$ equation 7
4: Backward on $\mathcal{L}_{\text{TRE}} = -\mathcal{L}_{\text{ascent}} + \mathcal{L}_{\text{noise}}$; cache $g_e \leftarrow \{\theta.\text{grad}\}$; zero gradients
5: Backward on $\lambda_s \mathcal{L}_{\text{preserve}} = \lambda_s(\mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{act}})$; cache $g_s \leftarrow \{\theta.\text{grad}\}$
6: **for** each parameter tensor p with both gradients present **do**
7: **if** $\langle g_e^{(p)},\, g_s^{(p)} \rangle < 0$ **then** $\widetilde{g}_e^{(p)} \leftarrow g_e^{(p)} - \dfrac{\langle g_e^{(p)},\, g_s^{(p)} \rangle}{\|g_s^{(p)}\|_2^2}\, g_s^{(p)}$
8: **else** $\widetilde{g}_e^{(p)} \leftarrow g_e^{(p)}$
9: **end if**
10: $p.\text{grad} \leftarrow \widetilde{g}_e^{(p)} + g_s^{(p)}$
11: **end for**
12: Apply optimizer step

Projection runs independently on each parameter tensor, so layers where erasure and preservation agree contribute their gradients unmodified while conflicting layers receive local correction.

Algorithm 1 summarizes the per-step procedure: the trainable model runs forward on a target batch and a safe batch, the frozen reference runs forward on the same safe batch, the four losses combine into an erasure gradient and a preservation gradient via two backward passes, and the per-tensor projection of eq. (8) produces the final update.

4 Experiments

4.1 Experimental Setup

Baselines and attack protocol. We compare GIFT against three concept-erasure methods operating on Stable Diffusion v1.4: ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), and IMMA (Zheng & Yeh, 2024). All defenses, including ours, are initialized from the standard `CompVis/stable-diffusion-v1-4` checkpoint and run using official implementations with default hyperparameters. More details in B.1. To test tamper resistance, the adversary fine-tunes the defended weights $\theta^\star$ on an attack set $\mathcal{D}_{\text{adv}}$. Our default attack uses LoRA (Hu et al., 2022) with rank 4, a 10^{-4} learning rate, batch size 1, and 5000 update steps. To ensure resistance generalizes across recipes, we additionally evaluate full-parameter fine-tuning and DreamBooth (Ruiz et al., 2023), and sweep attack hyperparameters (section 4.4). We report additional comparison with baselines at C.2.

Datasets and concepts. We evaluate on two concept families: artistic styles (14 artists) and objects (10 ImageNet classes). For each concept, we generate 200 images using templated prompts. For art styles, we use `"A high quality painting in ⟨artist_name⟩ style"`. For objects we use `"A high-quality image of a ⟨object⟩"`. These are split equally into disjoint immunization ($\mathcal{D}_t$) and attack ($\mathcal{D}_{\text{adv}}$) sets. To anchor safe generation ($\mathcal{D}_s$), we use 139 benign prompt-image pairs from Li et al. (2024). For safe-generation evaluation, we adopt a 1000 text-image subset of MS-COCO (Lin et al., 2014) following (Zhang et al., 2024a). More details in B.2

Metrics. For objects, we measure top-1 and top-5 Attack Success Rate (ASR) using a ResNet-50 (He et al., 2016) classifier. We pair this with CLIP-T (Radford et al., 2021) (prompt alignment) and the LAION aesthetic predictor (Schuhmann et al., 2022) (AS) to detect when an attack yields classifier-recognized but visually unusable images. For style concepts, we measure CLIP-I against $\mathcal{D}_{\text{adv}}$ (visual similarity to the artist), CLIP-T, LPIPS (Zhang et al., 2018) against $\mathcal{D}_{\text{adv}}$ (perceptual distance), and AS. Stronger resistance is indicated by lower CLIP-I, lower CLIP-T, lower AS, and higher LPIPS on $\mathcal{D}_{\text{adv}}$. On the COCO subset, we report CLIP-T and AS to quantify fidelity on safe prompts. More details about the metrics are in B.3.

Table 1: **Art-style main results.** Lower CLIP-T, CLIP-I, and AS together with higher LPIPS indicate stronger resistance; higher CLIP-T and AS on safe concepts indicate better preservation. Best per column in **bold**, second is underlined. GIFT achieves the strongest erasure pre-attack and remains farthest from the artist distribution post-attack while maintaining safe-concept generation.

Method	Target – Before FT Attack				Target – After FT Attack				Safe Concepts	
	CLIP-T↓	CLIP-I↓	AS↓	LPIPS↑	CLIP-T↓	CLIP-I↓	AS↓	LPIPS↑	CLIP-T↑	AS↑
SD (Rombach et al., 2022)	0.290	0.768	5.274	0.681	–	–	–	–	0.260	5.280
ESD (Gandikota et al., 2023)	0.232	0.685	5.232	0.740	0.285	0.794	5.071	0.664	0.251	5.252
UCE (Gandikota et al., 2024)	0.250	0.714	5.188	0.722	0.283	0.766	5.187	0.675	**0.258**	**5.285**
IMMA (Zheng & Yeh, 2024)	0.194	0.673	4.609	**0.895**	0.256	0.775	5.042	0.704	0.246	5.159
GIFT (ours)	**0.161**	**0.663**	**3.891**	0.744	**0.202**	**0.700**	**4.049**	**0.756**	0.257	5.231

Table 2: **Object erasure and resistance to fine-tuning attack.** For each method, we obtain ten object defended checkpoints, then probe each by generating 100 images for each object and report classifier-based attack success rate (ASR) alongside CLIP-T and AS, before and after a 5000-step LoRA attack. Lower ASR-1, ASR-5, CLIP-T, and AS indicate stronger resistance; higher CLIP-T and AS on safe concepts indicate better preservation. Best per column in **bold**, second is underlined.

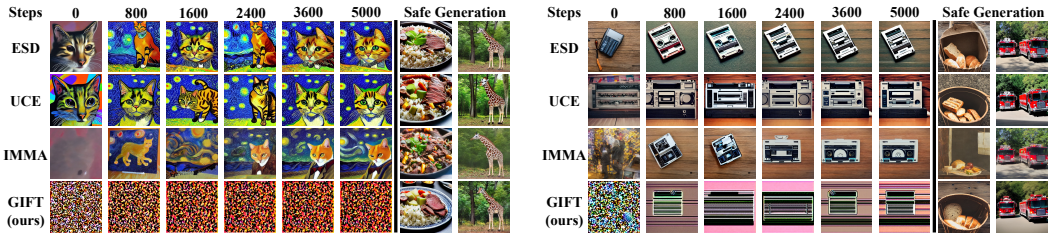
Method	Target – Before FT				Target – After FT				Safe Concepts	
	ASR-1↓	ASR-5↓	CLIP-T↓	AS↓	ASR-1↓	ASR-5↓	CLIP-T↓	AS↓	CLIP-T↑	AS↑
SD (Rombach et al., 2022)	0.739	0.899	0.247	5.238	–	–	–	–	0.260	5.280
ESD (Gandikota et al., 2023)	0.047	0.089	**0.161**	5.020	0.748	0.928	0.250	5.286	0.248	**5.250**
UCE (Gandikota et al., 2024)	0.232	0.531	0.207	5.077	0.723	0.956	0.252	5.256	**0.252**	5.249
IMMA (Zheng & Yeh, 2024)	**0.000**	0.001	0.162	4.633	0.450	0.689	0.229	5.192	0.236	5.049
GIFT (ours)	**0.000**	**0.000**	0.162	**4.020**	**0.144**	**0.297**	**0.225**	**4.140**	**0.252**	5.213

Implementation details. GIFT restricts trainable parameters to the projection matrices of *all* attention modules (both self and cross-attention). We train for 600 steps using AdamW (Loshchilov & Hutter, 2019) with a 10^{-4} learning rate, batch size 1, and a safe-weight $\lambda_s = 2.5$. All experiments run on a single 48GB NVIDIA L40S GPU. See B.5.

4.2 Experimental Results

Art-style erasure and resistance. Table 1 reports erasure quality and attack resistance aggregated across the 14 artist checkpoints. To evaluate, we probe each model to generate 100 images with the prompt "`a painting of a cat in ⟨artist⟩ style`" before and after a 5000-step LoRA attack. Pre-attack, GIFT achieves the strongest erasure across CLIP-T, CLIP-I, and AS. While IMMA scores higher on LPIPS but with smaller gains on the other metrics. Post-attack, ESD and UCE regress close to the standard SD baseline performance, confirming that concepts erased by these methods resurface easily under fine-tuning attacks. IMMA demonstrates partial resistance but still drifts toward the target artist distribution as shown by the increase in the post-attack metrics. In contrast, GIFT retains the lowest CLIP-T, CLIP-I, and AS, and the highest LPIPS remaining furthest from the artist distribution even after an attack budget that fully breaks baseline defenses. On the safe COCO subset, GIFT matches the baselines on CLIP-T with a modest AS drop, a small trade-off relative to the resistance gain. Figure 2a presents qualitative results for this gap: under a 5000-step LoRA attack on Van Gogh's style, prior defenses quickly recover the target concept (before 800 steps), whereas GIFT holds throughout while preserving fidelity on unrelated prompts from the COCO subset. We report SDXL results at C.1

Object erasure and resistance. Table 2 summarizes performance across the 10 ImageNet object classes. Pre-attack, both GIFT and IMMA reduce ASR-1 to zero. However, GIFT also achieves zero ASR-5 and the lowest AS. Post-attack, the performance gap widens significantly. ESD and UCE recover most of the concept, with ASR-1 and ASR-5 rebounding near and even bypassing the undefended SD baseline. IMMA resists more strongly but still surrenders the majority of previously



(a) Van Gogh style erasure.

(b) Cassette player erasure.

Figure 2: Qualitative comparison of concept-erasure defenses under a 5000-step LoRA fine-tuning attack. Columns show generations at different attack steps. The two rightmost columns of each panel show safe-prompt generations. ESD, UCE, and IMMA quickly recover the target concept, while GIFT keeps generations noisy across the full attack budget without sacrificing safe-prompt fidelity.

Table 3: **Erasure loss ablation.** Each row ablates a $\mathcal{L}_{\text{TRE}}$ term.

	Safe Generation		After FT			
Variant	CLIP-T↑	AS↑	CLIP-T↓	CLIP-I↓	AS↓	LPIPS↑
SD v1.4 (no defense)	0.260	5.28	0.29	0.74	5.02	0.70
$\mathcal{L}_{\text{ascent}}$ only	0.256	5.08	0.24	0.73	**3.86**	0.70
$\mathcal{L}_{\text{noise}}$ only	**0.258**	**5.30**	0.27	0.72	4.71	0.73
GIFT ($\mathcal{L}_{\text{ascent}} + \mathcal{L}_{\text{noise}}$)	0.257	5.23	**0.22**	**0.70**	4.11	**0.76**

Table 4: **Safe-preservation ablation.** Each row ablates one component of $\mathcal{L}_{\text{preserve}}$.

	Safe Generation		After FT			
Variant	CLIP-T↑	AS↑	CLIP-T↓	CLIP-I↓	AS↓	LPIPS↑
$\mathcal{L}_{\text{prior}}$ only	0.256	5.01	0.20	0.67	3.99	0.76
$\mathcal{L}_{\text{act}}$ only	0.256	5.02	0.20	0.64	4.02	0.84
prior + act, no surgery	0.255	4.93	0.20	0.72	4.06	0.73
GIFT (all three)	**0.257**	**5.23**	0.22	0.70	4.11	0.76

erased cases. GIFT maintains the lowest ASR-1 and ASR-5 by a substantial margin, alongside the lowest CLIP-T and AS while preserving fidelity on unrelated prompts from the COCO subset.

4.3 Ablation Studies

We isolate the contribution of each design choice in GIFT by ablating one axis at a time on the Claude Monet style. We hold the remaining settings at their defaults from section 4.1.

Table 3 ablates the two terms that compose $\mathcal{L}_{\text{TRE}}$. Using $\mathcal{L}_{\text{ascent}}$ alone leaves the concept semantically reachable while degrading safe generation. $\mathcal{L}_{\text{noise}}$ alone also allows the attack to restore the concept. GIFT combines both terms and dominates on three of four post-attack metrics while maintaining safe generation quality. Both terms attack the static-potential and reachability factors of the relearning trajectory respectively (section 3.3), and removing either one surrenders the corresponding factor to the adversary.

Table 4 removes one component of $\mathcal{L}_{\text{preserve}}$ at a time. Using $\mathcal{L}_{\text{prior}}$ alone or $\mathcal{L}_{\text{act}}$ alone reduces safe generation quality. Disabling gradient surgery while keeping both preservation losses yields the largest single drop, with safe generation AS falling to 4.93. The conflict between erasure and preservation gradients (section 3.5) thus drives measurable collateral damage, and projecting out the conflicting component recovers the lost utility. The reduced variants record slightly stronger post-attack numbers because the weakened preservation lets the erasure update push harder on $\mathcal{D}_t$, but the gain costs roughly 0.2 AS on safe generation, an unfavorable trade for a deployable defense.

Table 5 varies the trainable-parameter set. Self-attention alone fails under attack, while full UNet training collapses safe generation quality. Cross-attention alone resists the attack but slightly degrades safe generation. Targeting both attention types balances utility with tamper resistance.

Table 6 sweeps the immunization learning rate. Although all of them resist re-learning, high learning rates degrade safe generation quality, while low ones allow the attack to relearn the target concept. The default 10^{-4} balances both axes. 5×10^{-4} produces moderately stronger post-attack resistance but pays a small safe generation penalty.

4.4 Different Attack Recipes

Table 5: **Parameter ablation.** Targeting attention modules preserves utility while resisting the attack.

	Safe Generation		After FT			
Trainable parameters	CLIP-T↑	AS↑	CLIP-T↓	CLIP-I↓	AS↓	LPIPS↑
Cross-attention only	0.257	5.02	0.21	0.65	4.09	0.71
Self-attention only	0.259	5.28	0.29	0.78	4.88	0.68
Full UNet	0.245	4.99	0.17	0.61	4.64	0.98
GIFT (cross + self)	0.257	5.23	0.22	0.70	4.11	0.76

Table 6: **Defender learning-rate ablation.** Effect of different learning rates on defense and safe generation.

	Safe Generation		After FT			
Learning rate	CLIP-T↑	AS↑	CLIP-T↓	CLIP-I↓	AS↓	LPIPS↑
1×10^{-3}	0.189	4.47	0.19	0.67	3.62	0.76
5×10^{-4}	0.248	5.05	0.17	0.64	3.96	0.83
1×10^{-4} (GIFT)	0.257	5.23	0.22	0.70	4.11	0.76
5×10^{-5}	0.258	5.13	0.27	0.72	4.64	0.70

Table 7: **Robustness to LoRA attack hyperparameters on Van Gogh.** We attack the same GIFT-immunized checkpoint while sweeping (a) attack learning rate, (b) attack rank, and (c) attack steps. Default attack uses rank 4, LR 10^{-4}, and 5,000 steps; (c) holds rank 16 and LR 5×10^{-4} as a stress test. Lower CLIP-T, CLIP-I, AS and higher LPIPS indicate stronger resistance. Tamper resistance holds across the full sweep.

(a) Attack LR

LR	CLIP-T↓	CLIP-I↓	AS↓	LPIPS↑
5×10^{-4}	0.17	0.70	3.72	0.84
1×10^{-4}	0.16	0.71	3.88	0.75
5×10^{-5}	0.18	0.68	3.70	0.82

(b) Attack rank

Rank	CLIP-T↓	CLIP-I↓	AS↓	LPIPS↑
16	0.16	0.70	3.77	0.82
8	0.16	0.70	3.75	0.85
4	0.16	0.71	3.88	0.75

(c) Attack steps

Steps	CLIP-T↓	CLIP-I↓	AS↓	LPIPS↑
5,000	0.17	0.69	3.62	0.80
10,000	0.18	0.69	3.59	0.81
25,000	0.17	0.67	3.57	0.79

Different fine-tuning attacks. We test whether GIFT's tamper resistance generalizes beyond LoRA by attacking the Van Gogh checkpoint with three fine-tuning recipes (fig. 3). LoRA uses rank 4 at LR 10^{-4} for 5000 steps. DreamBooth and full-parameter fine-tuning both also run for 5000 steps at LR 10^{-4}, The rest of the hyperparameters are matching the defaults from the `diffusers` library. Across all three methods, generations remain visually unrelated to the target style, demonstrating that the immunization holds under fine-tuning recipes structurally distinct from the one used during defense.

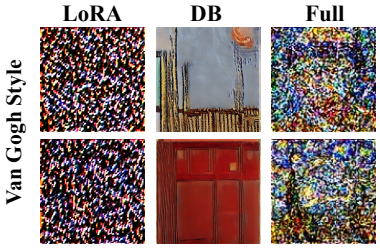


Figure 3: **GIFT resists fine-tuning attacks beyond LoRA.**

Different LoRA Parameters. We stress-test the Van Gogh checkpoint immunized with the default settings from section 4.1 against a grid of stronger LoRA attacks. We sweep three axes independently: rank, learning rate, and steps. To simulate a determined adversary, we further combine the strongest setting from each axis (rank 16, LR 5×10^{-4}, 25000 steps) into a single attack (table 7).

5 Conclusion

We presented GIFT, a tamper-resistant immunization framework for text-to-image diffusion models. Prior erasure methods widen the loss gap on the target concept but leave its representational structure intact, so a small adversarial budget recovers the concept through fine-tuning. GIFT instead attacks both factors of the relearning trajectory: gradient ascent on the denoising loss inflates the static-potential barrier, and representation noising at attention modules degrades reachability by pulling target activations toward Gaussian noise. A complementary preservation objective protects safe-concept generation through an output-level denoising prior and an activation anchor to the frozen base model, and gradient surgery resolves the conflict between erasure and preservation updates by projecting out the component of the erasure gradient that opposes preservation. Across fourteen artistic styles and ten ImageNet object classes, GIFT delivers stronger initial erasure than ESD, UCE, and IMMA, and remains substantially farther from the target distribution after a 5000-step LoRA attack while preserving fidelity on safe-concept generation. The resistance holds across LoRA, DreamBooth, and full-parameter fine-tuning, and across sweeps over learning rate, rank, and step count. We hope this work motivates immunization as a first-class objective for safer open-weight diffusion models and to other modalities.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript and for writing simple shell scripts to automate experiment running. We have not used generative AI tools for generating experimental results, tables, figures, or conducting the literature review. We have reviewed all AI-assisted work and verified the accuracy and content of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the implementation details, training configurations, evaluation protocols, and hyperparameters necessary to reproduce our experiments. We have taken the following steps to make this work reproducible. The complete source code is included as supplementary material: it contains the immunization code for both backbones, the three fine-tuning attacks used in the paper, the evaluation metrics, and two end-to-end scripts to run immunization. The concept list, the prompts used to build each concept's defender and attacker splits, and the script that generates and the generated image sets will be released on publication. For the baselines we report the settings we used in the Appendix.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Alessandro Achille, Glen Mbeng, and Stefano Soatto. Dynamics and reachability of learning tasks. *arXiv preprint arXiv:1810.02440*, 2018.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations (ICLR)*, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006.

Ana-Maria Cretu, Klim Kireev, Amro Abdalla, Wisdom Obinna, Raphael Meier, Sarah Adel Bargal, Elissa M. Redmiles, and Carmela Troncoso. Evaluating concept filtering defenses against child sexual abuse material generation by text-to-image models, 2026. URL https://arxiv.org/abs/2512.05707.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

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.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2131–2141. IEEE, 2025.

Sensen Gao, Xiaojun Jia, Yihao Huang, Ranjie Duan, Jindong Gu, Yang Bai, Yang Liu, and Qing Guo. Hts-attack: Heuristic token search for jailbreaking text-to-image models, 2024. URL https://arxiv.org/abs/2408.13896.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11516–11524, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Peter Henderson, Eric Mitchell, Christopher D. Manning, Dan Jurafsky, and Chelsea Finn. Self-destructing models: Increasing the costs of harmful dual uses of foundation models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36:17170–17194, 2023.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 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 (ECCV)*, 2024a.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*, 2023.

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.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. In *International Conference on Machine Learning (ICML)*, 2024.

Boheng Li, Renjie Gu, Junjie Wang, Leyi Qi, Yiming Li, Run Wang, Zhan Qin, and Tianwei Zhang. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Boheng Li, Renjie Gu, Junjie Wang, Leyi Qi, Yiming Li, Run Wang, Zhan Qin, and Tianwei Zhang. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. *Advances in Neural Information Processing Systems*, 38:17754–17789, 2026.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 4807–4821, 2024.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision (ECCV)*, 2016.

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 *European Conference on Computer Vision (ECCV)*, 2014.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ag3o2T51Ht.

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. In *International Conference on Machine Learning (ICML)*, 2021.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation noising: A defence mechanism against harmful finetuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5234–5247, 2024b.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, et al. Tamper-resistant safeguards for open-weight llms. *arXiv preprint arXiv:2408.00761*, 2024.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*, 2015.

Lixu Wang, Shichao Xu, Ruiqi Xu, Xiao Wang, and Qi Zhu. Non-transferable learning: A new approach for model ownership verification and applicability authorization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=tYRrOdSnVUy.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *International Conference on Machine Learning (ICML)*, 2024.

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 897–912, 2024. doi: 10.1109/SP54263.2024.00123.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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, 2024b. URL https://arxiv.org/abs/2310.11868.

Amber Yijia Zheng and Raymond A Yeh. Imma: Immunizing text-to-image models against malicious adaptation. In *European Conference on Computer Vision*, pp. 458–475. Springer, 2024.

Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. Making harmful behaviors unlearnable for large language models. *arXiv preprint arXiv:2311.02105*, 2023.

Appendix

A Theoretical Justification for Representation Noising

This section justifies why representation noising in eq. (5) attacks the relearning trajectory at both factors. The argument adapts Rosati et al. (2024a, Theorem 1) from autoregressive language modeling to diffusion: a single bridging lemma transports the cross-entropy machinery to the Gaussian-MSE setting, after which the original mutual-information argument applies.

A.1 Setup

We follow the notation of section 3. Let X_t denote the random variable (x_t, t, p) drawn from the target-concept distribution $\mathcal{D}_t$ and the diffusion timestep distribution, $Y = \epsilon$ the noise-prediction target, and $\hat{Y}_\theta = \epsilon_\theta(X_t)$ the network's prediction. The activations at the L attention output projections form the stack $Z_\theta = \{z_\ell(\theta;\, X_t)\}_{\ell=1}^{L}$ and depend on X_t deterministically through the network. The network therefore induces the Markov chain $X_t \to Z_\theta \to \hat{Y}_\theta$, and the data-processing inequality (Cover & Thomas, 2006) gives

$$I(X_t;\, \hat{Y}_\theta) \;\le\; I(X_t;\, Z_\theta). \tag{9}$$

This inequality means: any control we exert on $I(X_t;\, Z_\theta)$ at attention modules transfers downstream to a control on $I(X_t;\, \hat{Y}_\theta)$. We also write $H_\theta(Y \mid \hat{Y}_\theta) = -\mathbb{E}[\log p_\theta(Y \mid \hat{Y}_\theta)]$ for the conditional cross-entropy of the target Y given the prediction $\hat{Y}_\theta$.

Trajectory decomposition. A harmful fine-tuning attack traces an SGD trajectory through parameter space. Achille et al. (2018, Eq. 10) characterize the probability of completing such a trajectory as

$$p\big(\theta^{(t^*)},\, t^* \,\big|\, \theta^\star,\, 0\big) \;\approx\; \underbrace{\exp\{-\Delta\mathcal{L}_{\mathrm{dn}}\}}_{p_{\mathrm{static}}} \cdot \underbrace{\int_{\theta^\star}^{\theta^{(t^*)}} \exp\left\{-\int_0^{t^*} |\nabla\mathcal{L}_{\mathrm{dn}}(\theta_t)|^2\, dt\right\} d\theta(t)}_{p_{\mathrm{reach}}}, \tag{10}$$

where $\Delta\mathcal{L}_{\mathrm{dn}} = \mathcal{L}_{\mathrm{dn}}(\theta^{(t^*)}) - \mathcal{L}_{\mathrm{dn}}(\theta^\star)$. The static-potential factor shrinks as the loss gap widens, and the reachability factor shrinks as gradient magnitudes along the path decrease. A tamper-resistant defense must therefore reduce both, and our claim is that representation noising does exactly this.

A.2 Gaussian-MSE Bridge

Rosati et al. (2024a) reason about cross-entropy loss because their setting is autoregressive language modeling. To apply their argument to diffusion, we connect the diffusion denoising loss to a Gaussian log-likelihood. The connection is standard: under an isotropic Gaussian noise model, MSE minimization coincides with maximum likelihood up to constants that do not depend on θ, so any optimization step that raises $\mathcal{L}_{\mathrm{dn}}$ raises the cross-entropy $H_\theta(Y \mid \hat{Y}_\theta)$ by a fixed scalar.

Lemma 1 (Gaussian-MSE log-likelihood equivalence)**.** *Let $p_\theta(Y \mid \hat{Y}_\theta) = \mathcal{N}(\hat{Y}_\theta,\, \sigma^2 I)$, the standard isotropic Gaussian noise model with mean $\hat{Y}_\theta$ and variance σ^2. Then the diffusion denoising loss equals the conditional cross-entropy of Y given $\hat{Y}_\theta$ up to a θ-independent additive constant:*

$$\mathcal{L}_{\mathrm{dn}}(\theta;\, X_t) \;=\; 2\sigma^2\, H_\theta(Y \mid \hat{Y}_\theta) \;+\; C, \tag{11}$$

where C depends only on σ and the dimension of Y.

Proof. The isotropic Gaussian log density gives $\log p_\theta(Y \mid \hat{Y}_\theta) = -\frac{1}{2\sigma^2}\|Y - \hat{Y}_\theta\|_2^2 - \frac{d}{2}\log(2\pi\sigma^2)$. Taking expectations over X_t, Y and rearranging yields eq. (11) with $C = -\sigma^2 d \log(2\pi\sigma^2)$. □

Two consequences follow. Since C does not depend on θ, the gradients of $\mathcal{L}_{\mathrm{dn}}$ and of $H_\theta(Y \mid \hat{Y}_\theta)$ point in the same direction up to the scalar $2\sigma^2$. And any rise in cross-entropy translates to a proportional rise in the denoising loss. The cross-entropy machinery of Rosati et al. (2024a) therefore applies to $\mathcal{L}_{\mathrm{dn}}$ as well.

A.3 Main Result

With the bridge in place, both factors of eq. (10) fall to the same observation: a smaller $I(X_t;\, Z_\theta)$ propagates through the Markov chain to a smaller $I(X_t;\, \hat{Y}_\theta)$, which simultaneously raises the loss gap and shrinks the gradient magnitudes available to the adversary.

Theorem 1 (Mutual-information minimization suppresses both factors)**.** *Reducing $I(X_t;\, Z_{\theta^\star})$ on $\mathcal{D}_t$ suppresses both p_{static} and p_{reach} in eq.* (10)*.*

Proof. We treat the two factors in turn.

Static potential. Equation (9) bounds $I(X_t;\, \hat{Y}_\theta)$ by $I(X_t;\, Z_\theta)$. A smaller $I(X_t;\, \hat{Y}_\theta)$ leaves the prediction with less information about X_t, which raises the cross-entropy $H_\theta(Y \mid \hat{Y}_\theta)$ on $\mathcal{D}_t$ (Rosati et al., 2024a, Theorem 1). Lemma 1 translates this rise in cross-entropy to a rise in $\mathcal{L}_{\mathrm{dn}}(\theta^\star;\, \mathcal{D}_{\mathrm{adv}})$. The loss gap $\Delta\mathcal{L}_{\mathrm{dn}}$ widens accordingly, and p_{static} shrinks.

Reachability. The reachability factor grows with $\|\nabla\mathcal{L}_{\mathrm{dn}}(\theta_t)\|$ along the trajectory (Achille et al., 2018, Eq. 14). Lemma 1 aligns $\nabla\mathcal{L}_{\mathrm{dn}}$ with $\nabla H_\theta(Y \mid \hat{Y}_\theta)$, whose magnitude scales with $I(X_t;\, \hat{Y}_\theta)$ (Wang et al., 2022, Theorem 2), and eq. (9) bounds this in turn by $I(X_t;\, Z_\theta)$. Reducing $I(X_t;\, Z_{\theta^\star})$ therefore caps the gradient magnitudes accessible along any attack trajectory rooted at $\theta^\star$, and p_{reach} shrinks. □

Theorem 1 justifies the design of $\mathcal{L}_{\mathrm{noise}}$. Acting on $I(X_t;\, Z_\theta)$ controls both factors of the relearning trajectory at once, whereas $\mathcal{L}_{\mathrm{ascent}}$ alone touches only p_{static}. The estimator in eq. (5) drives the conditional distribution $p_\theta(z_\ell \mid X_t)$ toward $\mathcal{N}(0, I)$ via moment matching, and the two distributions match exactly when $I(X_t;\, z_\ell) = 0$. Pairing $\mathcal{L}_{\mathrm{noise}}$ with $\mathcal{L}_{\mathrm{ascent}}$ therefore attacks the static potential through the loss directly while attacking reachability through the representations, which is the design we evaluate empirically in section 4.

B Experimental Setup Details

This appendix expands the experimental setup summarized in section 4.1. We detail baselines, datasets, evaluation metrics, fine-tuning attack recipes, and the implementation choices behind GIFT.

B.1 Baselines and Diffusion Backbone

We immunize the `CompVis/stable-diffusion-v1-4` checkpoint, since the concept-erasure baselines we compare against operate on this backbone. Initializing every defense from the same checkpoint isolates the immunization mechanism from differences in the underlying generator. We benchmark against three representative methods that span the design space of diffusion concept erasure: gradient-based fine-tuning (ESD), closed-form parameter editing (UCE), and bi-level immunization (IMMA).

SD v1.4 (Rombach et al., 2022). Stable Diffusion v1.4 acts as the undefended reference. We report its target-concept metrics to establish the upper bound an adversary aims to reach and its safe-concept metrics to define the fidelity ceiling any defense should approach. The checkpoint also serves as the initialization θ_0 for every defense, including GIFT.

ESD (Gandikota et al., 2023). ESD fine-tunes the UNet so the conditional score on a target prompt matches the unconditional score under classifier-free guidance, which suppresses the concept while leaving the rest of the model approximately unchanged. The method exposes two configurations: ESD-x trains only the cross-attention projections (intended for concepts tied to a specific token, such as artist names or objects), while ESD-u trains the non-cross-attention parameters of the UNet (intended for broader concepts such as NSFW content). Following the recommendations in the original paper, we use ESD-x for both the artistic styles and object classes. We run the official implementation with default hyperparameters.

UCE (Gandikota et al., 2024). UCE applies a closed-form edit to the cross-attention key and value projection matrices, redirecting target prompt embeddings toward neutral anchor embeddings without any gradient-based fine-tuning. The closed-form solution comes from a least-squares formulation that simultaneously minimizes the projection error on a set of preserved concepts and the alignment error on the target concepts. Because the edit is non-iterative, UCE runs in seconds and avoids the optimization instability of gradient-based methods. We run the official implementation with default hyperparameters.

IMMA (Zheng & Yeh, 2024). IMMA learns a tamper-resistant initialization through bi-level optimization: the inner loop simulates an adversary fine-tuning the model on the target concept, and the outer loop updates the defended weights to maximize the post-adaptation loss. The procedure produces a checkpoint that, by construction, places the target concept in a region of parameter space from which standard fine-tuning recovers it slowly. IMMA represents the closest prior art to GIFT, since both methods explicitly target the relearning trajectory rather than the loss at $\theta^\star$ alone. IMMA provides separate scripts per adaptations method (*e.g.*, LoRA, DreamBooth, Textual Inversion). In our experiments, we run the official LoRA implementation with default hyperparameters.

We obtain each defended checkpoint by running the official implementation with default hyperparameters, which guarantees that any reported gap reflects the methods themselves rather than tuning differences.

B.2 Datasets and Concepts

Artistic styles. We curate 14 artists spanning different styles: Andy Warhol, Claude Monet, Frida Kahlo, Gustav Klimt, Hokusai, Jackson Pollock, Leonardo da Vinci, Michelangelo, Pablo Picasso, Piet Mondrian, Rembrandt, Salvador Dalí, Vincent van Gogh, and Yayoi Kusama. For each artist we use the baseline model SDv1.4 to generate 200 images from the undefended Stable Diffusion v1.4 model with the prompt "`A high quality painting in ⟨artist_name⟩ style`". We split the 200 images equally into a 100-image immunization set $\mathcal{D}_t$ and a disjoint 100-image attack set $\mathcal{D}_{\text{adv}}$, ensuring no overlap between defender and adversary data. To probe each defended checkpoint, we generate 100 images using the prompt "`a painting of a cat in ⟨artist⟩ style`" before and after the attack and compute pre- and post-attack metrics on these generations.

Object categories. We select 10 ImageNet classes following Zheng & Yeh (2024); Gandikota et al. (2023): cassette player, english springer, french horn, garbage truck, gas pump, military uniform, plane, tench, toaster, and violin. For each class we generate 200 images using the template "`A high-quality image of a ⟨object⟩`". As with the art styles, we split each set into a 100-image $\mathcal{D}_t$ and a 100-image $\mathcal{D}_{\text{adv}}$. We probe each defended checkpoint with 100 generations using the prompt "`a photo of a ⟨object⟩`" for pre- and post-attack evaluation.

Safe prior data. For the preservation set $\mathcal{D}_s$, we adopt the 139 prompt-image pairs released by Li et al. (2025). The set covers three benign categories: 30 images of random objects with simple captions (e.g., "`photo of banana`"), 50 images of benign human activities (e.g., "`Young woman on a scooter in the city`"), and 59 images of human portraits (e.g., "`a photography of a man with a white shirt and tie smiling`"). We use this set as $\mathcal{D}_s$ across all experiments.

Safe-generation evaluation. To measure fidelity on broad, unrelated prompts, we adopt the 10,000 text-image subset of MS-COCO (Lin et al., 2014) released by Zhang et al. (2024a). We generate one image per prompt with each defended checkpoint and compute aggregated metrics on the resulting set.

B.3 Metrics

CLIP. We use CLIP ViT-B/32 (Radford et al., 2021) for two complementary measurements. CLIP-T scores the cosine similarity between a generated image and its prompt embedding, capturing prompt-image alignment. CLIP-I scores the cosine similarity between a generated image and the

corresponding training images in $\mathcal{D}_{\text{adv}}$, capturing how closely the generation matches the target concept distribution.

Aesthetic score. We use the LAION aesthetic predictor (Schuhmann et al., 2022) to obtain a human-aligned visual quality estimate (AS). On target prompts, a low AS combined with a low CLIP-T flags structurally degraded outputs that no longer resemble the concept; on safe prompts, AS tracks generation quality on benign content.

LPIPS. LPIPS (Zhang et al., 2018) measures perceptual distance between a generation and the reference set $\mathcal{D}_{\text{adv}}$ in the feature space of a pretrained network. Higher LPIPS values reflect generations that drift further from the target distribution and indicate stronger erasure.

Attack Success Rate (objects only). For object concepts, we measure top-1 and top-5 Attack Success Rate (ASR) using a ResNet-50 classifier (He et al., 2016) pretrained on ImageNet. For each defended checkpoint we generate 100 images per class and pass them through the classifier; ASR-1 and ASR-5 report the fraction of generations whose ground-truth class falls in the top-1 and top-5 predictions, respectively. We aggregate ASR across the 10 classes for each method.

B.4 Fine-Tuning Attack Recipes

We test tamper resistance using three fine-tuning recipes that span the practical attack surface for an adversary holding the defended weights.

LoRA fine-tuning. LoRA (Hu et al., 2022) injects low-rank trainable matrices $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$ into linear layers, so each layer's effective weight becomes $W + BA$ with the original W frozen. Only A and B receive gradient updates, which dramatically reduces memory footprint and trainable parameter count. We choose LoRA as the default attack because it captures the realistic threat model: an adversary fine-tuning on consumer hardware. Full-parameter fine-tuning, by contrast, requires substantially more memory and tends to overfit small attack sets quickly. We run the LoRA attack using the official Diffusers training script.[1]

DreamBooth. DreamBooth (Ruiz et al., 2023) personalizes a diffusion model on a small image set by fine-tuning the full UNet under a class-specific prior preservation loss, binding the target concept to a unique identifier token. Adversaries adopt DreamBooth when they want strong concept reproduction from a handful of reference images. We run DreamBooth using the Diffusers reference implementation.[2]

Full-parameter fine-tuning. Full-parameter fine-tuning updates every weight in the UNet on $\mathcal{D}_{\text{adv}}$ with the standard denoising objective. The recipe demands more compute than LoRA and proves harder to tune without overfitting, but it represents the strongest attack a well-resourced adversary can mount. We run it with the Diffusers text-to-image script.[3]

B.5 Implementation Details

Immunization training. We initialize GIFT from Stable Diffusion v1.4 and run algorithm 1 for 600 steps with AdamW (Loshchilov & Hutter, 2019), a learning rate of 10^{-4}, batch size 1, and a safe-weight $\lambda_s = 2.5$. We update only the projection matrices of attention modules (both self- and cross-attention). Table 6 sweeps the learning rate over $\{10^{-3}, 5\times10^{-4}, 10^{-4}, 5\times10^{-5}\}$, and table 5 ablates the trainable-parameter set across cross-attention only, self-attention only, both attention types, and the full UNet.

[1] https://github.com/huggingface/diffusers/blob/main/examples/text_to_image/train_text_to_image_lora.py

[2] https://github.com/huggingface/diffusers/blob/main/examples/dreambooth/train_dreambooth.py

[3] https://github.com/huggingface/diffusers/blob/main/examples/text_to_image/train_text_to_image.py

Attack hyperparameters. Unless otherwise specified, we attack each defended checkpoint with LoRA at rank 4, learning rate 10^{-4}, batch size 1, and 5000 update steps under a cosine learning-rate schedule. To verify that resistance does not hinge on a specific attack budget, we additionally sweep the learning rate over $\{5\times10^{-4},\ 10^{-4},\ 5\times10^{-5}\}$, the LoRA rank over $\{4, 8, 16\}$, and the step count over $\{5\text{k}, 10\text{k}, 25\text{k}\}$ in section 4.4.

Hardware. All immunization runs and fine-tuning attacks execute on a single NVIDIA L40S GPU with 48 GB of memory.

C Additional Results

C.1 Results on SDXL

The experiments in the main paper use Stable Diffusion v1.4. To test whether GIFT depends on that particular backbone, we repeat the experiments on SDXL base 1.0 at 1024×1024 resolution, using the same 14 artistic styles and 10 object classes, the same defender and attacker splits, and the same 5000-step LoRA fine-tuning attack. All other hyperparameters are unchanged.

Table 8: **GIFT on SDXL.** Averages over 14 artistic styles and 10 object classes.

		Before FT		After FT			Safe Concepts	
	Method	CLIP-T ↓	AS ↓	CLIP-T ↓	AS ↓	ASR-1 ↓	CLIP-T ↑	AS ↑
Styles	SDXL (no erasure)	0.288	5.86	–	–	–	0.262	5.41
	GIFT (ours)	0.147	4.22	0.199	4.20	–	0.248	5.06
Objects	SD-XL (no erasure)	0.264	5.36	–	–	0.87	0.262	5.41
	GIFT (ours)	0.162	4.27	0.185	4.13	0.01	0.243	5.17

Table 8 shows that the behaviour observed on SD1.4 carries over to SDXL. Before fine-tuning, GIFT suppress the target concept: CLIP-T falls from 0.288 to 0.147 on artistic styles and from 0.264 to 0.162 on objects. After a 5000-step LoRA attack, the GIFT keeps CLIP-T at 0.199 on styles and 0.185 on objects, and top-1 accuracy recovers only to 0.01. It is also worth noting the utility degradation on the safe concepts, which is a limitation of GIFT. We note that resistance on SDXL is at least as strong as on SDv1.4 under a matched attack budget, which indicates that GIFT does not rely on properties specific to the smaller backbone.

Figure 4 shows representative generations. The top row is the undefended SDXL model, the middle row is the same prompts after immunization, and the bottom row is after the fine-tuning attack. We can observe that immunization persists after attack.

Figure 5 shows that GIFT maintains safe utility on the SDXL family after immunization. The top row is the undefended SDXL model and the bottom row is the immunized SDXL checkpoint using GIFT. Prompts were sampled from the COCO dataset.

C.2 Additional Baselines

In this section we compare our immunization framework against two other recent works on the same topic: ResAlign (Li et al., 2026) and Meta-Unlearn (Gao et al., 2025).

ResAlign. (Li et al., 2026) addresses the issue that safety-driven unlearning methods largely fail to retain their effectiveness once the model is fine-tuned downstream, even on entirely benign data. It treats the downstream fine-tuning process as an implicit optimization problem and applies a Moreau envelope-based reformulation, which allows the gradient of the post-fine-tuning behavior to be estimated without solving the inner problem explicitly. The unlearning objective is then optimized against that estimate, so the erasure is placed where a subsequent fine-tune is least able to undo it. A meta-learning strategy simulates a diverse distribution of fine-tuning scenarios so that the resulting robustness generalizes beyond any single downstream setting. We use the official implementation and follow the released training script.

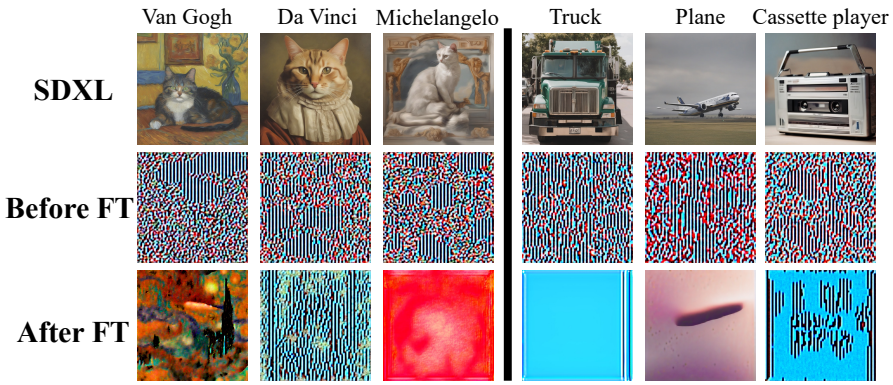


Figure 4: **GIFT erases concepts and resists recovery under fine-tuning attacks on SDXL.** Top: the undefended SDXL baseline generates target concepts across three artistic styles and three object classes. Middle: after immunization with GIFT, the same prompts collapse into structureless noise. Bottom: even after a 5000-step LoRA fine-tuning attack, recovered generations remain heavily distorted and far from the original concepts. The robustness generalizes across both concept families and to the SDXL model family.

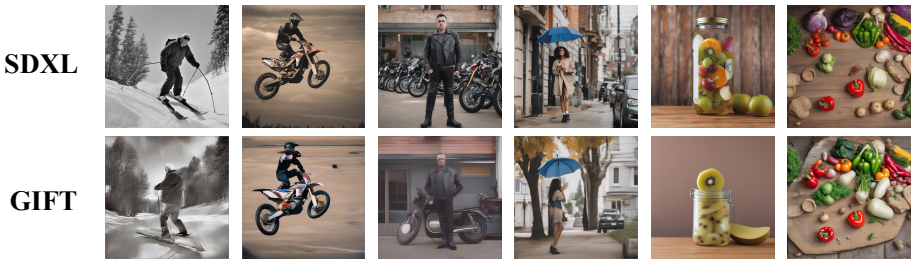


Figure 5: **SDXL Safe Generation.** An immunized SDXL checkpoint is still able to generate safe concepts properly. Prompts chosen randomly from COCO dataset.

Meta-Unlearn (Gao et al., 2025) targets a complementary failure mode: after a concept is unlearned, benign concepts related to it remain in the model, and a malicious party can exploit those remaining concepts to relearn the erased one through targeted fine-tuning. The method adds a meta objective on top of an existing unlearning procedure so that the unlearned model behaves normally when used as is, but the retained related concepts are triggered to self-destruct if the model is subsequently fine-tuned on the unlearned concept. The objective is therefore not to make erasure harder to reverse, but to make reversing it destroy the capability the attacker needs. We use the official implementation and run the ESD-u-1 variant defined in the paper, that is, unlearning of all UNet parameters except cross-attention at an erase scale of $\eta = 1$, for 1000 steps at batch size 10 and a learning rate of 10^{-5}, with the released γ values.

Table 9 and Figure 6 show that neither baseline achieves both objectives at once, while GIFT does. ResAlign (Li et al., 2026) suppresses the concept effectively before the fine-tuning attack and leaves safe generation intact, but it does not survive the attack: after fine-tuning its CLIP-T returns to 0.250 against 0.256 for the base SD1.4 model, and the recovered samples in Figure 6 are clean, recognizable cassette players. Meta-Unlearn (Gao et al., 2025) resists the attack in the sense that its post-attack CLIP-T of 0.150 is no higher than its pre-attack value, but it does so by sacrificing the model itself: its safe CLIP-T falls to 0.121 and its aesthetic score to 3.91, against 0.260 and 5.28 for the base SD1.4 model, and the corresponding column in Figure 6 shows that safe prompts also produce structureless output. Its post-attack robustness is therefore not protection of the concept

Table 9: **GIFT vs. Additional Baselines.** ResAlign and Meta-Unlearn each sacrifice one of the two objectives: robustness to fine-tuning or preservation of safe generation quality. GIFT is the only method that suppresses the protected concept before and after fine-tuning while preserving safe generation. Both baselines are far more expensive per concept. Erase time is the wall-clock cost of immunizing a single concept on one NVIDIA L40S; the attack and evaluation stages are identical across methods and are excluded.

Method	Before fine-tuning		After fine-tuning		Safe (COCO)		Time ↓
	CLIP- ↓T	AS ↓	CLIP-T ↓	AS ↓	CLIP-T ↑	AS ↑	
SD1.4	0.256	5.27	–	–	0.260	5.28	–
Meta-Unlearning (Gao et al., 2025)	0.165	3.87	0.150	4.45	0.121	3.91	8.2 h
ResAlign (Li et al., 2026)	0.170	4.03	0.250	5.51	0.263	5.26	1.9 h
GIFT	0.170	4.12	0.229	4.36	0.253	5.03	6 min

but an absence of any usable generation to recover. It is worth noting that safe generations always happen before attempting to attack the immunized checkpoint, so Meta-Unlearn's argument that the model will self-destruct when an attack is attempted is not valid in this case. GIFT is the only method that suppresses the concept before fine-tuning, keeps it suppressed afterwards, and preserves safe generation throughout, as the safe column in Figure 6 shows. It also does so at a small fraction of the cost, requiring 6 minutes per concept against 1.9 hours for ResAlign (Li et al., 2026) and 8.2 hours for Meta-Unlearn (Gao et al., 2025) on identical hardware.

D Limitations and Ethical Considerations

GIFT has three limitations worth noting. First, our evaluation covers artistic styles and ImageNet object classes; extending to harder concept families such as NSFW content or specific identities remains future work. Second, immunization requires access to the target concept and a curated safe set, so defenders must anticipate which concepts to remove before release. Third, our resistance guarantees hold across the LoRA, DreamBooth, and full-parameter attacks we test, but a determined adversary with a larger compute budget or a novel adaptation recipe could still recover the concept. On the ethical side, immunization is a defender-side tool, but the choice of which concepts to remove rests with the model provider and carries downstream consequences for artists, communities, and downstream users. It is a tool that could be used to enforce sensorship. Tamper-resistance should complement rather than replace transparent governance over what gets erased and why.

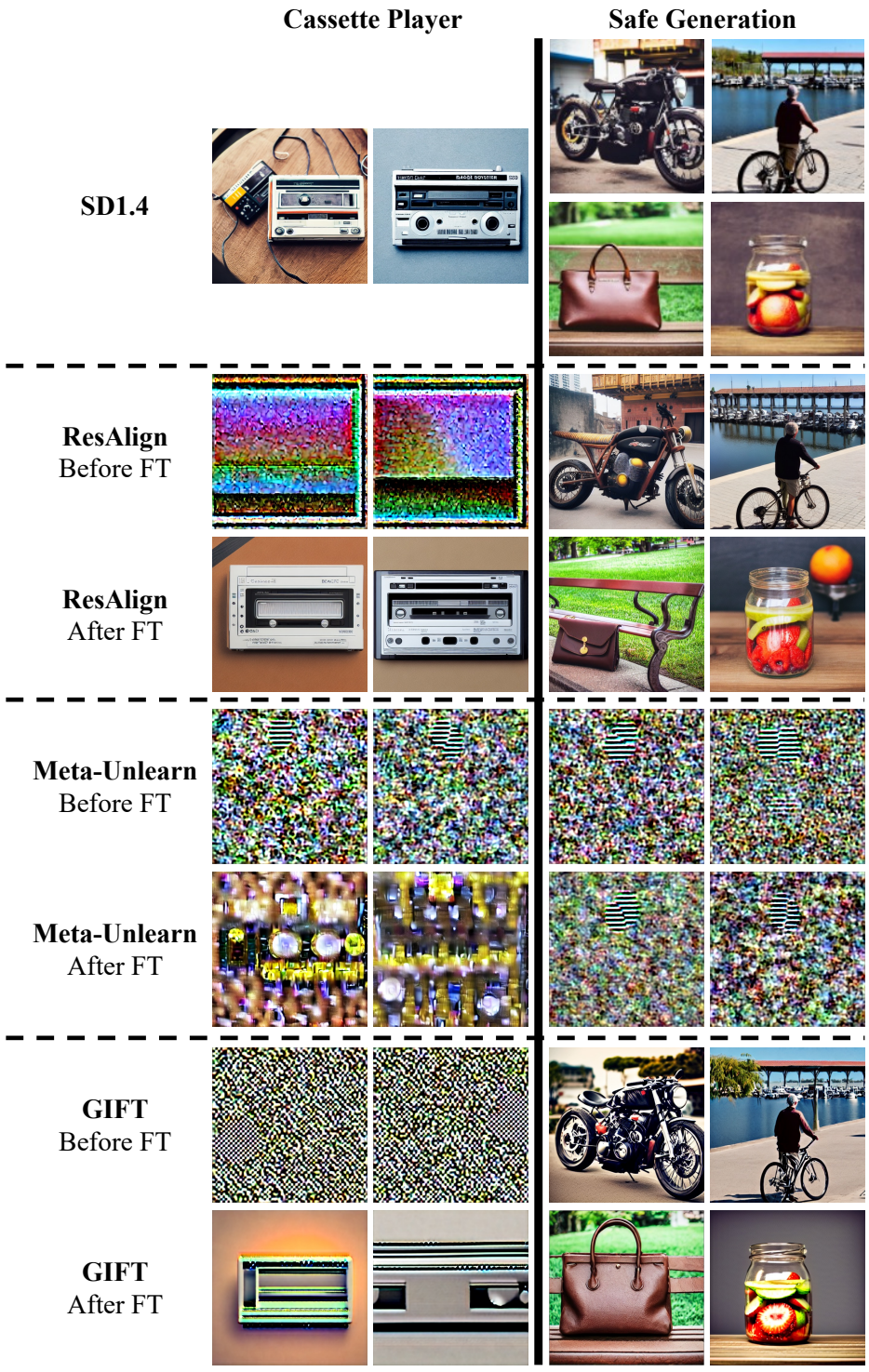


Figure 6: **GIFT vs. Additional Baselines on Cassette Player.** The left side shows the model's outputs for the protected concept cassette player, before and after the fine-tuning attack. The right side shows generations of safe concepts from the corresponding immunized checkpoints. GIFT preserves the protected concept after the attack while maintaining generation of safe concepts. In contrast, ResAlign and Meta-Unlearn either fail to preserve the protected concept after fine-tuning or substantially degrade the model's generation quality. GIFT maintains both robustness to fine-tuning and safe generation quality.