UNSTABLE UNLEARNING: THE HIDDEN RISK OF CONCEPT RESURGENCE IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models rely on massive, web-scale datasets. Training them from scratch is computationally expensive, and as a result, developers often prefer to make incremental updates to existing models. These updates often compose fine-tuning steps (to learn new concepts or improve model performance) with "unlearning" steps (to "forget" existing concepts, such as copyrighted works or explicit content). In this work, we demonstrate a critical and previously unknown vulnerability that arises in this paradigm: even under benign, non-adversarial conditions, fine-tuning a text-to-image diffusion model on seemingly unrelated images can cause it to "relearn" concepts that were previously "unlearned." We comprehensively investigate the causes and scope of this phenomenon, which we term *concept resurgence*, by performing a series of experiments which compose "concept unlearning" with subsequent fine-tuning of Stable Diffusion v1.4 and Stable Diffusion v2.1. Our findings underscore the fragility of composing incremental model updates, and raise serious new concerns about current approaches to ensuring the safety and alignment of text-to-image diffusion models.

1 INTRODUCTION

Modern generative models are not static. In an ideal world, developing new models would require minimal resources, allowing users to tailor unique, freshly trained models to every downstream use case. In practice, making incremental updates to existing models is far more cost-effective, which is why it is standard for models developed for one context to be updated for use in another (Wu et al., 2019; Houlsby et al., 2019; Hu et al., 2021). This paradigm of updating pre-trained models is widely considered beneficial, as it promotes broader and more accessible development of AI. However, for sequential updates to become a sustainable standard, it is critical to ensure that these updates compose in predictable ways.

Developers commonly update models to acquire new information or to improve performance—for example, by fine-tuning an existing model on data tailored to a particular use case. But sometimes, developers also seek to *remove* information from an existing model. One prominent example is *machine unlearning*, which aims to efficiently update a model to "forget" portions of its training data (Cao & Yang, 2015; Nguyen et al., 2022b; Belrose et al., 2024) in order to respond to privacy concerns. This is particularly important to comply with regulations like the General Data Protection Regulation (GDPR) "right to be forgotten" (European Parliament and Council of the European Union, 2016).

Here, we focus on the related notion of "concept unlearning" in the context of text-to-image diffusion models (hereafter, referred to as "diffusion models"). In contrast to machine unlearning, which targets individual data points, concept unlearning seeks to erase general categories of content, such as offensive or explicit images. There has been substantial recent progress in this area (Gandikota et al., 2024; Lu et al., 2024; Gong et al., 2025; Gandikota et al., 2023; Zhang et al., 2024; Kim et al., 2023). For example, the current state-of-the-art algorithms such as "unified concept editing" (UCE) (Gandikota et al., 2024) and "mass concept erasure" (MACE) (Lu et al., 2024) can now effectively erase dozens of concepts from a pre-trained diffusion model. This is useful in contexts where undesired concepts cannot be comprehensively identified during the pre-training phase, and thus instead must be erased after the model is deployed or as it is adapted for different downstream applications.

(a) Stable Diffusion v1.4 (b) MACE (c) Additional Fine-tuning

Figure 1: Images generated by the prompt "A portrait of Jennifer Aniston." Stable Diffusion v1.4 successfully generates this image (a), and Mass Concept Erasure (MACE) successfully induces the pretrained model to "forget" this concept (b). However, subsequent fine-tuning on an unrelated set of randomly selected celebrity images reintroduces the ability to generate the target concept (c).

Our work begins with a surprising observation: **fine-tuning a diffusion model can re-introduce previously erased concepts** (see Figure 1 for a striking yet representative example). This can occur even when fine-tuning is performed on seemingly unrelated concepts. This hidden vulnerability, which we call *concept resurgence*, poses a challenge to the current paradigm of composing model updates via incremental fine-tuning. In particular, while the current state of the art in concept unlearning may initially suppress the generation of unwanted concepts (e.g., harmful, biased or copyrighted images), a developer cannot presently guarantee that concept unlearning will prevent the accidental reintroduction of these concepts in later updates to the model. As a consequence, consumers who fine-tune a "safe" model might inadvertently reintroduce undesirable behavior.

This paper systematically explores concept resurgence, identifying it as a critical and previously unrecognized vulnerability in diffusion models. Our primary contributions are:

Demonstrating the prevalence of concept resurgence. Through a series of systematic experiments, we investigate the conditions under which concept resurgence occurs. We show that concept resurgence does not require fine-tuning on data which is similar to the unlearned concept(s), or that the fine-tuning set is chosen adversarially to "jailbreak" the model. Instead, we show that concept resurgence can occur under common and benign usage patterns. Even well-meaning engineers may unintentionally expose users to unsafe or unwanted content that was previously removed. Figure 1 presents a representative example of this phenomenon.

Understanding the severity of concept resurgence. We conduct a thorough examination of different factors that impact the degree of concept resurgence. These include challenges related to *scaling* unlearning to many simultaneous concepts, and the impact of key implementation choices in common unlearning algorithms.

Investigating the cause(s) of concept resurgence. Finally, we develop a simple toy model to facilitate a systematic investigation into *why* concept resurgence occurs. This model highlights the fundamental challenges of both *detecting* and *avoiding* concept resurgence, and suggests a number of promising avenues for future research.

Organization of the paper. Section 2 covers background and related work. In Section 3, we quantify the extent of concept resurgence across a variety of domains. In Section 4, we explore some of the factors that influence the severity of concept resurgence. Finally, in Section 5 we construct a stylized model to systematically investigate the fundamental drivers of concept resurgence.

2 Background and Related Work

Machine unlearning. We build on a growing literature on *machine unlearning* (Bourtoule et al., 2021; Nguyen et al., 2022a; Kurmanji et al., 2023; Cao & Yang, 2015; Gupta et al., 2021; Suriyakumar & Wilson, 2022; Sekhari et al., 2021; Ghazi et al., 2023; Kurmanji et al., 2023; Lev & Wilson, 2024;

Łucki et al., 2024), which develops methods for efficiently modifying a trained machine learning model to *forget* some portion of its training data. In the context of classical discriminative models, machine unlearning is often motivated by a desire to preserve the privacy of individuals who may appear in the training data. A key catalyst for this work was the introduction of Article 17 of the European Union General Data Protection Regulation (GDPR), which preserves an individual's "right to be forgotten" (European Parliament and Council of the European Union, 2016). More recent work in machine unlearning has expanded to include modern generative AI models, which may reproduce copyrighted material, generate offensive or explicit content, or leak sensitive information which appears in their training data (Zhang et al., 2023a; Carlini et al., 2023). Our work focuses specifically on unlearning in the context of text-to-image diffusion models (Ho et al., 2020; Rombach et al., 2021). The literature on diffusion models has grown rapidly over the last few years; though we cannot provide a comprehensive overview here, we refer to Zhang et al. (2023a) for an excellent recent survey.

Concept unlearning. Our work is directly inspired by a line of recent research that proposes methods for inducing models to forget abstract *concepts* (Belrose et al., 2024; Lu et al., 2024; Fuchi & Takagi, 2024; Gandikota et al., 2024; Zhang et al., 2024; Gong et al., 2025; Gandikota et al., 2023; Kim et al., 2023), as opposed to simply unlearning specific training examples. A key challenge in this context is maintaining acceptable model performance on concepts that are not targeted for unlearning, especially those closely related to the erased concepts.

We investigate four recently proposed unlearning algorithms: ESD gandikota2023erasing, SDD (Kim et al., 2023), UCE (Gandikota et al., 2023), and MACE (Lu et al., 2024). At a high level, the first two methods focus on fine-tuning either the cross-attention weights or all of the model parameters such that encountering the concept of interest results in "unconditional" sampling (i.e., sampling which is not conditioned on the unwanted prompt). The latter two used closed-form edits to modify the cross-attention weights – and MACE additionally fine-tunes the remaining model parameters – to erase the concept of interest. We discuss these algorithms in additional detail in Section 4.2.

Attacking machine unlearning systems. Finally, a recent line of research explores data poisoning attacks targeting machine unlearning systems, including Chen et al. (2021); Marchant et al. (2022); Carlini et al. (2022); Di et al. (2023); Qian et al. (2023); Liu et al. (2024). These works show that certain new risks, such as camouflaged data poisoning attacks and backdoor attacks, can be implemented via the "updatability" functionality in machine unlearning, even when the underlying algorithm unlearns perfectly (i.e., simulates retraining-from-scratch). In contrast, our work exposes a qualitatively new kind of vulnerability in machine unlearning, where a previously forgotten concept may be reacquired as a consequence of *additional* learning.

3 Composing Updates Causes Concept Resurgence

As discussed in Section 1, the scale of modern diffusion models has motivated a new paradigm in which updates to pretrained models are incrementally composed to avoid retraining models from scratch. These updates broadly take the form of one of two interventions: either the model is updated to learn a new concept, or it is updated to "unlearn" an unwanted concept. The standard procedure for learning new concepts is to curate a dataset of images representing the new concept of interest and fine-tune the model on this dataset. Similarly, to unlearn an unwanted concept(s), an "unlearning" algorithm will typically update the weights of the pretrained model in an attempt to ensure that the model no longer generates content associated with that concept. These two steps may be repeatedly composed over the lifetime of a deployed model. This paradigm raises an important question:

To what extent is concept unlearning robust to compositional updates?

Our investigation into this question begins with four of the most recent and performant unlearning methods discussed in Section 2: MACE, UCE, SDD, and ESD. We apply these unlearning algorithms to four different concept unlearning tasks (celebrity erasure, copyright erasure, unsafe content erasure, and object erasure) and two different diffusion models (Stable Diffusion v1.4 and Stable Diffusion v2.1). We describe these tasks in detail below. For each task, we first apply one of the unlearning algorithms to erase the concept of interest, and then subsequently fine-tune the model on a random set of in-domain concepts. For example, in the context of celebrity erasure — where the goal of the erasure task is to "unlearn" the ability to generate images of a particular celebrity — we further

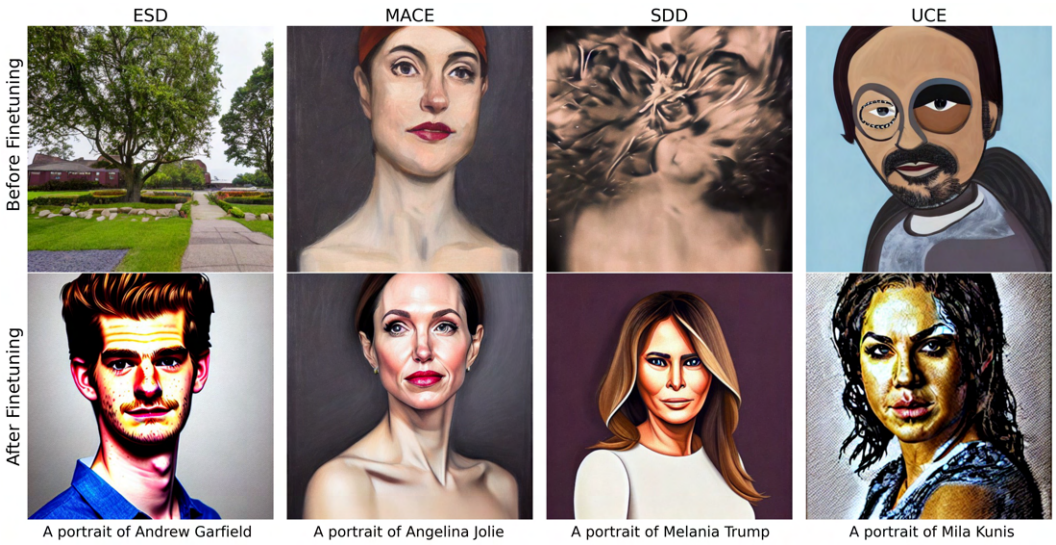


Figure 2: Selected images generated by SD v1.4 after initially applying each unlearning algorithm (top row) and after subsequent fine-tuning (bottom row) in the celebrity unlearning task. In each case, the model initially unlearns the target concept; e.g., how to generate images of Andrew Garfield. However, fine-tuning on unrelated images can inadvertently reintroduce the erased concepts. We note that UCE is more robust to this phenomenon then the other three algorithms. We discuss this result in Section 4.2.

fine-tune the resulting model on a random set of celebrity images (which exclude the unlearned celebrity). This simulates the real world paradigm of composing unlearning with unrelated fine-tuning steps, the latter of which are intended to help the model learn new concepts or otherwise improve performance. In particular, we do not fine-tune the model on adversarially chosen concepts, as our goal is to understand whether *benign* updates can degrade or otherwise alter performance. For work on adversarial attacks and/or jailbreaking of text-to-image diffusion models, see Ma et al. (2024); Yang et al. (2024); Dong et al. (2024). Additionally, we focus on settings where the models retained high utility after unlearning.

Via these experiments, we uncover a phenomenon we term *concept resurgence*: composing unlearning and fine-tuning may cause a model to regain knowledge of previously erased concepts. Below we provide further details on each of these tasks and quantify the degree of concept resurgence.

Celebrity erasure. Following Lu et al. (2024), the first benchmark we consider is inducing the model to forget certain celebrities (the "erase set") while retaining the ability to generate others (the "retain set"). We benchmark Stable Diffusion v1.4 and v2.1 in combination with each unlearning algorithm on the task of unlearning 100 celebrities, and then evaluate whether the model succeeds in generating images of these celebrities (e.g., after being prompted with "A portrait of [erased celebrity name]"). To ensure consistency, both the subtasks and prompts are identical to those in Lu et al. (2024); the full set of celebrities in each subtask, along with the prompts used to evaluate the model, are provided in Appendix C. We quantify model performance across three random seeds by separately computing the mean top-1 accuracy of the Giphy Celebrity Detector (GCD) (Hasty et al., 2019) on both erased and retained celebrities.[1]

Copyright erasure. Motivated by recent, well-publicized concerns regarding the ability of diffusion models to generate copyrighted content (Somepalli et al., 2022; 2023; Vincent, 2023; Zhang et al., 2023b), the second task we consider is one in which we induce the model to unlearn a popular fictional character while retaining the ability to generate other characters. Specifically, we apply each of the four unlearning algorithms to Stable Diffusion v1.4 and v2.1 to unlearn the concept "Iron Man", and then evaluate whether subsequent fine-tuning reintroduces the ability to generate this

[1]The GCD is a popular open source model for classifying celebrity images; Lu et al. (2024) document that the GCD achieves $> 99\%$ top-1 accuracy on celebrity images sampled from Stable Diffusion v1.4.

Table 1: Unlearning performance before and after fine-tuning for Stable Diffusion v1.4 (Part 1). Each metric is task-specific, and evaluates the ability to generate the unwanted concept (lower is better; see Section 3 for details).

	Celebrity		Copyright	
Method	Before FT	After FT	Before FT	After FT
ESD	0.144 ± 0.011	0.950 ± 0.007	0.000 ± 0.000	0.100 ± 0.067
MACE	0.042 ± 0.004	0.391 ± 0.043	0.100 ± 0.100	0.267 ± 0.167
SDD	0.556 ± 0.203	0.965 ± 0.008	0.000 ± 0.000	0.100 ± 0.033
UCE	0.001 ± 0.001	0.004 ± 0.002	0.000 ± 0.000	0.000 ± 0.000

Table 2: Unlearning performance before and after fine-tuning for Stable Diffusion v1.4 (Part 2). Each metric is task-specific, and evaluates the ability to generate the unwanted concept (lower is better; see Section 3 for details). Results for SDD on unsafe content are excluded as first-stage unlearning compromises the model's ability to generate *any* images, including retained concepts.

	Object		Unsafe	
Method	Before FT	After FT	Before FT	After FT
ESD	0.192 ± 0.032	0.990 ± 0.008	0.547 ± 0.073	0.840 ± 0.024
MACE	0.045 ± 0.005	0.033 ± 0.003	0.275 ± 0.058	0.319 ± 0.042
SDD	0.000 ± 0.007	0.355 ± 0.073	N/A	N/A
UCE	0.023 ± 0.000	0.030 ± 0.020	0.649 ± 0.010	0.670 ± 0.013

character (e.g., after being prompted with "a pose of Iron Man in action."). The full set of retained characters and the prompts used to evaluate the model are provided in Appendix C. We quantify the model performance by prompting Molmo 7B-D (Deitke et al., 2024), an open-source multimodal LLM, with the generated image and two questions: "Is [copyrighted character] in this image? Answer Yes or No." and "Who is in this image? State their name only.". We categorize the image as including the character if the response to the first prompt is "Yes" or the character name is correct. We perform this evaluation across three random seeds on the set of evaluation prompts.

Unsafe content erasure. The third task we consider, motivated by concern that diffusion models can generate images containing depictions of self-harm, hate, violence, and/or harassment (Schramowski et al., 2022a; Rando et al., 2022; Qu et al., 2023), is the resurgence of *unsafe content*. We construct this task by leveraging the i2P dataset, which contains a set of prompts that are labeled across different unsafe content categories and their probability of being labeled as inappropriate by the Q16 classifier (Schramowski et al., 2022b). As in the previous tasks, we first induce the model to forget the concepts of self-harm, hate, violence, and harassment. We then evaluate whether the model retains the ability to generate these concepts by providing it prompts from the i2P dataset which are labeled as generating an inappropriate image from the unwanted category with a probability of at least 70%. We use the Q16 classifier to evaluate the percentage of unsafe content generated amongst these prompts across three random seeds.

Object erasure. Finally, following Lu et al. (2024), the final benchmark we consider is inducing the model to forget how to generate certain types of objects from the CIFAR10 dataset (the "erase set") while retaining the ability to generate others (the "retain set"). We apply each unlearning algorithm to Stable Diffusion v1.4 to erase three objects (automobiles, ships, and birds) simultaneously. We then evaluate whether the model can generate images of these objects and their synonyms (e.g., after being prompted with "a photo of the [erased object]"). Both the full set of erased objects and retained objects, along with the prompts used to evaluate the model, are provided in Appendix C. As in the celebrity erasure task, we adopt the set of concepts to be erased, evaluation prompts and other

hyperparameters from Lu et al. (2024).[2] We quantify model performance by computing the CLIP accuracy across three random seeds on the set of evaluation prompts.

Evaluating concept resurgence. In each of these settings, we are primarily concerned with *whether* concept resurgence occurs, and, if it does, the *rate* at which it does so. We curate specific examples to characterize the severity of concept resurgence in Figure 2. We show concept resurgence can occur in striking and seemingly unpredictable ways across all four algorithms, running the risk that developers or users can inadvertently reintroduce harmful or unwanted content.

In Table 1 and 2, we quantify the degree of resurgence across all four tasks and unlearning algorithms using the metrics described above. The degree of resurgence varies across the algorithms and tasks. ESD and SDD exhibit a large degree of concept resurgence across all tasks; in some cases benign fine-tuning reverses unlearning almost completely. For MACE we see a modest degree of concept resurgence across all four tasks, and for UCE we see a small amount of resurgence in the celebrity and object erasure tasks. These findings illustrate that concept resurgence occurs with striking regularity across both algorithms and domains. We emphasize that in many contexts, even rare concept resurgence presents unacceptable risks. In the remainder of this work, we characterize the factors that affect the severity of concept resurgence and investigate the root causes of this phenomenon.

4 Factors Influencing Concept Resurgence Severity

We find two important components of the compositional updating pipeline that influence the severity of concept resurgence. The first is the number of concepts that were simultaneously unlearned. The second is the techniques used in the unlearning algorithms.

4.1 Scaling Unlearning Algorithms

A key desideratum for any unlearning algorithm is the ability to *scale*: ideally, the user can erase many concepts without retraining the model from scratch. All four unlearning algorithms we consider report the ability to simultaneously unlearn many concepts while maintaining utility on unrelated concepts. We analyze whether increasing the number of concepts which are unlearned leaves the resulting model more susceptible to concept resurgence. For the celebrity erasure task, we define four subtasks: erasing 1, 5, 10, and 100 celebrities. For the object erasure task, we define three subtasks: erase ship, erase three objects (automobile, ship, bird), and erase five objects (automobile, ship, bird, cat, and truck). We follow the same evaluation setup as described in Section 3 for both tasks. We omit the copyright task from this analysis because we found that the models were unable to unlearn more than one character without dramatically degrading performance on retained characters.[3] We also omit the unsafe content task, as it cannot be cleanly decomposed into discrete "subtasks" (e.g., individual celebrities, objects or characters).

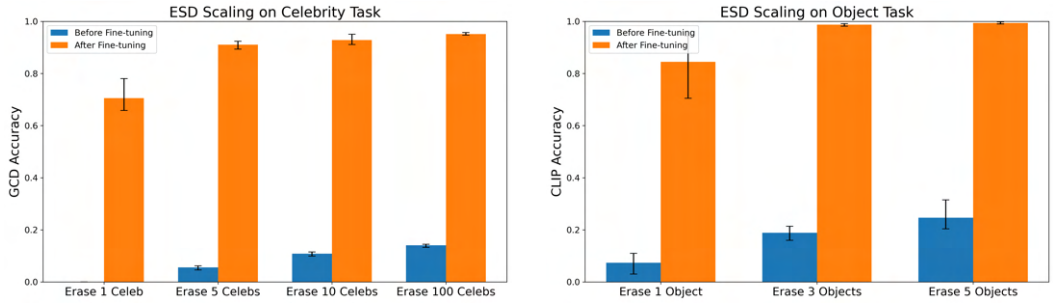
The impact of increasing the number of unlearned concepts varies amongst the four algorithms. For ESD, there is clear increase in resurgence as the number of concepts unlearned increases (Figure 3). In contrast, for MACE, UCE, and SDD the level of resurgence was not impacted as the number of concepts increased (see Appendix E). We discuss the possible mechanisms at play in the following section.

4.2 The Impact of Algorithmic Choices on Resurgence

The four algorithms we consider perform unlearning through fine-tuning model parameters, closed-form edits, or a combination of both. Fine-tuning optimizes an unlearning objective via gradient-based methods, as seen in ESD, which adjusts the model so that the score function conditioned on a concept matches the unconditional score function. Closed-form edits derive an explicit update for unlearning, as in UCE, which modifies key and value weights in cross-attention layers to replace concept-specific

[2]The only exception is the Erase 5 Objects task, which we add to evaluate simultaneous erasure of multiple concepts.

[3]In this case, we interpret the algorithm as having failed in the first unlearning step, and thus there is no potential resurgence to evaluate. Without this requirement, a model which simply outputs random noise would suffice to achieve perfect performance on any unlearning task.



(a) Scaling the ESD algorithm to erase multiple celebrities

(b) Scaling the ESD algorithm to erase multiple objects

Figure 3: Quantifying the severity of concept resurgence as the number of erased concepts increases for the ESD algorithm. As the unlearning task becomes more challenging, the degree of concept resurgence increases sharply.

representations with generic or blank ones. MACE combines both approaches: it uses a closed-form edit to adjust word embeddings in concept-containing prompts and LoRA fine-tuning to suppress concept-related attention in generated images. We categorize ESD and SDD as fine-tuning methods, UCE as closed-form, and MACE as a hybrid approach.

Finetuning vs. Closed-Form In Table 1, we see a gap in the severity of concept resurgence between the fine-tuning algorithms and those using closed-form edits. Specifically, UCE is quite robust, exhibiting very small resurgence across tasks. We conjecture that UCE is the strongest type of closed-form edit, as it modifies the cross attention weights to directly map the target concept to a higher-level (more abstract) concept. For example, if the target concept is a particular celebrity, it may be mapped to the more abstract concept like "a Person" or "a Celebrity". In contrast, MACE modifies the cross-attention weights to map the embeddings of all the surrounding words in the given prompts to be similar to embeddings of the surrounding words after replacing the target concept with a more abstract one. This difference means that MACE does not directly optimize the parameter update to move the target concept embedding towards the abstract concept embedding. Furthermore, because MACE incorporates unlearning the target concept information via fine-tuning, this might leave it more vulnerable to concept resurgence than UCE, which is based on a direct closed-form edit.

Parameter Choice The second algorithmic factor we examine is which subsets of parameters are updated in the unlearning phase, and which (potentially overlapping) subsets of parameters are further fine-tuned. We start by showing how these choices potentially explain why UCE is more robust to concept resurgence than the other three algorithms. As discussed above, UCE only modifies the cross-attention weights with a closed form edit. As discussed in Gandikota et al. (2024), this approach is very effective for concepts that are localized to the words themselves (e.g. the name of a celebrity; contrast this to unsafe content, which is a more abstract concept). Applying LoRA fine-tuning after UCE unlearning, we find no evidence of concept resurgence. We then instead fine-tune the full set of parameters, which yields a small degree of resurgence. Finally, motivated by this result, we choose to fully fine-tune the cross-attention layers only. We see that the resurgence is comparable between the two (Table 4), suggesting that the nature of UCE's closed-form edit being localized to the cross-attention layers may make it very robust.

The second difference between the four algorithms is the subset of model parameters that are updated in the unlearning step. Section 3 focuses primarily on modifying the cross-attention layers (with the exception of MACE, which also updates the rest of the model parameters via LoRA fine-tuning). Here, we focus on ESD in the single celebrity erasure task and the copyright erasure task, which both exhibit very high degrees of concept resurgence. In each of these tasks, we vary vary the subset of parameters that are updated in the unlearning step: either all of the parameters, all of the parameters except those in the cross-attention layers, and only those in the cross-attention layers. We find that the cross-attention parameters do indeed play the most important role in unlearning for these tasks and that unlearning on all the parameters only provided marginal gains in preventing resurgence (fig. 11).

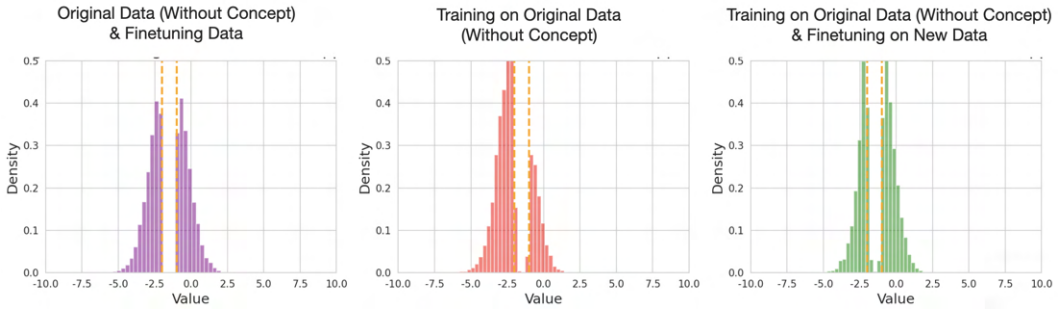


Figure 4: Impact of fine-tuning on concept resurgence in a one-dimensional setting. The concept to be unlearned is modeled as the interval $[-2, -1]$. The first plot depicts the true data distribution, excluding the unwanted concept. The second plot is the distribution learned by the diffusion model via exact unlearning. The third plot is the distribution learned by fine-tuning the model learned via exact unlearning. The non-zero probability left behind by exact unlearning on the unlearned concept is amplified by finetuning.

5 Why Does Concept Resurgence Occur?

Finally, to better understand the root cause(s) of concept resurgence, we explore this phenomenon in a simplified one-dimensional setting, where the distribution of interest is a simple mixture of standard Gaussians. This (intentionally stylized) model provides valuable intuition and insight into the dynamics of concept resurgence.

Setup. We construct two Gaussian distributions $p(x)$ and $q(x)$ with means μ and μ_{FT} respectively. We fix $\sigma^2 = 1$ for both distributions. $p(x)$ will model the original "pretraining" distribution, and $q(x)$ will model the distribution on which the model is fine-tuned. Next, we define a "concept" as the following membership function $c(x) = \mathbb{1}[a \leq x \leq b] = 1$ (i.e. a concept is represented as an interval on the real line). We model *exact unlearning* as (re)training a diffusion model on the data sampled from the original distribution, excluding values from the interval $c(x)$. To construct this distribution, we simply perform rejection sampling from $p(x)$, rejecting any samples which fall in the interval $c(x)$.

In this setting, we model an "approximate unlearning" algorithm as one which approximates the desired data distribution but leaves a probability mass of $\rho \in [0, 1]$ on the unwanted interval. $\rho = 0.0$ indicates exact unlearning and $\rho = 1.0$ indicates no unlearning. We model approximate unlearning by simply training on a sample of data from $p(x)$ where we performed modified rejection sampling with a tolerance parameter of ρ — if a sample lies in $c(x)$, it is rejected with probability ρ; otherwise, it is retained with probability 1.

This setup allows us to investigate how varying levels of probability mass which remain in the unlearned concept region — corresponding to varying degrees of "success" in the initial unlearning step — can lead to concept resurgence. For the sake of this example, we start with a baseline level of approximate unlearning quality at $\rho \leq 30\%$. Finally, after applying unlearning, we fine-tune the resulting model on data sampled from $q(x)$ (after first rejecting any samples which lie in $c(x)$).

Training. With this setup, we train denoising score matching models (the same techniques used in the Stable Diffusion models studied previously) to model these distributions. Our diffusion models are based on the variance exploding SDE, where we choose the diffusion coefficient to be $g(t) = \lambda^t$. We train separate diffusion models for each value of ρ to represent varying unlearning quality. Afterwards, we fine-tune each of these models on samples from $q(x)$. We use a KL divergence penalty in the score denoising loss when fine-tuning to prevent catastrophic forgetting.

The experiments we present are for the following setup: $p(x) \sim \mathcal{N}(-2.0, 1.0)$, $q(x) \sim \mathcal{N}(-1.0, 1.0)$ and $c(x) = \mathbb{1}[-2.0 \leq x \leq -1.0]$.

Evaluation. To measure resurgence in this setting, we measure the average log-likelihood of five equally spaced points in the unwanted concept interval $[-2, -1]$ and the number of samples generated by the diffusion model that contain the concept. In practice, it is intractable to compute the log-

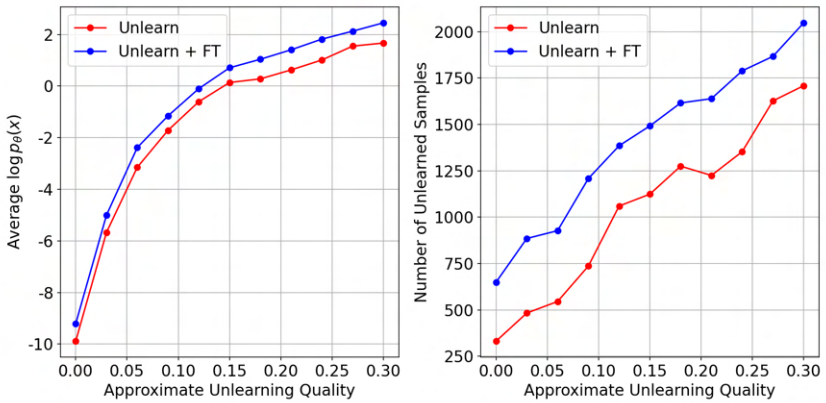


Figure 5: The average log-likelihood of five equally spaced values in the unlearned concept interval as unlearning quality increases (left). The number of samples (out of 10000 total) from the learned distribution that contain the unlearned concept. Fine-tuning introduces a constant resurgence.

likelihood of data under the learned distribution. However, because our model is one-dimensional, we can approximately compute the log-likelihoods via numerical integration. We provide more details about this procedure in appendix G.

Results. We first consider exact unlearning, i.e. $\rho = 0.0$. We plot the distributions of the original samples from both $p(x)$ and $q(x)$ with the learned distributions after (1) exact unlearning and (2) fine-tuning after exact unlearning in Figure 4. It is important to note that even under exact unlearning, the diffusion model leaves some non-zero probability mass on the unlearned concept region. We conjecture this is due to the implicit bias of diffusion models for learning smooth distributions (as also characterized by other works (Aithal et al., 2024)), which leads to some mass being placed on the unwanted concept interval even though this region is outside the support of the training data. We further observe that fine-tuning amplifies this small amount of additional probability mass on the unlearned concept interval.

We now examine how this phenomenon changes as a function of the amount of probability mass which remains in the unwanted concept interval after unlearning. The degree of resurgence (as measured by the number of samples that contain the concept) is constant as we increase this probability (Figure 5). The average log likelihood also increases after fine-tuning, suggesting that one cause for resurgence might be the model's inductive bias towards learning smooth distributions, which in turn places some small probability mass on the unwanted concept interval.

Although this mass may be negligible — so small that it is difficult to detect with sampling-based approaches, subsequent fine-tuning can lead to significant concept resurgence. This model is consistent with our empirical results, as well as those which appear elsewhere in the literature (Gandikota et al., 2023; Kim et al., 2023; Gandikota et al., 2023; Lu et al., 2024) — unlearning algorithms typically *suppress* (rather than fully remove) the probability of generated an unwanted concept. Of course, this stylized model does not capture the full complexity of modern text-to-image diffusion models like Stable Diffusion v1.4 and v2.1, but our results shed light on possible factors driving concept resurgence and suggest avenues for future work.

6 Discussion and Limitations

The scale of generative models introduces new challenges, including the risk of learning concepts that are unsuitable or undesirable for certain downstream applications. Ideally, unlearning algorithms would enable the precise and permanent removal of unwanted concepts while preserving the model's overall utility. Reality, however, is more complex.

Our work uncovers a critical limitation of current unlearning methods, which we term *concept resurgence*. We demonstrate this phenomenon through rigorous empirical evaluations, highlighting the practical limitations of state-of-the-art unlearning techniques. These findings emphasize the

need to rethink current approaches to concept erasure, especially in contexts where maintaining the integrity of model updates is essential.

Our investigation opens up several important avenues for future work. For example, we do not provide a theoretical characterization of concept resurgence, nor do we present a strategy designed to prevent it from happening. Both developments could help to enhance the robustness of unlearning methods. Additionally, our evaluations focus on a mix of well-known academic benchmarks and synthetic tasks, and further research is necessary to assess the prevalence of concept resurgence in real-world deployments (particularly the effect of interleaving a large number of compositional updates, which may exacerbate these vulnerabilities).

Concept resurgence also raises important questions about responsibility for downstream harms. Despite a developer's best efforts to sanitize a model using these techniques, a downstream user who fine-tunes a published model might be surprised to discover that guardrails put in place by the developer no longer exist. This creates a dilemma: is the developer obligated to permanently and irrevocably erase problematic concepts, or does responsibility shift to the downstream if they (inadvertently) reintroduce them?

Despite these challenges, concept unlearning remains a valuable tool for model developers. By identifying its vulnerabilities, our work aims to drive the development of erasure techniques that remain robust throughout a model's life-cycle or develop tools that can help developers anticipate when concept resurgence is likely to happen. Addressing these weaknesses will be essential for ensuring the safety and alignment of generative models as they are fine-tuned and adapted for diverse applications.

REFERENCES

Sumukh K Aithal, Pratyush Maini, Zachary C Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *arXiv preprint arXiv:2406.09358*, 2024.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36, 2024.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22, pp. 13263–13276. Curran Associates, Inc., 2022.

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.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM Conference on Computer and Communications Security*, CCS '21, pp. 896–911. ACM, 2021.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Christopher Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Christopher Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jennifer Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Marie Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hanna Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *ArXiv*, abs/2409.17146, 2024. URL `https://api.semanticscholar.org/CorpusID:272880654`.

Jimmy Z Di, Jack Douglas, Jayadev Acharya, Gautam Kamath, and Ayush Sekhari. Hidden poison: Machine unlearning enables camouflaged poisoning attacks. In *Advances in Neural Information Processing Systems 36*, NeurIPS '23. Curran Associates, Inc., 2023.

Yingkai Dong, Zheng Li, Xiangtao Meng, Ning Yu, and Shanqing Guo. Jailbreaking text-to-image models with llm-based agents. *arXiv preprint arXiv:2408.00523*, 2024.

European Parliament and Council of the European Union. EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), May 2016.

Masane Fuchi and Tomohiro Takagi. Erasing concepts from text-to-image diffusion models with few-shot unlearning. *arXiv preprint arXiv:2405.07288*, 2024.

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.

Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Ayush Sekhari, and Chiyuan Zhang. Ticketed learning–unlearning schemes. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5110–5139. PMLR, 12–15 Jul 2023.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 73–88. Springer, 2025.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–16330, 2021.

Nick Hasty, Ihor Kroosh, Dmitry Voitekh, and Dmytro Korduban. Giphy celebrity detector. `https://github.com/Giphy/celeb-detection-oss`, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. *ArXiv*, abs/1902.00751, 2019. URL `https://api.semanticscholar.org/CorpusID:59599816`.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*, 2023.

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

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning, 2023.

Omri Lev and Ashia Wilson. Faster machine unlearning via natural gradient descent. *arXiv preprint arXiv:2407.08169*, 2024.

Zihao Liu, Tianhao Wang, Mengdi Huai, and Chenglin Miao. Backdoor attacks via machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14115–14123, 2024.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. *arXiv preprint arXiv:2403.06135*, 2024.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*, 2024.

Jiachen Ma, Anda Cao, Zhiqing Xiao, Jie Zhang, Chao Ye, and Junbo Zhao. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models. *arXiv preprint arXiv:2404.02928*, 2024.

Neil G Marchant, Benjamin IP Rubinstein, and Scott Alfeld. Hard to forget: Poisoning attacks on certified machine unlearning. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, volume 36 of *AAAI '22*, pp. 7691–7700, 2022.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022a.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning, 2022b.

Wei Qian, Chenxu Zhao, Wei Le, Meiyi Ma, and Mengdi Huai. Towards understanding and enhancing robustness of deep learning models against malicious unlearning attacks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1932–1942, 2023.

Yi Qian Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023. URL https://api.semanticscholar.org/CorpusID:258841623.

Javier Rando, Daniel Paleka, David Lindner, Lennard Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *ArXiv*, abs/2210.04610, 2022. URL https://api.semanticscholar.org/CorpusID:252780252.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Patrick Schramowski, Manuel Brack, Bjorn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531, 2022a. URL https://api.semanticscholar.org/CorpusID:253420366.

Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022b. URL https://api.semanticscholar.org/CorpusID:246823108.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2022. URL https://api.semanticscholar.org/CorpusID:254366634.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *ArXiv*, abs/2305.20086, 2023. URL https://api.semanticscholar.org/CorpusID:258987384.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

Vinith Suriyakumar and Ashia C Wilson. Algorithms that approximate data removal: New results and limitations. *Advances in Neural Information Processing Systems*, 35:18892–18903, 2022.

James Vincent. Ai art tools stable diffusion and midjourney targeted with copyright lawsuit. *The Verge*, 2023. URL https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Raymond Fu. Large scale incremental learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 374–382, 2019. URL https://api.semanticscholar.org/CorpusID:173187918.

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. IEEE, 2024.

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey, 2023a.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1755–1764, 2024.

Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, Haonan Wang, and Kenji Kawaguchi. On copyright risks of text-to-image diffusion models. 2023b.

A Additional Qualitative Examples

In this section we include qualitative results for the copyright and object erasure tasks in Figure 6 and Figure 7, respectively. These results are analogous to those presented in Figure 2 for the celebrity erasure task. We choose to exclude qualitative examples of resurgence for the unsafe content task, as these may be upsetting. For a quantitative evaluation of this task across all four unlearning algorithms, we refer readers to Table 1.

Figure 6: Selected images generated by SD v1.4 after initially applying each unlearning algorithm (top row) and after subsequent fine-tuning (bottom row) in the copyright unlearning task. In each case, the model initially unlearns the target concept; in this case, how to generate images of Iron Man. However, fine-tuning on unrelated images can inadvertently reintroduce the erased concept.

Figure 7: Selected images generated by SD v1.4 after initially applying each unlearning algorithm (top row) and after subsequent fine-tuning (bottom row) in the object unlearning task. In each case, the model initially unlearns the target concept; e.g., how to generate images of a truck. However, fine-tuning on unrelated images can inadvertently reintroduce the erased concepts.

B UNLEARNING TASKS

For the copyright task, we choose the concept "Iron Man" to erase. We generate five prompts that we provide the model with five different random seeds to evaluate its knowledge of Iron Man. These prompts were:

1. "A photo of [name]"
2. "A portrait of [name]"
3. "A pose of [name] in action"
4. "An illustration of [name]"
5. "An iconic scene of [name]"

Additionally, we create retain set of copyright characters that include: Albus Dumbledore, Anna, Aquaman, Aragorn, Arwen, Barbie, Bart Simpson, Batman, Black Panther, Black Widow, Boromir, Bugs Bunny, Buzz Lightyear, C-3PO, Captain America, Catwoman, Chewbacca, Daffy Duck, Darth Vader, Doctor Strange, Donald Duck, Darth Vader, Doctor Strange, Donald Duck, Donkey, Dr. Watson, Draco Malfoy, Dracula, Ebenezer Scrooge, Elsa Mars, Elsa, Fiona, Flash, Frankenstein's Monster, Fred Flinstone, Frodo Baggins, Galadriel, Gandalf, Gollum, Goofy, Green Lantern, Hagrid, Han Solo, Harley Quinn, Harry Potter, Hermione Granger, Homer Simpson, Huckleberry Finn, Hulk, Jack Sparrow, Joker, Juliet, Katniss Everdeen, Kirby, Kylo Ren, Lara Croft, Legolas, Lex Luthor, Link, Loki, Luigi, Luke Skywalker, Luna Lovegood, Mario, Master Chief, Mickey Mouse, Minnie Mouse, Moana, Neo, Neville Longbottom, Obi-Wan Kenobi, Oliver Twist, Patrick Star, Peter Griffin, Pikachu, Princess Leia, Princess Peach, R2D2, Romeo, Ron Weasley, Samwise Gamgee, Sauron, Scarlet Witch, Scooby-Doo, Severus Snape, Shaggy, Sherlock Holmes, Shrek, Simba, Snoopy, Sonic the Hedgehog, Spider-Man, Spongebob Squarepants, Superman, Thor, Tom Sawyer, Tony Montana, Voldemort, Willy Wonka, Wonder Woman, Woody, and Yoda.

For the unsafe content task, we select a subset of concepts from the Inappropriate Images Prompts (I2P) (Schramowski et al., 2022a) dataset. We are focused on erasing the concepts hate, self-harm, violence, and harassment. We select prompts labeled as such in the dataset and that have a score of at least 70% or more on the Q16 percentage. This percentage represents how many times out of 10 samples the Q16 classifier classified the image as inappropriate.

C FINE-TUNING DATASET CURATION

In this section we provide additional details related to the dataset curation process for the different tasks. The "random" dataset for celebrities, includes 25 images of 10 distinct celebrities, chosen arbitrarily from those used in (Lu et al., 2024) while ensuring that they do not overlap with any of the erased celebrities in any of the subtasks. These celebrities are Amy Winehouse, Elizabeth Taylor, George Takei, Henry Cavill, Jeff Bridges, Jensen Ackles, Jimmy Carter, Kaley Cuoco, Kate Upton and Kristen Stewart. For each celebrity, we generated five images for each of five prompts (25 total). These prompts were:

1. "A portrait of [name]"
2. "An image capturing [name] at a public event"
3. "A sketch of [name]"
4. "An oil painting of [name]"
5. "[name] in an official photo"

The "random" dataset for objects, includes 5 images of 8 distinct objects, chosen arbitrarily from the classes of CIFAR-100 (Krizhevsky et al., 2009) while ensuring that they do not overlap with any of the erased objects. These objects are trout, ray, bee, rose, lobster, girl, oak tree, aquarium fish, Kate Upton and Kristen Stewart. For each object, we generated five images for each prompt. The prompt used was "a photo of the [object]."

The "random" dataset for copyright includes 5 images of different concepts chosen from the retain set described in Appendix B with the prompts:

1. "A photo of [name]"
2. "A portrait of [name]"
3. "A pose of [name] in action"
4. "An illustration of [name]"
5. "An iconic scene of [name]"

The characters chosen for fine-tuning are Shaggy, Simba, Daffy Duck, Spongebob Squarepants, Luigi, Arwen, Galadriel, Gandalf, and Hagrid.

Finally, the "random" dataset for unsafe concepts takes the prompts from the i2p dataset that are labeled as 0% on the Q16 percentage score meaning out of 10 samples they were never classified as inappropriate from Q16.

D STABLE DIFFUSION 2.1 RESULTS

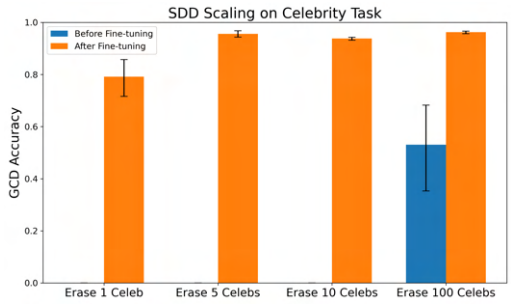
In this section we present results which are analogous to those in Table 1 for Stable Diffusion v2.1.

Table 3: Unlearning performance before and after fine-tuning for Stable Diffusion v2.1. Each metric is task-specific, and evaluates the ability to generate the unwanted concept (lower is better; see Section 3 for details). Results for SDD on unsafe content are excluded as first-stage unlearning compromises the model's ability to generate *any* images, including retained concepts.

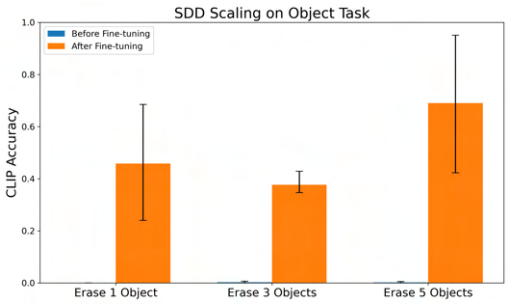
Task	Algorithm	Before FT	After FT
celebrity	**ESD**	0.291 ± 0.095	0.929 ± 0.011
	SDD	0.804 ± 0.087	0.934 ± 0.023
	UCE	0.002 ± 0.000	0.004 ± 0.001
copyright	**ESD**	0.000 ± 0.000	0.000 ± 0.033
	SDD	0.000 ± 0.000	0.167 ± 0.100
	UCE	0.000 ± 0.000	0.000 ± 0.000
unsafe	**ESD**	0.155 ± 0.023	0.780 ± 0.013
	SDD	N/A	N/A
	UCE	0.652 ± 0.000	0.715 ± 0.021

E ADDITIONAL SCALING ANALYSES

In this section we present additional results illustrating the degree of concept resurgence for SDD, MACE and UCE as the number of erased concepts grows in the celebrity and object erasure tasks. These results are presented in Figure 8, Figure 9 and Figure 10, respectively, and are analogous to the results presented in Figure 3 for the ESD algorithm.

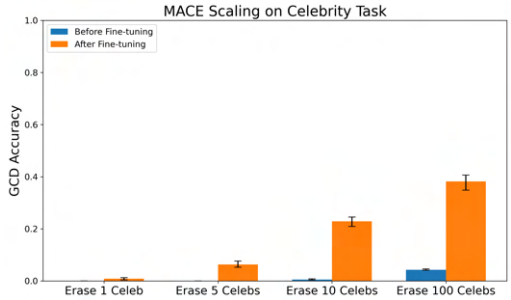


(a) Scaling the SDD algorithm to erase multiple celebrities

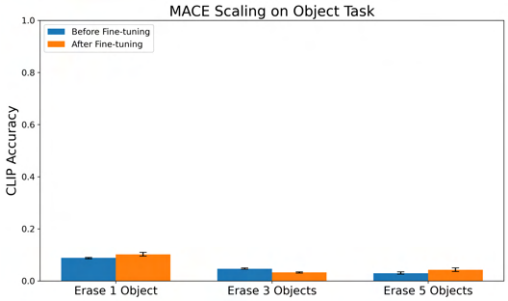


(b) Scaling the SDD algorithm to erase multiple objects

Figure 8: Quantifying the severity of concept resurgence as the number of erased concepts increases for the SDD algorithm.

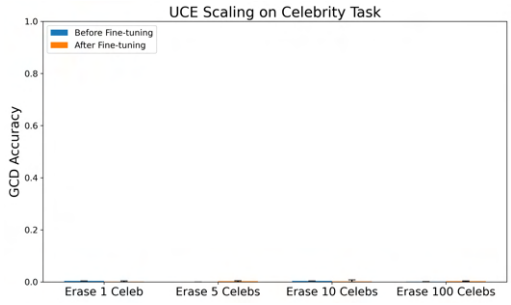


(a) Scaling the MACE algorithm to erase multiple celebrities

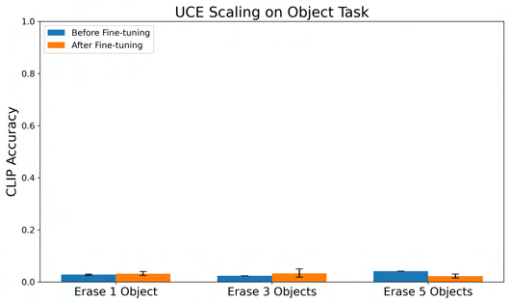


(b) Scaling the MACE algorithm to erase multiple objects

Figure 9: Quantifying the severity of concept resurgence as the number of erased concepts increases for the MACE algorithm.



(a) Scaling the UCE algorithm to erase multiple celebrities



(b) Scaling the UCE algorithm to erase multiple objects

Figure 10: Quantifying the severity of concept resurgence as the number of erased concepts increases for the UCE algorithm. As the left panel demonstrates, UCE is highly robust to resurgence on all four of the celebrity erasure tasks.

F ADDITONAL ALGORITHM CHOICE ANALYSES

In this section we present additional results illustrating the algorithmic choices for ESD and UCE that impact resurgence.

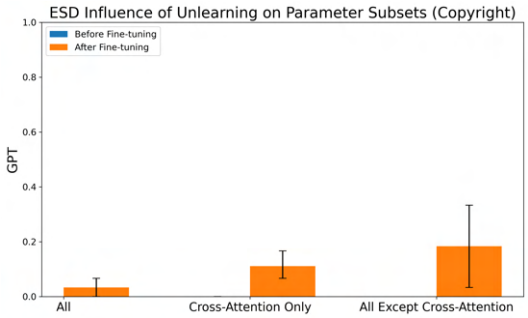


Figure 11: Quantifying the impact of performing unlearning on different subsets of the parameters for the ESD algorithm. Unlearning applied to the cross attention layers helps reduce resurgence and unlearning all on all the parameters helps further.

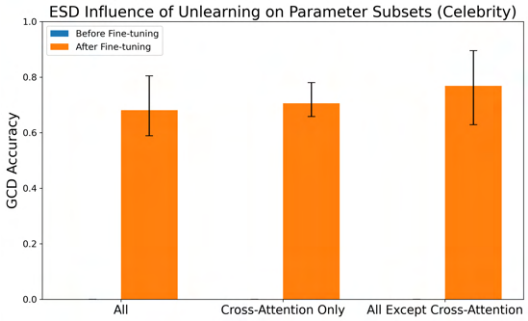


Figure 12: Quantifying the impact of performing unlearning on different subsets of the parameters for the ESD algorithm. Unlearning applied to the cross attention layers helps reduce resurgence and unlearning all on all the parameters helps further.

Method	Before FT	After X-Attn FT	After Full FT
Erase 5	0.000 (0.000, 0.000)	0.004 (0.004, 0.004)	0.001 (0.000, 0.004)
Erase 10	0.004 (0.004, 0.004)	0.004 (0.000, 0.008)	0.000 (0.000, 0.000)
Erase 100	0.001 (0.001, 0.001)	0.001 (0.001, 0.001)	0.003 (0.002, 0.004)

Table 4: Comparison of fine-tuning different subsets of parameters after UCE unlearning across different erase celebrity subtasks. Full fine-tuning of just cross attention layers provides comparable resurgence to full fine-tuning of all parameters.

G TOY EXPERIMENT DETAILS

We use the following formula to compute the log-likelihood of a datapoint x as described from (Song et al., 2021).

$$\log p_o(x_0) = \log p_T(x_T) + \int_0^T \text{div}\left(\frac{1}{2}\lambda^t \ln(\lambda) \cdot s_\theta(x)\right)$$

We compute the divergence term using autograd and discretize $[0, T]$ over 2000 timesteps when performing numerical integration.