Unstable Unlearning: The Hidden Risk of Concept Resurgence in Diffusion Models

Vinith M. Suriyakumar*,† Rohan Alur*,† Ayush Sekhari†

Manish Raghavan† Ashia C. Wilson†

†Massachusetts Institute of Technology

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 "mass concept erasure" (the current state of the art for unlearning in text-to-image diffusion models [31]) with subsequent fine-tuning of Stable Diffusion v1.4. 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 common for models developed for one context to be updated for use in another [43, 19, 20]. 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 a novel dataset tailored to a particular use case. But developers sometimes seek to *remove* information from an existing model. One prominent example is *machine unlearning*, which aims to efficiently update a trained model to "forget" portions of its training data [3, 36, 1], in order to respond to privacy concerns. This is especially important for compliance with regulations like the General Data Protection Regulation (GDPR) "right to be forgotten" [11].

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,

*Equal Contribution; Corresponding authors: {vinithms@mit.edu, ralur@mit.edu}

which targets particular 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. For example, the current state-of-the-art in "mass concept erasure" (MACE) can now effectively erase dozens of concepts from a pre-trained diffusion model [31]. 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.

Our work begins with a surprising observation: **fine-tuning a diffusion model can re-introduce previously erased concepts** and this can occur even when fine-tuning is performed on seemingly unrelated concepts (see Figure 1 for a striking yet representative example). 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 content.

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

- **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.

- **Understanding the causes of concept resurgence.** We conduct a thorough examination of both the unlearning algorithms and the fine-tuning data involved in concept resurgence. We quantify the extent of the problem across a range of standard benchmarks, finding that the degree of concept resurgence is closely related to the choice of mapping concept (i.e., the more generic / unrelated the concept, the less resurgence) and degree of regularization imposed during unlearning.

Organization of the paper. Section 2 is dedicated to background and related work. In Section 3, we quantify the extent of concept resurgence across a variety of standard benchmarks. We investigate the influence of unlearning and fine-tuning algorithms in Section 4, followed by an analysis of the role of fine-tuning data in Section 5. Finally, Section 6 discusses limitations and outlines directions for future research.

2 Background and Related Work

Machine unlearning. We build on a growing literature on *machine unlearning* [2, 35, 25, 3, 16, 41, 39, 14, 25, 27, 32], 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

(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).

forgotten" [11]. 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 [45, 5]. Our work focuses specifically on unlearning in the context of diffusion models [18, 38]. The literature on diffusion models has grown rapidly over the last few years; though we cannot provide a comprehensive overview here, we refer to [45] 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* [1, 31, 12, 13], 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. At the time of this work, Lu et al. [31] (MACE) is the state of the art in terms of both erasure performance and image generation quality after unlearning.

MACE: Mass Concept Erasure in diffusion models. We build directly on the recent work of Lu et al. [31] for Mass Concept Erasure (MACE) in diffusion models. Broadly speaking, MACE fine-tunes a model to erase certain target phrase (e.g., "an image of a ship") and their related concepts (e.g., "an image of a boat") by using a combination of cross-attention refinement and low rank adaptation (LoRA) [20]. Cross-attention refinement modifies the "key" embeddings associated with each token co-existing in the target phrase with the corresponding "key" embedding of co-existing words in a more generic phrase. The second step, LoRA fine-tuning, perturbs the weights of the model to minimize activations in regions which correspond to the target phrase; these regions are identified by segmenting the image using Grounded-SAM [23, 29]. These perturbations are learned via low rank adapatation (LoRA) of the model parameters [20]. Finally, the LoRA modules corresponding to each erased concept are combined to produce a final model by formulating the "fusion" of multiple LoRA modules as a quadratic programming problem. For additional detail on MACE we refer the reader to [31].

Attacking machine unlearning systems. Finally, a recent line of research explores data poisoning attacks targeting machine unlearning systems, including Chen et al. [7], Marchant et al.

Figure 2: Representative images generated by SD v1.4 in each celebrity unlearning task. We first apply MACE to erase a set of 1,5,10 or 100 celebrities from the pretrained model; e.g., to "forget" how to generate images of Angelina Jolie (top row). We then fine-tune the resulting model on images of unrelated celebrities (bottom row). While the model initially "unlearns" how to generate each erased celebrity, subsequent fine-tuning reintroduces some of the unlearned concepts.

[34], Carlini et al. [4], Di et al. [8], Qian et al. [37], Liu et al. [30]. 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 erasure robust to compositional updates?

We begin our investigation with Stable Diffusion v1.4, and separately apply MACE to perform a celebrity erasure and object erasure task. We describe these tasks in detail below. For each task,

we fine-tune the model on a random set of in-domain concepts after MACE has been applied. 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 fine-tune the resulting model on a random set of celebrity images (which exclude the unlearned celebrity). This is intended to simulate 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 [33, 44, 9].

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 [31], the first benchmark we consider is inducing the model to forget how to generate certain celebrities (the "erase set") while retaining the ability to generate others (the "retain set"). We benchmark Stable Diffusion v1.4 on four subtasks in which we apply MACE to unlearn 1, 5, 10 or 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 [31]; the full set of celebrities in each subtask, along with the prompts used to evaluate the model, are provided in Appendix C.1. We quantify model performance across three random seeds by separately computing the mean top-1 accuracy of the Giphy Celebrity Detector (GCD) [17] on both erased and retained celebrities.[1] Additionally, for one random seed, we quantify the model's ability to continue to generate general concepts using the CLIP Score (i.e. the cosine similarity between the prompts and generated images) computed on a random sample of 5000 captions from MSCOCO [28] (COCO-5K) and the FID between the generated images from pretrained Stable Diffusion and those from our experiments for COCO-5K.

Object erasure. Following [31], the second 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 MACE to Stable Diffusion v1.4 across four subtasks. In the first three, we apply MACE to erase knowledge of a single object (automobiles, ships and bird, respectively). In the fourth, we apply MACE to a more challenging task in which the goal is to *simultaneously* unlearn a set of five concepts (automobiles, ships, birds, cats, and trucks). In each subtask, 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.1. As in the celebrity erasure task, we adopt the set of concepts to be erased, evaluation prompts and other hyperparameters from [31].[2] We quantify model performance by computing the CLIP accuracy across three random seeds on the set of evaluation prompts. Following [31], we do not compute FID-5K and CLIP-5K for the object erasure task; COCO-5K is itself composed of common objects, and the goal of this evaluation is to assess performance on generic concepts unrelated to the erasure task.

[1] The GCD is a popular open source model for classifying celebrity images; [31] document that the GCD achieves $> 99\%$ top-1 accuracy on celebrity images sampled from Stable Diffusion v1.4.

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

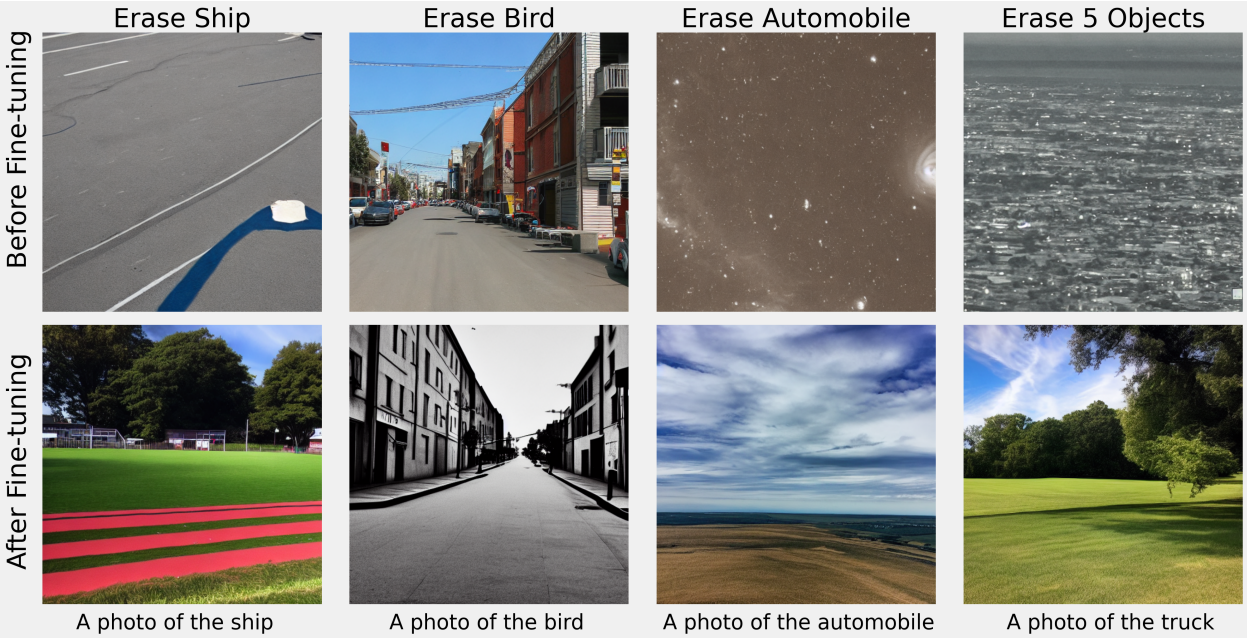

Figure 3: Representative images generated by SD v1.4 in each object unlearning task before (top row) and after (bottom row) subsequent fine-tuning on images of unrelated objects. In the "Erase 5 Objects" task, the model simultaneously unlearns five concepts (one of which is the ability to generate images of trucks). Unlike the celebrity erasure task (Figure 2), none of the unlearned concepts reappear in this set of representative images. However, as Figure 4 demonstrates, the vulnerability persists on certain concepts and prompts.

As discussed above, object erasure and celebrity erasure are two of the four tasks considered in [31]. We discuss our choice to exclude the other two (artistic style and explicit content erasure) in Appendix A.

Evaluating concept resurgence. We present representative examples to characterize the degree of "typical" concept resurgence in Figure 2 and Figure 3, and curate specific examples of this vulnerability in Figure 4.[3]

As Figure 2 demonstrates, concept resurgence can occur *in degrees*, as some concepts are not reintroduced at all (e.g., Melania Trump), and others are only partially reintroduced (e.g., Barack Obama). Furthermore, Figure 3 demonstrates that concept resurgence can be rare in some contexts; indeed, none of the representative images we sample reintroduce the unlearned concepts in the object erasure task. However, as both Figure 2 and Figure 4 demonstrate, concept resurgence can occur in striking and seemingly unpredictable ways, running the risk that developers or users can inadvertently reintroduce harmful or unwanted content.

In Figure 5, we quantify the degree of resurgence across both the object and celebrity erasure tasks using the metrics described above. As suggested by the qualitative results, the degree of resurgence is substantially larger in the celebrity erasure task, particularly as the number of unlearned celebrities grows large (intuitively, such tasks are "harder" than unlearning a smaller number of celebrities, as the model must simultaneously unlearn many concepts without degrading model

[3]We say these are "representative" examples because we choose the *first* image generated for each prompt after fixing the random seed at 0, and use the same set of unlearned concepts, hyperparameters and prompts studied in [31]. Thus, these figures highlight the degree of concept resurgence in a "typical" case.

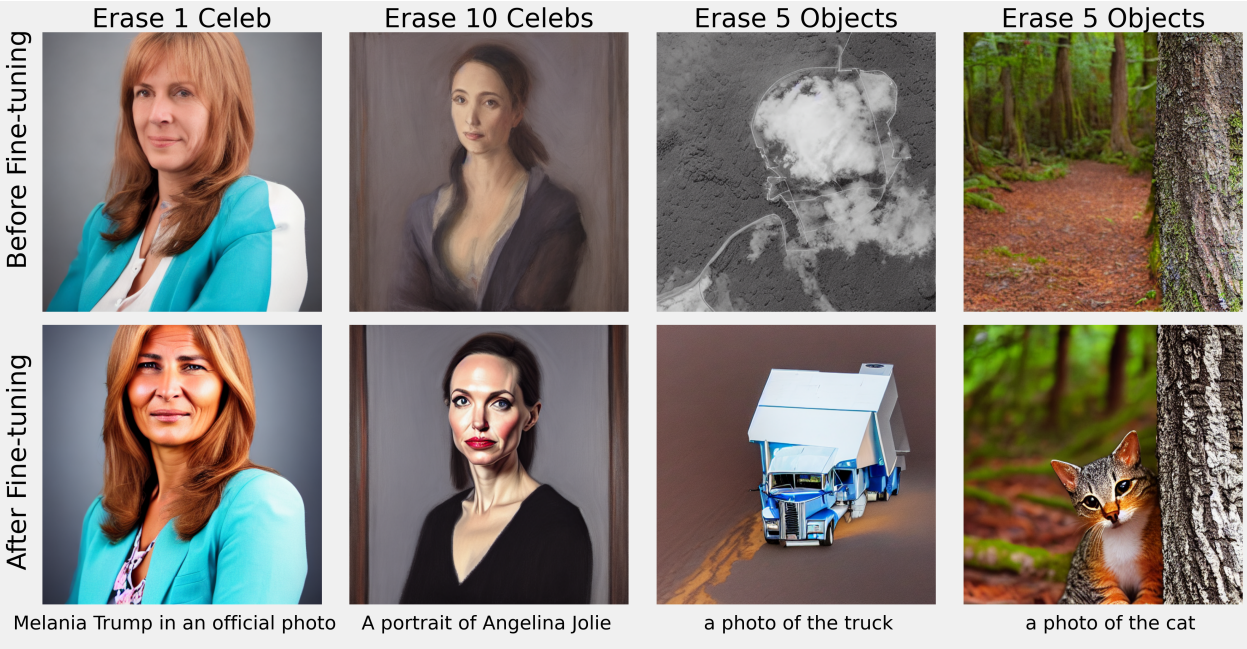

Figure 4: Selected images generated by SD v1.4 in each unlearning task before (top row) and after (bottom row) subsequent fine-tuning. In each task, the model initially unlearns the target concept; e.g., how to generate images of Melania Trump. However, fine-tuning on unrelated images can inadvertently reintroduce the erased concepts. In the "Erase 5 Objects" task, the model simultaneously unlearns five concepts (one of which is the ability to generate images of cats).

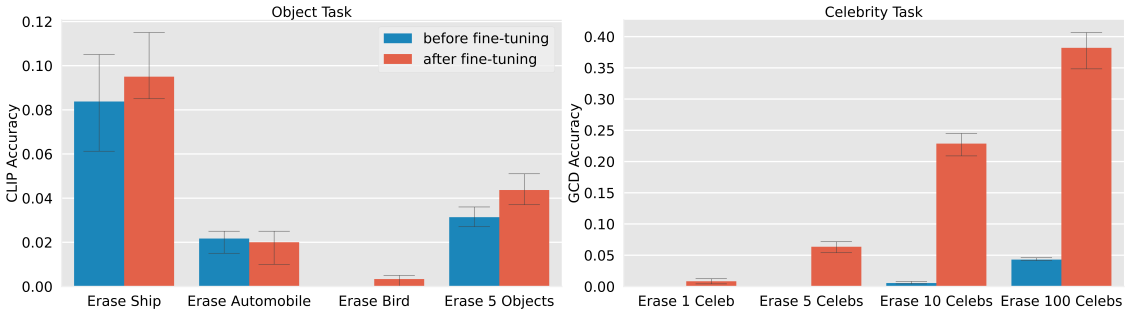

Figure 5: Quantifying the degree of concept resurgence in object and celebrity erasure. The celebrity erasure task demonstrates particularly severe resurgence as the number of unlearned celebrities grows large. Resurgence occurs to a more limited degree in the object erasure task.

performance on other, unrelated tasks). The vulnerability appears in the object erasure task as well, albeit to a lesser degree. We emphasize however that in many contexts, even rare concept resurgence presents unacceptable risks. In the remainder of this work, we seek to systematically characterize *when* and *why* this resurgence occurs, focusing first on algorithmic choices (Section 4) and then on characteristics of the fine-tuning dataset (Section 5).

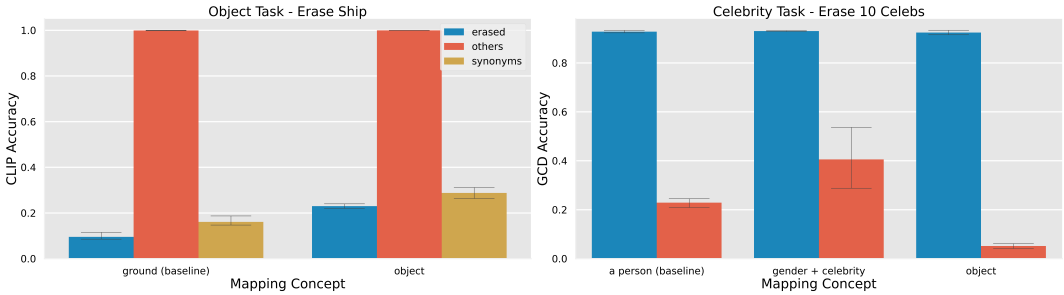

Figure 6: Effect of different generic mapping concepts on concept resurgence. More generic and unrelated concepts greatly reduce resurgence in both tasks, with near-total elimination in the celebrity erasure task when using objects as the mapping concept.

4 Algorithmic Factors Driving Concept Resurgence

The compositional updating pipeline involves several algorithmic choices that can either contribute to or mitigate the risk of concept resurgence. We focus on common choices made across most concept unlearning algorithms for diffusion models, using MACE as our baseline technique to assess their impact. These choices include regularization in the cross-attention refinement update, mapping the embedding of the concept to be unlearned to more general concepts, and the selection of the algorithm used for fine-tuning. In the following two sections we focus our attention on the `erase 10 celebrities` and `erase ship` tasks as we vary other components of the training pipeline.

At a high level, we hypothesize that resurgence occurs because unlearning does not update the model parameters to be sufficiently "far away" from the pretrained weights. Thus, although unlearning may initially suppress the generation of unwanted concepts, even a modest degree of fine-tuning tends to shift model weights towards their initial state, and thus reintroduce seemingly erased concepts. To validate this hypothesis, we first assess the impact of three algorithmic factors on concept resurgence: choice of mapping concept, regularization, and the fine-tuning algorithm.

4.1 Mapping Concepts

A key algorithmic component in many approaches to concept unlearning in diffusion models, including MACE and Unified Concept Editing (UCE) [13] is modifying the cross-attention mechanism. This mechanism is responsible for encoding the prompt into an embedding that the diffusion process conditions on. In particular, MACE proposes to modify the cross-attention weights found via the solution to the following optimization problem:

$$\min_{\mathbf{W}'_k} \underbrace{\sum_{i=1}^{n} \|\mathbf{W}'_k \cdot \mathbf{e}_i^{\text{co}} - \mathbf{W}_k \cdot \mathbf{e}_i^{\text{gen}}\|_2^2}_{\substack{\text{guides embeddings of model} \\ \text{towards those of generic concept}}} + \lambda_1 \underbrace{\sum_{i=n+1}^{n+m} \|\mathbf{W}'_k \cdot \mathbf{e}_i^{\text{prior}} - \mathbf{W}_k \cdot \mathbf{e}_i^{\text{prior}}\|_2^2}_{\substack{\text{ensures embeddings of other concepts} \\ \text{remain accurate}}}, \tag{1}$$

where $\mathbf{W}_k$ is the set of pretrained weights, $\mathbf{e}_i^{\text{co}}$ is the embedding of the i^{th} word that co-exists in the prompt with the concept to be unlearned (e.g. the photo of {concept}), and $\mathbf{e}_i^{\text{gen}}$ is the embedding of the i^{th} word that co-exists with the concept to be unlearned if that concept was replaced with its more generic concept (e.g. *the photo of {generic concept}*). Finally, $\mathbf{e}_i^{\text{prior}}$ is an embedding for a concept that we would like to preserve. Typically, the preserved concepts are generic ones that the model should still be able to generate after unlearning. The second term

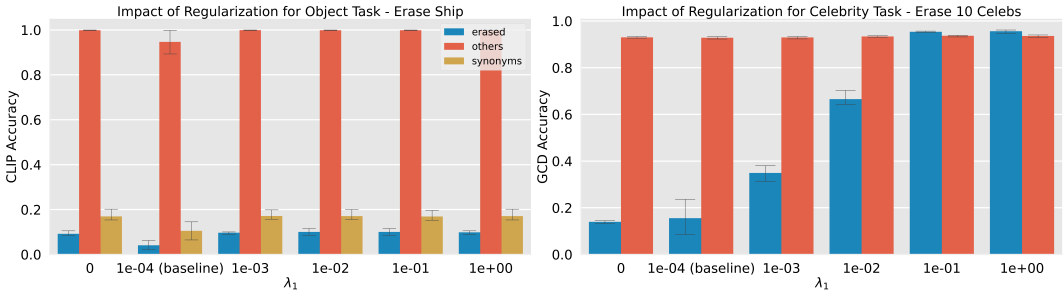

Figure 7: Impact of regularization in the cross-attention refinement weight update in both the object erasure (left) and celebrity erasure (right) tasks. Increasing regularization increases concept resurgence in the celebrity erasure task, but has little impact on the object erasure task.

acts as a regularizer that keeps the new weights close to the original weights and thus prevent performance degradation on unrelated concepts.

In this section we focus on the choice of the generic *mapping concept*. This impacts both the cross-attention refinement objective and the final update applied to the model, which fuses multiple LoRA modules corresponding to each unlearned concept. As described in Eq. (1), the embeddings of the words which co-occur with the concept to be unlearned are mapped to the corresponding embeddings for the mapping concept. To test our initial hypothesis, we select mapping concepts which are increasingly unrelated to the concept to be unlearned. This acts as a proxy for moving the initial model farther and farther away from its initial pretrained weights. For `erase 10 celebrities` we experiment with the following mapping concepts: "a person" (the baseline), "an object" and "a {male, female} celebrity," the latter two of which correspond to more general and more specific concepts, respectively. For `erase ship` we experiment with "ground" (the baseline) and "object." We present these results in Figure 6.

As illustrated in Figure 6, concept resurgence is quite sensitive to the choice of mapping concept. In the object erasure task, mapping "ship" to a different but *specific* concept ("ground") reduces resurgence more effectively than mapping it to a general concept like ("object"). In the celebrity erasure task, mapping each celebrity to a more specific concept (e.g., Jennifer Aniston $\rightarrow$ "a female celebrity") leads to more severe resurgence than if erased celebrities are mapped to more generic concepts. These findings are consistent with the hypothesis that while unlearning can initially suppress the generation of the unlearned concepts, it may do so through small changes in the parameter space, which can be easily undone by even modest degrees of further fine-tuning.

4.2 Regularization

In this section we study the impact of the regularization parameter λ_1 (recall Equation (1)) and assess its impact in the degree of concept resurgence. We run the MACE algorithm with $\lambda_1 \in \{0, .0001, .001, .01, .1, 1\}$ in the `erase 10 celebrities` and `erase ship` tasks. Intuitively, small values of λ_1 allow for large model updates, and thus prioritize unlearning the unwanted concepts at the expense of possibly degrading overall model performance. In contrast, large values of λ_1 only allow the unlearning algorithm to make small perturbations to the initial model weights.

As Figure 7 demonstrates, the degree of regularization is correlated with the degree of concept resurgence, particularly in the celebrity erasure task. Once the regularization parameter surpasses $\lambda =$ 1e-03 we see that more than half of the images generated in the forget set are accurately classified as their concept, indicating that larger updates to the initial pretrained weights are needed to ensure the prevention of concept resurgence. In the object task, we find that the regularization does not

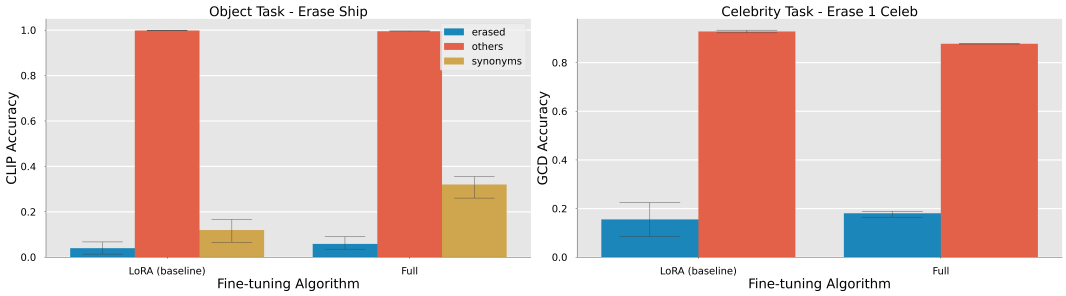

Figure 8: The choice of fine-tuning algorithm, whether using LoRA or full parameter fine-tuning, does not affect the extent of concept resurgence in either the object erasure tasks (left) or celebrity erasure tasks (right).

impact concept resurgence. Based on our results in Sec. 4.1, it is clear that the impact of mapping concepts on the phenomena of concept resurgence is first order. The object erasure task uses the mapping concept *ground* which is completely unrelated to ship. Celebrity erasure uses *person* as the mapping concept, which is much more related. Decreasing regularization helps prevent concept resurgence when the mapping concepts chosen are more related to the unlearned concepts.

4.3 Fine-Tuning Algorithm

Finally, we investigate the impact of fine-tuning the unlearned models with low rank adaptation (LoRA) instead of traditional full-parameter fine-tuning algorithms. LoRA is the most common method for fine-tuning large scale diffusion models (like Stable Diffusion v1.4). Although full parameter fine-tuning is far less efficient than LoRA, and thus not always feasible to implement in practice, this experiment seeks to investigate the effect of full parameter fine-tuning on concept resurgence. In particular, full parameter fine-tuning will tend to make larger incremental updates to the model parameters than LoRA under otherwise similar starting conditions.

Fig. 8 demonstrates that concept resurgence is largely insensitive to the choice of fine-tuning algorithm, and, if anything, that full fine-tuning can *exacerbate* the degree of concept resurgence in some cases. This provides additional evidence that vulnerability to concept resurgence is largely due to the initial unlearning algorithm rather than choices made at the fine-tuning stage. In particular, these results, in conjunction with the regularization and mapping concept experiments above, lend support to the hypothesis that larger parameter updates are required in the unlearning phase to prevent concept resurgence.

5 Data-Dependent Factors Driving Concept Resurgence

The second component of the pipeline we investigate is the data the unlearned models are fine-tuned on. As discussed in Section 3, we choose not to focus on adversarial dataset constructions (including e.g., simply directly fine-tuning a model on concepts which were previously unlearned); We refer the reader to [33, 44, 9] for works on attacking or jailbreaking text-to-image diffusion models. Instead, we consider the kind of dataset constructions that occur as part of common and benign use. For example, the end user of an open source diffusion model may want to fine-tune the model to acquire new concepts which were excluded from the pretraining set and/or improve its performance on particular tasks of interest. In Section 3, we demonstrated that fine-tuning on random in-domain concepts (e.g., fine-tuning on randomly chosen celebrities that were not part of the erased set) can lead to concept resurgence. In this section, we seek to further investigate the role

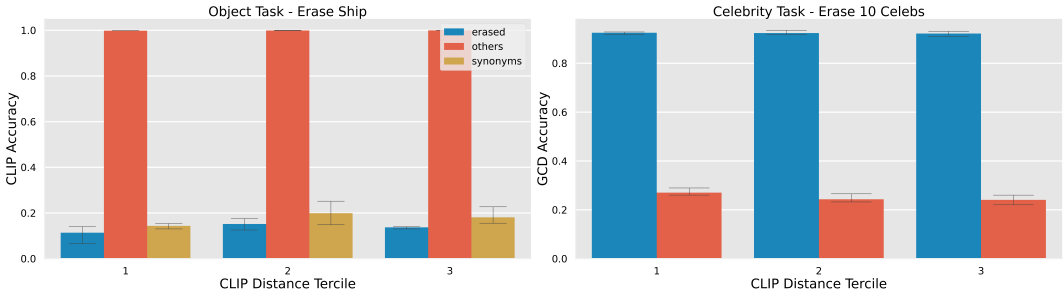

Figure 9: The degree of concept resurgence remains constant in both tasks, object (left) and celebrity (right) erasure regardless of the similarity (as measured by CLIP distance) of the fine-tuning dataset to the unlearned concepts.

of dataset construction in concept resurgence. First, we consider fine-tuning on in-domain concepts with varying levels of similarity to the concepts which are initially unlearned. Intuitively, it may be that fine-tuning on more related concepts can exacerbate the degree of concept resurgence. We then turn to fine-tuning on *out-of-domain* concepts; e.g., fine-tuning on images of randomly chosen objects after a celebrity unlearning task. We describe these experiments in more detail below.

5.1 CLIP Distance

First, we consider fine-tuning on unrelated but in-domain concepts as described above. We further segment these concepts by thresholding the CLIP distance to the unlearned concepts, which we use as a proxy for how "related" the fine-tuning dataset is to the unlearned concepts. In particular, for each task we find a publicly available dataset with hundreds of concepts in the same domain (e.g., for the celebrity unlearning task, this is a dataset of celebrity images). We describe these datasets in detail in Appendix C.2. For each concept in the corresponding dataset, we compute its minimum CLIP cosine similarity over the set of unlearned concepts. We then partition this fine-tuning set into three evenly sized subsets based on the percentiles of the CLIP cosine similarity. Finally, we randomly sample 10 concepts from each tercile to create different fine-tuning datasets which vary in their degree of "relatedness" to the unlearned concepts. We present these results in Figure 9.

As Figure 9 demonstrates, the degree to which the fine-tuning set is related to the unlearned concepts — at least as measured by CLIP cosine similarity — does not appear to meaningfully correlate with the degree of concept resurgence. This finding underscores the danger of concept resurgence even when fine-tuning on relatively unrelated data.

5.2 Out-of-Domain Concepts

Finally, to better understand the scope of concept resurgence, we curate an additional set of fine-tuning datasets which contain "out-of-domain" concepts which are wholly unrelated to those in the unlearning task. In particular, for each unlearning task, we fine-tune the resulting model on the random concept datasets from the other two unlearning tasks (e.g., for the celebrity unlearning task we further fine-tune the resulting model on the same randomly selected objects used as the initial fine-tuning set in the object erasure task).

As Figure 10 demonstrates, the degree of concept resurgence does vary with the domain of the fine-tuning set, particularly for the celebrity task. In particular, fine-tuning on in-domain images appears to exacerbate the risk of concept resurgence, while fine-tuning on out-of-domain images can mitigate it. Thus, although we found in the previous section that the degree of concept resurgence

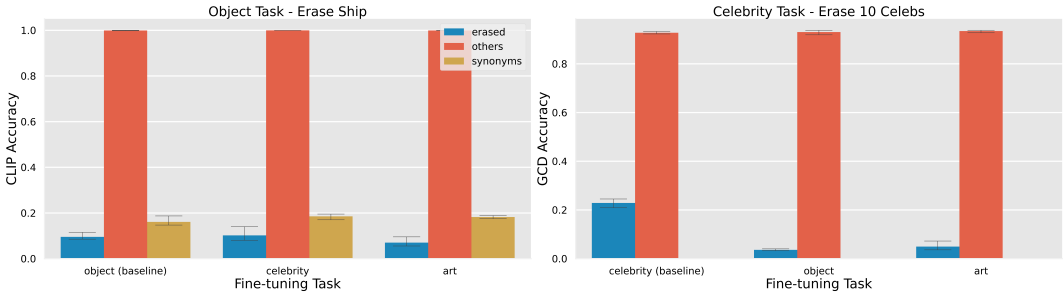

Figure 10: Concept resurgence is much less prevalent when fine-tuning on out-of-domain concepts. In both the object (left) and celebrity (right) tasks the degree of concept resurgence reduces when fine-tuning on out-of-domain concepts compared to in domain ones.

is relatively insensitive to variation in the fine-tuning set *within* each domain, it appears that fine-tuning on images which are wholly unrelated to the unlearned concepts can be safer than fine-tuning on more closely related images. This finding thus lends additional support to the hypothesis laid out in Section 4.

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 methods would allow model developers to precisely and permanently remove 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, though our evaluations focus on well-known academic benchmarks, 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 the vulnerabilities we demonstrate here).

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.

Acknowledgements

A.W. was funded in part by the Lister Brothers Career Development Professorship at MIT. V.S was funded in part by an MIT GenAI impact grant. A.S. acknowledges support from the Simons Foundation and NSF through award DMS-2031883, as well as from the DOE through award DE-SC0022199. R.A. and M.R acknowledge support from A Stephen A Schwarzman College of Computing Seed Grant.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Stephen Casper, Zifan Carl Guo, Shreya Mogulothu, Zachary Marinov, Chinmay Anand Deshpande, Rui-Jie Yew, Zheng Dai, and Dylan Hadfield-Menell. Measuring the success of diffusion models at imitating human artists. *ArXiv*, abs/2307.04028, 2023.

[7] 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.

[8] 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.

[9] 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.

[10] Michaela Drouillard, Ryan Spencer, Nikee Allen, and Tegan Maharaj. Quantifying likeness: A simple machine learning approach to identifying copyright infringement in (ai-generated) artwork. In *Proceedings of the 2024 ICML GenLaw Workshop*, 2024.

[11] 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.

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

[13] 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.

[14] 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.

[15] Jessica Grose. A.i. is making the sexual exploitation of girls even worse. *The New York Times*, 2024.

[16] 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.

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

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

[19] 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.

[20] 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.

[21] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

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

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

[26] Issie Lapowsky. The race to prevent 'the worst case scenario for machine learning'. *The New York Times*, 2023.

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

[28] Tsung-Yi Lin, Michael Maire, Serge J. 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*, 2014.

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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*, 2022.

[36] 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, 2022.

[37] 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.

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

[39] 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.

[40] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models, 2024.

[41] 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.

[42] Joshua Tenenbaum and William Freeman. Separating style and content. In M.C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

[43] 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.

[44] 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.

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

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

A Excluding Artistic Style and Explicit Content Erasure

As discussed in Section 3, celebrity and object erasure are two of the four benchmarks considered in [31]. The other two considered are artistic style erasure — unlearning the ability to generate images in the style of specific artists, e.g., due to copyright concerns — and explicit content erasure, particularly to suppress nudity. We exclude artistic style erasure due to the difficulty of quantifying the *degree* of a particular style in an image. In particular, unlike celebrity or object erasure, artistic style is not localized to specific regions in an image, and instead is a holistic (and partially subjective) property of the model output. For example, [10] note that "courts have emphasized the importance of considering the 'total concept and overall feel' [for determining whether copyright infringement has occurred], rather than relying on mechanical dissection or quantitative measures alone."

Characterizing artistic style replication (and copyright infringement more broadly) is rich topic in its own right, and we refer to [42, 40, 46, 6] for additional background.

We further exclude the explicit content benchmark due to the sensitive and unpredictable nature of the images which may be generated by the model, the lack of agreed upon standards for conducting such evaluations responsibly, and recent well publicized examples of the real-world harm that can result from synthetic but realistic nude images [26, 15]. Instead, we use the object and celebrity erasure tasks as representative but benign benchmarks on which to conduct our evaluations.

B Impact of Fine-tuning on Retained Concepts

Below we first examine the analogue of Figure 5 on the retained set, which is presented in Figure 11. Consistent with [31], we find that MACE preserves model performance on the set of retained concepts, and furthermore, subsequent fine-tuning does not degrade performance on this set.

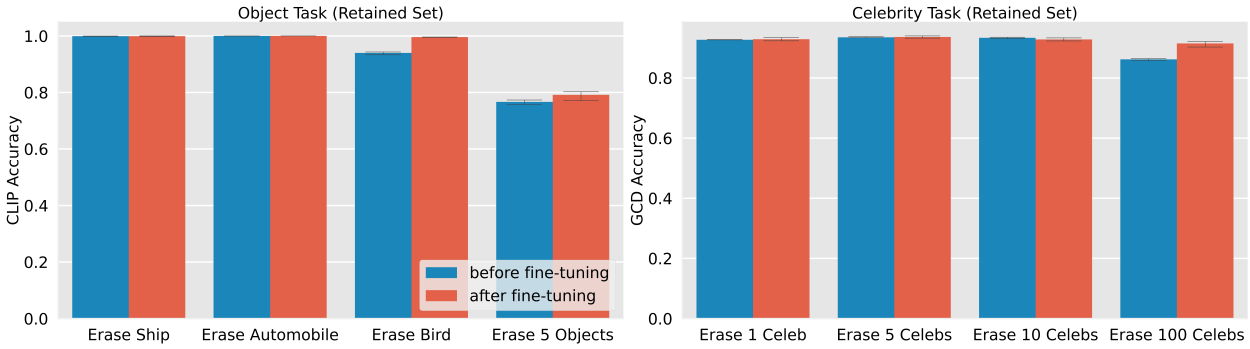

Figure 11: The performance on the retain set remains almost constant before and after fine-tuning in both tasks. It slightly increases when the number of concepts being erased is at its largest in both tasks (i.e. when erasing 5 objects and 100 celebrities).

As an additional sense check, we examine the CLIP and FID scores on random objects sampled from COCO-5K (as described in Section 3) before and after fine-tuning in the celebrity erasure tasks. These are presented in Figure 12 and Figure 13, respectively. We see that the CLIP scores remain almost identical, while the FID scores increase (i.e., degrade) after fine-tuning. The results of these three figures are thus broadly consistent with fine-tuning not degrading performance across a variety of tasks; if anything, concept resurgence can occur even if overall performance (i.e., on unrelated tasks) *decreases* slightly.

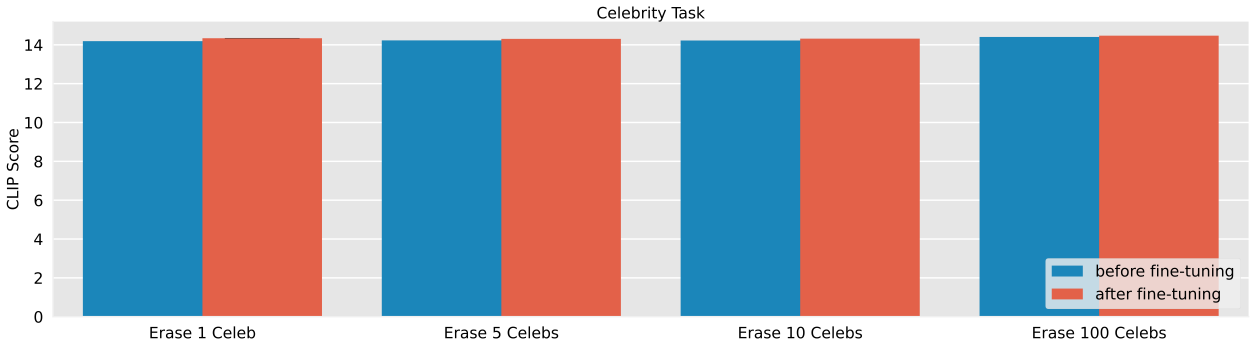

Figure 12: The CLIP score on unrelated objects sampled from COCO-5K remains almost constant before and after fine-tuning in the celebrity erasure task.

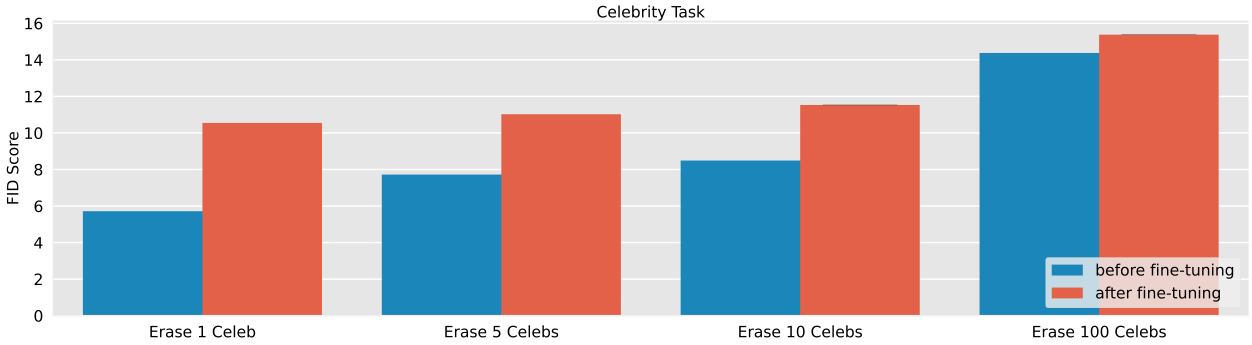

Figure 13: The FID score on unrelated objects sampled from COCO-5K degrades slightly after fine-tuning in the celebrity erasure task.

C Fine-tuning Dataset Curation

C.1 Random

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 [31] 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 [24] 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]."

C.2 CLIP Distance

In this section we provide additional details related to the dataset curation process for Sec. 5 for the different tasks. For celebrities, we start with all of the celebrities from the CelebA-HQ-Dialog [22, 21] dataset. We compute the CLIP embeddings of the 10177 celebrities from this dataset and the 10 celebrities being unlearned. Using these CLIP embeddings we compute the cosine similarity between every unlearned celebrity and the 10177 celebrities in CelebA-HQ-Dialog. We find the minimum and maximum similarity to be 0.17 and 0.80 respectively. We then construct terciles in this interval based on the minimum similarity between the celebrity in CelebA-HQ-Dialog and the unlearned celebrity, ensuring that at least 10 of the celebrities in CelebA-HQ-Dialog fall into each tercile. We then sample 10 celebrities from each tercile and generate a fine-tuning dataset with those celebrities in the same way as the random dataset.

Tercile	Cosine Similarity Interval	Celebrities
1	0.17 - 0.37	Elize Du Toit, Heather Marie Mardsen, Soleil Moon Frye, Eniko Mihalik, Mia Wasikowska, Ruslaan Mumtaz, Petra Cubonova, Karin Dor, Kathyrn Erbe, Justine Mattera
2	0.37 - 0.58	Delta Goodrem, Babs Jongh, Tom Green, Melissa Haro, Ratan Tata, Danielle Darrieux, Eike Batista, Johnny Borrell, Scott Stiner, Amy Davidson
3	0.58 - 0.80	Tamara Ecclestone, Bryan Cranston, Gregg Sulkin, Sigrid Agren, Ty Pennington, Noemie Lenoir, Jana Ina, Jonathan Tucker, Valerie Bertinelli

Table 1: Celebrity concepts used in each of the fine-tuning datasets for the CLIP distance experiments.

For objects, we start with all of the artists from the CIFAR100 [24] dataset. We compute the CLIP embeddings of the 100 objects from this dataset and the 5 objects being unlearned. Using these CLIP embeddings we compute the cosine similarity between every unlearned object and the 100 artists in CIFAR100. We find the minimum and maximum similarity to be 0.68 and 0.84 respectively. We then construct terciles in this interval based on the minimum distance between the object in CIFAR100 and the 5 unlearned objects, ensuring that at least 10 of the objects in CIFAR100 fall into each tercile. We then sample 10 objects from each tercile and generate a fine-tuning dataset with those objects in the same way as the random dataset.

Tercile	Cosine Similarity Interval	Objects
1	0.68 - 0.74	tulip, plain, bowl, pine tree, mountain, house, crab, willow tree, motorcycle, mushroom
2	0.75 - 0.79	streetcar, maple tree, seal, orange, cup, flatfish, sunflower, shark, hamster, aquarium fish
3	0.80 - 0.83	tiger, tank, turtle, cloud, orchid, road, elephant, rocket, bee, raccoon

Table 2: Object concepts used in each of the fine-tuning datasets for the CLIP distance experiments.