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

Anonymous authors
Paper under double-blind review

Abstract

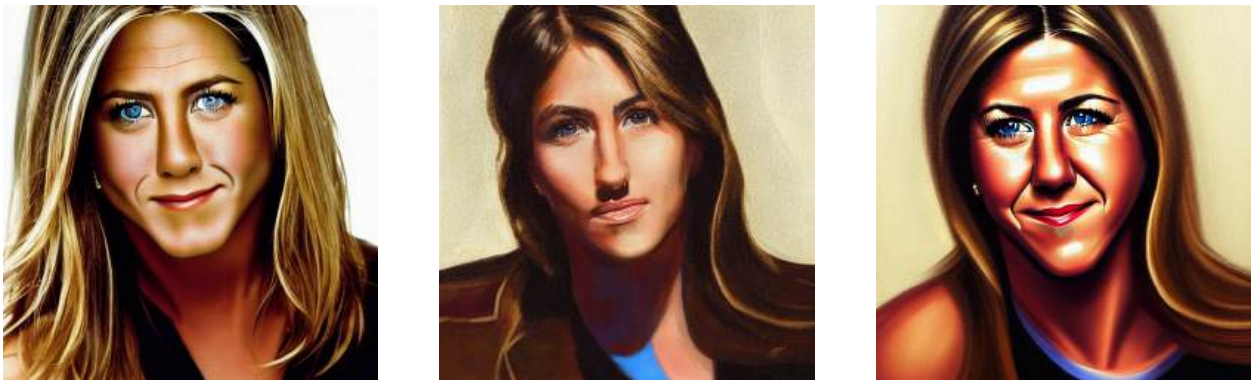
Text-to-image diffusion models rely on massive, web-scale datasets. Training them from scratch is computationally expensive, and as a result, developers often prefer to make incremental updates to existing models. These updates often compose fine-tuning steps (to learn new concepts or improve model performance) 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 (46; 20; 21). 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 (3; 31; 1) 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" (10).

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 (14; 27; 16; 13; 49; 22). For example, the current state-of-the-art algorithms such as "unified concept editing" (UCE) (14) and "mass concept erasure" (MACE) (27) 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 and when users prompt the model to generate a completely unrelated concept. 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.** We analyze a linear score-based diffusion model to understand, in a provable setting, *why* concept resurgence occurs after unlearning. Our analysis identifies two key factors that govern the strength of resurgence during fine-tuning: (1) the projection overlap between the forgotten subspace and the gradient directions introduced during fine-tuning, and (2) a curvature-limited sensitivity bound that quantifies how small gradient components in low-curvature subspaces can induce disproportionately large parameter updates. Crucially, our results show that some degree of resurgence is *inevitable* whenever there is nonzero overlap between the fine-tuning gradient subspace and the forgotten subspace, even if the overlap is small. Moreover, resurgence is most pronounced at early diffusion steps where gradients are strongest, but can also be amplified at intermediate-to-late steps when curvature is low and residual alignment persists.

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 provably investigate the fundamental drivers of concept resurgence.

2 Background and related work

Machine unlearning. We build on a growing literature on *machine unlearning* (2; 32; 24; 3; 17; 42; 39; 15; 24; 25; 28), 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" (10). 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 (48; 5). Our work focuses specifically on unlearning in the context of text-to-image diffusion models (19; 36). The literature on diffusion models has grown rapidly over the last few years; though we cannot provide a comprehensive overview here, we refer to (48) 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; 27; 12; 14; 49; 16; 13; 22), 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 seven recently proposed unlearning algorithms: ESD (13), SDD (22), UCE (13), MACE (27), SalUn (11), SHS (44), and EraseDiff (45). At a high level, ESD and SDD 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). EraseDiff performs unlearning similarly via a bi-level optimization problem. MACE and UCE 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. SalUn and SHS both start by identify the most influential parameters related to the concepts being unlearned and then finetune those parameters. 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 (6; 30; 4; 8; 33; 26). 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 seven of the most recent and performant unlearning methods discussed in Section 2: MACE, UCE, SDD, ESD, SalUn, SHS, and EraseDiff. 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 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 (29; 47; 9). Additionally, we focus on settings where the models retained high utility after unlearning. We describe the fine-tuning datasets and training details in Appendix C.

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 (27), 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 (27); 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) (18) 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 (40; 41; 43; 50), 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 seven 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 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 (7), 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 (37; 35; 34), 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 (38). 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.

[1]The GCD is a popular open source model for classifying celebrity images; (27) document that the GCD achieves $> 99\%$ top-1 accuracy on celebrity images sampled from Stable Diffusion v1.4.

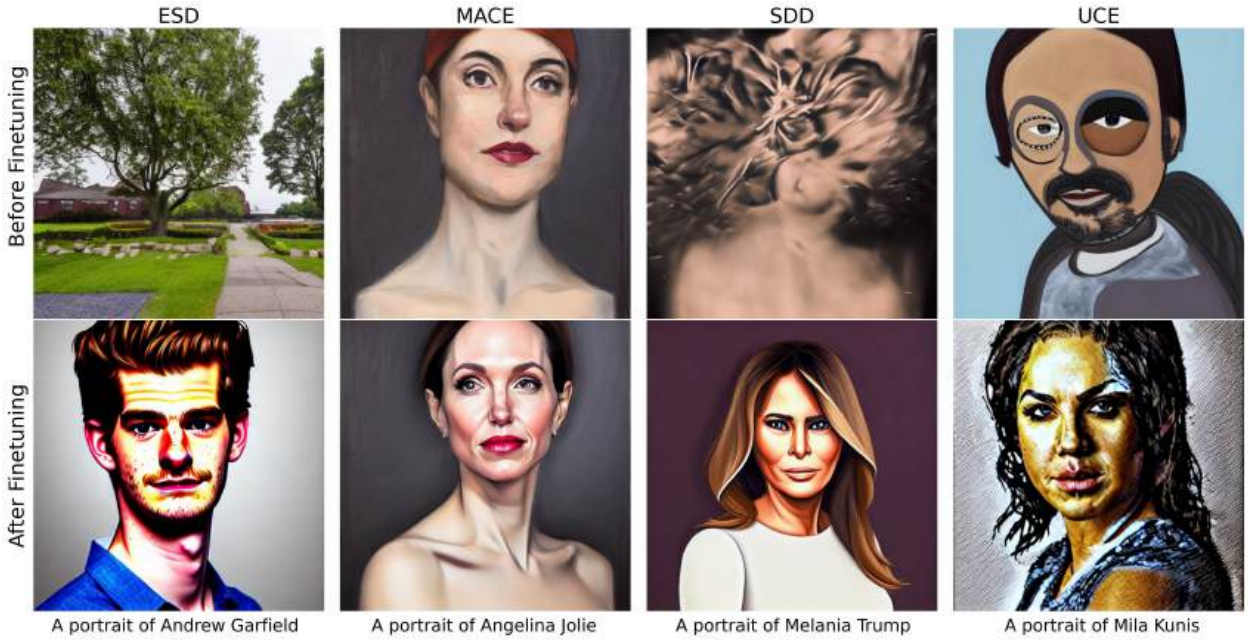


Figure 2: Selected images generated by SD v1.4 after initially applying each unlearning algorithm (top row) and after subsequent fine-tuning (bottom row) in the celebrity unlearning task. In each case, the model initially unlearns the target concept, e.g., how to generate images of Andrew Garfield. However,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

Object erasure. Finally, following (27), 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 (27).[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 seven algorithms, running the risk that developers or users can inadvertently reintroduce harmful or unwanted content.

In Table 1, 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, SDD, SalUn, SHS, and EraseDiff all 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.

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

Method	Celebrity		Copyright	
	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
EraseDiff	0.000 ± 0.000	0.693 ± 0.019	0.000 ± 0.000	0.367 ± 0.033
SHS	0.075 ± 0.019	0.893 ± 0.054	0.000 ± 0.000	0.133 ± 0.033
SalUn	0.363 ± 0.082	0.939 ± 0.056	0.000 ± 0.033	0.100 ± 0.067

(a) Celebrity and Copyright Tasks

Method	Object		Unsafe	
	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
EraseDiff	0.002 ± 0.002	0.995 ± 0.001	0.317 ± 0.181	0.876 ± 0.017
SHS	0.399 ± 0.274	0.999 ± 0.001	0.403 ± 0.058	0.848 ± 0.024
SalUn	0.831 ± 0.531	0.913 ± 0.065	0.840 ± 0.217	0.872 ± 0.008

(b) Object and Unsafe Tasks

Table 1: Unlearning performance before and after fine-tuning for Stable Diffusion v1.4. 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.

Incidental Concept Resurgence In conducting our object experiments, we uncover an even more concerning type of concept resurgence – the model will output an unlearned concept when prompted to generate an image of a retained concept. This means that a user can be prompting the model for an unrelated concept, and an unlearned concept is generated. We term this *incidental concept resurgence*. For example, when generating an image of an airplane that was retained, the model generates an image of an automobile that was unlearned before fine-tuning (example shown in Figure 3 and in Appendix A). Furthermore, we calculate the percentage of prompts on which this phenomenon occurs across all seven algorithms for our erase-three and erase-five object tasks. We find that ESD, UCE, MACE, and SDD all share this vulnerability on at least one of the tasks. Meanwhile, SalUn, SHS, and EraseDiff appear robust (Table 4).

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 seven 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 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

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

Figure 3: Images generated by the prompt "A photo of an airplane." Stable Diffusion v1.4 successfully generates this image (a), and Mass Concept Erasure (MACE) which unlearned {`cat`, `truck`, `automobile`, `ship`, `bird`}, partially generates this concept with the wing on the ground. However, subsequent fine-tuning on an unrelated set of randomly selected object images reintroduces the ability to generate the target concept when prompting with a completely ***unrelated*** concept ((c).

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 is only noticeable for ESD. For ESD, there is clear increase in resurgence as the number of concepts unlearned increases (Figure 9). In contrast, for the other six algorithms, the level of resurgence was not impacted as the number of concepts increased (see Appendix E).

4.2 The Impact of Algorithmic Choices on Resurgence

The seven 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 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

[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.

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 (14), 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 3), 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 seven algorithms is the subset of model parameters that are updated in the unlearning step. Section 3 focuses primarily on modifying the either the cross-attention layers (with the exception of MACE, which also updates the rest of the model parameters via LoRA fine-tuning) or the automatically selected parameter subset (i.e. SalUn and SHS). 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 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 (Figure 16).

Finetuning Hyperparameters Finally, we investigate how hyperparameter choices such as dataset size and number of fine-tuning steps impact the severity of resurgence. In Appendix J, we show that even with much smaller amounts of data or smaller amounts of fine-tuning steps that resurgence still occurs. For example, for MACE on our erase 10 celebrities task, only 20 fine-tuning steps are necessary for resurgence to occur with 250 samples.

5 Why Does Concept Resurgence Occur?

To better understand the conditions under which forgotten concepts can resurface during fine-tuning, we analyze a *linear score-based diffusion model*. Our results illustrate that *any non-zero overlap* between the subspace of the forgotten concept and the subpsace spanned by the fine-tuning gradients is sufficent to induce resurgence in this setting. Let $\mathcal{C} \subset \mathbb{R}^d$ denote the erased concept subspace and let $\mathcal{D}_{\text{FT}}$ be the fine-tuning dataset whose per-example gradients span a subspace $\mathcal{S}$. We assume that the orthogonal projection of $\mathcal{S}$ onto $\mathcal{C}$, denoted $P_{\mathcal{C}}(\mathcal{S})$, is nonzero (i.e., the subspaces are not orthogonal). This assumption reflects the realistic likelihood of incidental alignment in high-dimensional settings. Although this overlap is sufficient to trigger some degree of resurgence, it does not account for the magnitude of the effect. Our goal is to characterize how this seemingly weak interaction can be amplified into meaningful resurgence. In particular, our analysis identifies two bounds that characterize how resurgence occurs in this model:

- *Gradient resurgence bound.* This bound identifies when fine-tuning gradients reappear in the forgotten subspace $\mathcal{C}$, despite prior unlearning. It shows that nonzero gradient mass arises in $\mathcal{C}$ whenever there is residual alignment between the fine-tuning subspace $\mathcal{S}$ and $\mathcal{C}$. Formally:

$$\|P_{\mathcal{C}}\left(\nabla_W \mathcal{L}_t\right)\|_F \geq 2\sqrt{1-\alpha_t} \cdot \sqrt{\gamma(\mathcal{S},\mathcal{C})},$$

  where $\gamma(\mathcal{S},\mathcal{C}) \triangleq \lambda_{\min}(P_{\mathcal{S}} P_{\mathcal{C}} P_{\mathcal{S}})$ measures the worst-case leakage from $\mathcal{S}$ into $\mathcal{C}$. This overlap ensures that even when concepts in $\mathcal{C}$ have been suppressed, fine-tuning gradients computed from noise-perturbed data can reintroduce them if they are not fully orthogonal to the directions encoded in the new task. Notably, this bound is most active at early diffusion timesteps, where $1-\alpha_t$ is large and thus amplifies the residual error when there is *any* amount of overlap.

- *Curvature-limited sensitivity.* This bound captures the model's geometric sensitivity to reactivation. Even if gradient mass in $\mathcal{C}$ is small, the induced update can be large if the curvature in those directions is low. Formally, for any update ΔW supported in $\mathcal{C}$, we have:

$$\|P_{\mathcal{C}}\Delta W\|_F \geq \frac{2\sqrt{1-\alpha_t}\cdot\sqrt{\gamma(\mathcal{S},\mathcal{C})}}{\alpha_t\lambda_{\max}^{\mathcal{C}}+(1-\alpha_t)},$$

  where $\lambda_{\max}^{\mathcal{C}} := \lambda_{\max}(P_{\mathcal{C}}\Sigma P_{\mathcal{C}})$ is the maximum variance in the forgotten subspace. This bound reveals a key amplification mechanism: low-variance directions are highly sensitive to reactivation, since small gradients can produce large updates when curvature is shallow.

Proposition 5.1 (Linear diffusion model resurgence)**.** *Assume a linear diffusion model with residual of the form*

$$\epsilon_W(x_t,t) := Wx_t - \epsilon$$

for some matrix $W \in \mathbb{R}^{d\times d}$, where $\epsilon \sim \mathcal{N}(0,I)$ is independent Gaussian noise. Let $\mathcal{C} \subset \mathbb{R}^d$ be a subspace, and let $\mathcal{D}_{\mathrm{FT}}$ be a fine-tuning dataset whose induced gradient directions span a subspace $\mathcal{S}$. Let $P_{\mathcal{C}} = U_{\mathcal{C}}U_{\mathcal{C}}^\top$ and $P_{\mathcal{S}} = U_{\mathcal{S}}U_{\mathcal{S}}^\top$ denote the orthogonal projection matrices onto $\mathcal{C}$ and $\mathcal{S}$, respectively. Define the leakage

$$\gamma(\mathcal{S},\mathcal{C}) \triangleq \lambda_{\min}(P_{\mathcal{S}}P_{\mathcal{C}}P_{\mathcal{S}}).$$

Let $x_0 \sim \mathcal{D}_{\mathrm{FT}}$ have covariance Σ, and define the forward-corrupted input as

$$x_t := \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon \quad \textit{so that} \quad \Sigma_t := \mathbb{E}[x_t x_t^\top] = \alpha_t\Sigma + (1-\alpha_t)I.$$

Let $\lambda_{\max}^{\mathcal{C}} \triangleq \lambda_{\max}(P_{\mathcal{C}}\Sigma P_{\mathcal{C}})$ and suppose $P_{\mathcal{C}}\mathcal{S} \neq 0$. Then if the prior unlearning was successful, i.e. $P_{\mathcal{C}}W = 0$, we obtain bounds characterizing resurgence:

1. **Gradient resurgence:** *The fine-tuning gradient projected into $\mathcal{C}$ satisfies:*

$$\|P_{\mathcal{C}}\left(\nabla_W\mathcal{L}_t\right)\|_F \geq 2\sqrt{1-\alpha_t}\cdot\sqrt{\gamma(\mathcal{S},\mathcal{C})}.$$

2. **Curvature-limited sensitivity:** *The update $\Delta W \in \mathbb{R}^{d\times d}$ in the weight matrix supported in the forgotten subspace $\mathcal{C}$ satisfies:*

$$\|P_{\mathcal{C}}\Delta W\|_F \geq \frac{2\sqrt{1-\alpha_t}\cdot\sqrt{\gamma(\mathcal{S},\mathcal{C})}}{\alpha_t\lambda_{\max}^{\mathcal{C}}+(1-\alpha_t)}.$$

We provide a proof of the gradient resurgence bound in Appendix G and the proof of the curvature-limited sensitivity bound to Appendix H. The basic idea for gradient bound is to notice the norm of the gradient is lower bounded by the norm of the Frobenius norm $A := \mathbb{E}[\epsilon_W(x_t,t)x_t^\top]$ restricted to $\mathcal{C}$ multiplied by the overlap term $\gamma(\mathcal{S},\mathcal{C})$ which simplifies to $1-\alpha_t$ when exact unlearning has occurred. The bound on the update uses a bound that relies on the loss function for the weight W being quadratic.

Together, these bounds clarify the structural and dynamical factors that govern resurgence after unlearning. Specifically, we identify two distinct contributors: (1) *conceptual overlap*, captured by the gradient resurgence bound, quantifies when fine-tuning gradients reappear in the forgotten subspace $\mathcal{C}$. The bound depends on the overlap between $\mathcal{C}$ and the supervision subspace $\mathcal{S}$, via the leakage term $\gamma(\mathcal{S},\mathcal{C})$, and scales linearly with $\sqrt{1-\alpha_t}$, reflecting the increased influence of noise-perturbed input at early diffusion steps. In the simplified setting where $P_{\mathcal{C}}W = 0$, this term isolates reactivation due purely to residual alignment. More generally, when signal is injected into W, the bound acquires an additional contribution proportional to $P_{\mathcal{C}}W\Sigma$, which can dominate in late timesteps and further amplify resurgence; (2) *amplification*, captured by the curvature-limited sensitivity bound, governs how strongly the model responds to gradient mass in $\mathcal{C}$. Even small gradients can induce large parameter updates when the loss curvature is low. This effect is most pronounced when $\mathcal{C}$ aligns with low-variance directions in the data (i.e., small $\lambda_{\max}^{\mathcal{C}}$), and when α_t is large, but not too close to 1 so that the noise has diminished but curvature remains anisotropic. In contrast, early timesteps ($\alpha_t \ll 1$) introduce strong isotropic curvature, suppressing updates and making reactivation less likely.

6 Discussion and Limitations

Our investigation opens several important directions for future work. First, our theoretical analysis is restricted to the linear setting, and it remains an open question whether similar characterizations of concept resurgence extend to nonlinear models. Exploring such extensions could inform new strategies for mitigating resurgence and improving the robustness of unlearning procedures. Second, our empirical evaluation is limited to standard academic benchmarks and synthetic settings. Further research is needed to assess the practical relevance of concept resurgence in real-world deployments, particularly in scenarios involving long-horizon or compositional fine-tuning, where interleaved updates may amplify 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.

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)*, pages 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*, pages 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, pages 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)*, pages 5253–5270, 2023.

[6] 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, pages 896–911. ACM, 2021.

[7] 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.

[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] 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.

[11] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

[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, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023.

[14] 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*, pages 5111–5120, 2024.

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

[16] 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*, pages 73–88. Springer, 2025.

[17] 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.

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

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

[20] 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.

[21] 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.

[22] 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.

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

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

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

[26] 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, pages 14115–14123, 2024.

[27] 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.

[28] 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.

[29] 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.

[30] 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*, pages 7691–7700, 2022.

[31] 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.

[32] 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.

[33] 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*, pages 1932–1942, 2023.

[34] 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.

[35] Javier Rando, Daniel Paleka, David Lindner, Lennard Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *ArXiv*, abs/2210.04610, 2022.

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

[37] 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)*, pages 22522–22531, 2022.

[38] 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*, 2022.

[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, 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)*, pages 6048–6058, 2022.

[41] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *ArXiv*, abs/2305.20086, 2023.

[42] 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.

[43] James Vincent. Ai art tools stable diffusion and midjourney targeted with copyright lawsuit. *The Verge*, 2023.

[44] Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision*, pages 367–384. Springer, 2024.

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

[46] 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)*, pages 374–382, 2019.

[47] 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)*, pages 897–912. IEEE, 2024.

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

[49] 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*, pages 1755–1764, 2024.

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