DATA UNLEARNING IN DIFFUSION MODELS

Silas Alberti* **Kenan Hasanaliyev*** **Manav Shah** **Stefano Ermon**

Stanford University
{alberti,kenanhas,manavs,ermon}@cs.stanford.edu

ABSTRACT

Recent work has shown that diffusion models memorize and reproduce training data examples. At the same time, large copyright lawsuits and legislation such as GDPR have highlighted the need for erasing datapoints from diffusion models. However, retraining from scratch is often too expensive. This motivates the setting of data unlearning, i.e., the study of efficient techniques for unlearning specific datapoints from the training set. Existing concept unlearning techniques require an anchor prompt/class/distribution to guide unlearning, which is not available in the data unlearning setting. General-purpose machine unlearning techniques were found to be either unstable or failed to unlearn data. We therefore propose a family of new loss functions called Subtracted Importance Sampled Scores (SISS) that utilize importance sampling and are the first method to unlearn data with theoretical guarantees. SISS is constructed as a weighted combination between simpler objectives that are responsible for preserving model quality and unlearning the targeted datapoints. When evaluated on CelebA-HQ and MNIST, SISS achieved Pareto optimality along the quality and unlearning strength dimensions. On Stable Diffusion, SISS successfully mitigated memorization on nearly 90% of the prompts we tested. We release our code online.[1]

1 INTRODUCTION

The recent advent of diffusion models has revolutionized high-quality image generation, with large text-to-image models such as Stable Diffusion (Rombach et al., 2022) demonstrating impressive stylistic capabilities. However, these models have been shown to memorize and reproduce specific training images, raising significant concerns around data privacy, copyright legality, and the generation of inappropriate content (Carlini et al., 2023; Cilloni et al., 2023). Incidents such as the discovery of child sexual abuse material in LAION (Thiel, 2023; Schuhmann et al., 2022) as well as the need to comply with regulations like the General Data Protection Regulation and California Consumer Privacy Act that establish a "right to be forgotten" (Hong et al., 2024; Wu et al., 2024), underscore the urgency of developing effective methods to remove memorized data from diffusion models.

Retraining on a new dataset is often prohibitively expensive, and the bulk of traditional machine unlearning techniques have been built for classical supervised machine learning (Cao & Yang, 2015; Ginart et al., 2019; Izzo et al., 2021; Bourtoule et al., 2021). Recently, a new wave of research on unlearning in diffusion models has emerged, but it has focused almost exclusively on *concept unlearning* in text-conditional models (Gandikota et al., 2023; Kumari et al., 2023; Zhang et al., 2023; Gandikota et al., 2024; Schramowski et al., 2023; Heng & Soh, 2023; Fan et al., 2024). These approaches aim to remove higher-level concepts, e.g., the styles of painters or nudity, rather than specific datapoints from the training data needed to battle unwanted memorization. This paper focuses on the problem of efficient machine unlearning in diffusion models with the objective of removing specific datapoints, a problem that we refer to as *data unlearning*.

Unlike concept unlearning, the data unlearning setting has a concrete gold standard: retraining without the data to be unlearned. The goal of data unlearning is to achieve unlearning performance as close as possible to retraining while using less computational resources. To quantify unlearning performance,

*Equal contribution
[1] https://github.com/claserken/SISS

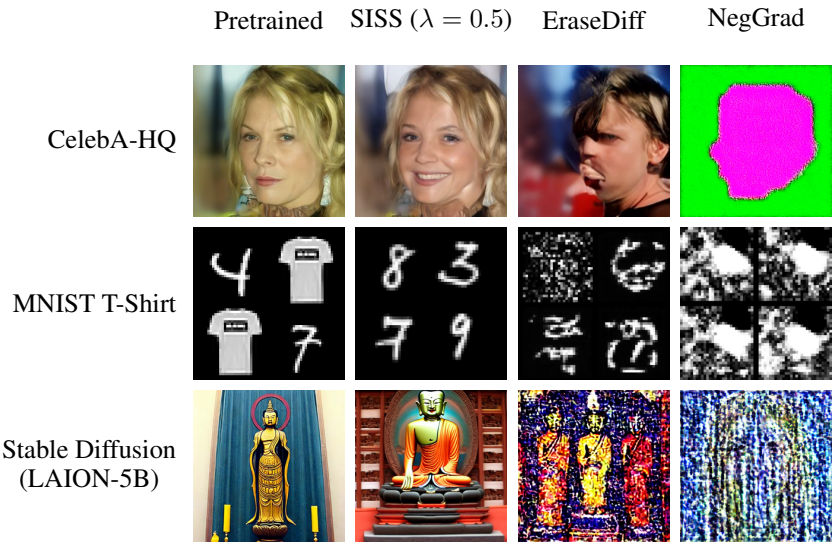


Figure 1: Examples of quality degradation across unlearning methods. On all 3 datasets, we find that our SISS method is the only method capable of unlearning specific training datapoints while maintaining the original model quality. See Tables 1, 2 and Figure 6a for complete quantitative results on quality preservation.

we focus on three separate areas: the degree of unlearning, the amount of quality degradation after unlearning, and the amount of compute needed. Examples of these areas are highlighted in Figures 1 and 2.

When applied to data unlearning, general-purpose machine unlearning techniques face certain limitations: naive deletion (fine-tuning on data to be kept) tends to be slow in unlearning, while NegGrad (gradient ascent on data to be unlearned) (Golatkar et al., 2020) forgets the data to be kept, leading to rapid quality degradation. State-of-the-art class and concept unlearning techniques do not apply to our setting because they require an anchor prompt/class/distribution to guide the unlearning towards which we do not assume access to. For example, Heng & Soh (2023) select a uniform distribution over pixel values to replace the 0 class on MNIST; however, the desired unlearning behavior would be to instead generate digits $1-9$ when conditioned on 0. Even if one selects the target distribution to be a random example/label from all other classes as Fan et al. (2024) do, it is not always clear what "all other classes" are in the text-conditional case. Furthermore, these prior experiments are targeting prompts/classes instead of datapoints and do not apply in the case of unconditional diffusion models. A notable exception is EraseDiff (Wu et al., 2024) which can unlearn data by fitting the predicted noise to random noise targets that are not associated with a prompt/class.

In this work, we derive an unlearning objective that combines the objectives of naive deletion and NegGrad. For further computational efficiency, we unify the objective through importance sampling, cutting the number of forward passes needed to compute it by half. We term our objective Subtracted Importance Sampled Scores (SISS). As seen in Figure 1, SISS allows for the computationally efficient unlearning of training data subsets while preserving model quality. It does so because the naive deletion component preserves the data to be kept, while the NegGrad component targets the data to be unlearned. The addition of importance sampling balances between the two components through a parameter λ, where $\lambda = 0$ and $\lambda = 1$ behave like naive deletion and NegGrad, respectively. We find that $\lambda = 0.5$ behaves as a mixture of the two, giving the desirable combination of quality preservation and strong unlearning.

We demonstrate the effectiveness of SISS on CelebA-HQ (Karras et al., 2018), MNIST with T-Shirt, and Stable Diffusion. On all 3 sets of experiments, SISS preserved the original model quality as shown in Figure 1. On CelebA-HQ with the objective of unlearning a celebrity face, SISS was Pareto-optimal with respect to the FID and SSCD similarity metric in Figure 2, cutting the latter by over half. The base model for MNIST with T-Shirt was trained on MNIST (Deng, 2012) augmented

with a specific T-shirt from Fashion-MNIST (Xiao et al., 2017). The objective was to unlearn the T-shirts, and SISS was again found to be Pareto-optimal with respect to the Inception Score and exact likelihood, increasing the latter by a factor of 8. Finally, on Stable Diffusion, we found SISS to successfully mitigate memorization on almost 90% of the prompts we tested.

2 Related Work

Machine Unlearning. Machine unlearning is the notoriously difficult problem of removing the influence of datapoints that models were previously trained on (Cao & Yang, 2015; Bourtoule et al., 2021; Shaik et al., 2024). Over the past years, it has received increased attention and relevance due to privacy regulation such as the EU's Right to be Forgotten (Ginart et al., 2019; Izzo et al., 2021; Golatkar et al., 2020; Tarun et al., 2023). The first wave of methods mostly approached classical machine learning methods like linear and logistic regression (Izzo et al., 2021), k-means clustering (Ginart et al., 2019), statistical query learning methods (Cao & Yang, 2015), Bayesian methods (Nguyen et al., 2020) or various types of supervised deep learning methods (Golatkar et al., 2020; Tarun et al., 2023). Some of these methods require modifications to the training procedure, e.g., training multiple models on distinct dataset shards (Bourtoule et al., 2021; Golatkar et al., 2024), whereas others can be applied purely in post-training such as NegGrad (Golatkar et al., 2020) and BlindSpot (Tarun et al., 2023). Recently, generative models have become a popular paradigm. While unlearning in this paradigm is less explored, there are early approaches looking at Generative Adversarial Networks (GANs) (Kong & Alfeld, 2023), language models (Liu et al., 2024; Yao et al., 2024) and on diffusion models via sharding (Golatkar et al., 2024).

Memorization in Diffusion Models. Large-scale diffusion models trained on image generation have recently attracted the attention of copyright lawsuits since they are prone to memorizing training examples (Somepalli et al., 2023a; Carlini et al., 2023; Somepalli et al., 2023b; Webster, 2023). Somepalli et al. (2023a) showed that Stable Diffusion (Rombach et al., 2022) exhibits verbatim memorization for heavily duplicated training data examples. Webster (2023) classified different types of memorization, introducing types of partial memorization. Carlini et al. (2023) discusses various black-box extraction and membership inference attacks and demonstrates them successfully on Stable Diffusion. Most recently, mitigation strategies have been introduced, e.g., by manually modifying the text prompts (Somepalli et al., 2023b) or taking gradients steps in prompt space to minimize the magnitude of text-conditional noise predictions (Wen et al., 2024).

Concept Unlearning in Diffusion Models. Recently, the problem of unlearning has become popular in the context of diffusion models, though almost exclusively in the form of concept unlearning: while the classical setting of machine unlearning deals with forgetting specific datapoints from the training set – which we call *data unlearning* for clarity – the setting of *concept unlearning* deals with forgetting higher level concepts in text-conditional models, e.g., nudity or painting styles (Shaik et al., 2024; Gandikota et al., 2023; Kong & Chaudhuri, 2024; Kumari et al., 2023; Zhang et al., 2024; Heng & Soh, 2023). Zhang et al. (2023) introduces Forget-Me-Not, a concept unlearning technique that minimizes the cross-attention map for an undesired prompt and also introduces ConceptBench as a benchmark. Gandikota et al. (2023) find an alternate approach to concept unlearning fine-tuning by fitting to noise targets that are biased away from the predicted noise with respect to an undesirable prompt. Similarly, Schramowski et al. (2023) also bias the noise away from an undesired prompt but do so only at inference time. UnlearnCanvas is a benchmark introduced by Zhang et al. (2024) to measure concept unlearning for artistic styles and objects. EraseDiff (Wu et al., 2024) discusses the data unlearning setting but only studies the settings of unlearning classes or concepts. Lastly, Li et al. (2024) indeed studies data unlearning but solely in the case of image-to-image models. To the authors' knowledge, the data unlearning setting remains a gap in the diffusion model literature.

3 Preliminaries

We define the data unlearning problem as follows: given access to a training dataset $X = \{x_1, x_2, \ldots, x_n\}$ with n datapoints and a diffusion model ϵ_θ that was pretrained on X, our goal is to unlearn a k-element subset $A = \{a_1, a_2, \ldots, a_k\} \subset X$. We refer to A as the *unlearning set*.

More specifically, we wish to efficiently unlearn A through *deletion fine-tuning* which moves θ towards a set of parameters θ' so that $\epsilon_{\theta'}$ is no longer influenced by A. Ideally, $\epsilon_{\theta'}$ should behave as if it were trained from scratch on $X \setminus A$. In practice, however, retraining can be computationally infeasible. The key research question in data unlearning is identifying strategies for obtaining models that (1) preserve quality, (2) no longer generate A unless generalizable from $X \setminus A$, and (3) are efficient.

Consider the standard DDPM forward noising process (Ho et al., 2020) where for a clean datapoint x_0, the noisy sample x_t at time t is given by

$$q(x_t|x_0) = \mathcal{N}(x_t; \gamma_t x_0, \sigma_t \mathbf{I}). \tag{1}$$

The parameters γ_t and σ_t are set by the variance schedule. A simple approach to data unlearning is by fine-tuning on $X \setminus A$ where the objective is to minimize the simplified evidence-based lower bound (ELBO):

$$L_{X\setminus A}(\theta) = \mathbb{E}_{p_{X\setminus A}(x)}\mathbb{E}_{q(x_t|x)} \left\|\epsilon - \epsilon_\theta(x_t, t)\right\|_2^2 \tag{2}$$

where p_S refers to the discrete uniform distribution over any set S. We refer to this approach as *naive deletion* because it does not involve sampling from the unlearning set A. Another general-purpose machine unlearning approach is NegGrad (Golatkar et al., 2020) which performs gradient ascent on A, maximizing

$$L_A(\theta) = \mathbb{E}_{p_A(a)}\mathbb{E}_{q(a_t|a)} \left\|\epsilon - \epsilon_\theta(a_t, t)\right\|_2^2. \tag{3}$$

NegGrad is effective at unlearning A but is unstable in that the predicted noise will grow in magnitude, and the model will eventually forget data from $X \setminus A$.

More recently, EraseDiff (Wu et al., 2024) unlearns by minimizing the objective

$$L_{X\setminus A}(\theta) + \lambda \mathbb{E}_{\epsilon_f \sim \mathcal{U}(\mathbf{0}, \mathbf{I}_d)}\mathbb{E}_{p_A(a)}\mathbb{E}_{q(a_t|a)} \left\|\epsilon_f - \epsilon_\theta(a_t, t)\right\|_2^2$$

through a Multi-Objective Optimization framework. Other state-of-the-art diffusion unlearning approaches such as SalUn (Fan et al., 2024) and Selective Amnesia (Heng & Soh, 2023) are designed only for conditional models and cannot be extended to the data unlearning setting. When unlearning a class c, they require either fitting to the predicted noise of a different class $c' \neq c$ or specifying a distribution $q(\mathbf{x} \mid \mathbf{c})$ to guide ϵ_θ towards when conditioned on c, neither of which we assume access to.

4 Proposed Method: Subtracted Importance Sampled Scores (SISS)

Assume the data unlearning setting from Section 3 with dataset X of size n and unlearning set A of size k. The naive deletion loss from Eq. 2 can be split up as

$$\begin{aligned}
L_{X\setminus A}(\theta) &= \mathbb{E}_{p_{X\setminus A}(x)}\mathbb{E}_{q(x_t|x)} \left\|\epsilon - \epsilon_\theta(x_t, t)\right\|_2^2 = \sum_{x\in X\setminus A} \frac{1}{n-k}\mathbb{E}_{q(x_t|x)} \left\|\epsilon - \epsilon_\theta(x_t, t)\right\|_2^2 \\
&= \sum_{x\in X} \frac{1}{n-k}\mathbb{E}_{q(x_t|x)} \left\|\epsilon - \epsilon_\theta(x_t, t)\right\|_2^2 - \sum_{a\in A} \frac{1}{n-k}\mathbb{E}_{q(a_t|a)} \left\|\epsilon - \epsilon_\theta(a_t, t)\right\|_2^2 \\
&= \frac{n}{n-k}\mathbb{E}_{p_X(x)}\mathbb{E}_{q(x_t|x)} \left\|\frac{x_t - \gamma_t x}{\sigma_t} - \epsilon_\theta(x_t, t)\right\|_2^2 \\
&\quad - \frac{k}{n-k}\mathbb{E}_{p_A(a)}\mathbb{E}_{q(a_t|a)} \left\|\frac{a_t - \gamma_t a}{\sigma_t} - \epsilon_\theta(a_t, t)\right\|_2^2.
\end{aligned} \tag{4}$$

By employing importance sampling (IS), we can bring the two terms from Eq. 4 together, requiring only one model forward pass as opposed to two forward passes through ϵ_θ on both x_t and a_t. IS restricts us to two choices: we either pick our noisy sample from $q(\cdot \mid x)$ or $q(\cdot \mid a)$. However, a

defensive mixture distribution (Hesterberg, 1995) parameterized by λ allows us to weigh sampling between $q(\cdot \mid x)$ and $q(\cdot \mid a)$, giving us the following SISS loss function:

$$\ell_\lambda(\theta) = \mathbb{E}_{p_X(x)}\mathbb{E}_{p_A(a)}\mathbb{E}_{q_\lambda(m_t|x,a)} \left[\frac{n}{n-k} \frac{q\left(m_t|x\right)}{(1-\lambda)q\left(m_t|x\right) + \lambda q\left(m_t|a\right)} \left\| \frac{m_t - \gamma_t x}{\sigma_t} - \epsilon_\theta(m_t, t) \right\|_2^2 - \frac{k}{n-k} \frac{q\left(m_t|a\right)}{(1-\lambda)q\left(m_t|x\right) + \lambda q\left(m_t|a\right)} \left\| \frac{m_t - \gamma_t a}{\sigma_t} - \epsilon_\theta(m_t, t) \right\|_2^2 \right] \tag{5}$$

where m_t is sampled from the mixture distribution q_λ defined by a weighted average of the densities of $q(m_t|x)$ and $q(m_t|a)$

$$q_\lambda\left(m_t|x,a\right) \coloneqq (1-\lambda)q\left(m_t|x\right) + \lambda q\left(m_t|a\right). \tag{6}$$

Employing IS and a defensive mixture distribution preserves the naive deletion loss. That is,

$$\ell_\lambda(\theta) = L_{X\setminus A}(\theta) \forall \lambda \in [0,1]. \tag{7}$$

We further prove that gradient estimators of the two loss functions used to update model parameters during deletion fine-tuning are also the same in expectation.

Lemma 1. *In expectation, gradient estimators of a SISS loss function $\ell_\lambda(\theta)$ and the naive deletion loss $L_{X\setminus A}(\theta)$ are the same.*

Proof. Follows from Eq. 7 and linearity of expectation. See Appendix A.2 for a complete proof.

Notice that the second term in Eq. 4 is the same as the NegGrad objective in Eq. 3 up to a constant. Hence, to boost unlearning on A, we increase its weight by a factor of $1+s$ where $s > 0$ is a hyperparameter, referred to as the *superfactor*. The final weighted SISS loss $\ell_{s,\lambda}(\theta)$ can be written as

$$\mathbb{E}_{p_X(x)}\mathbb{E}_{p_A(a)}\mathbb{E}_{q_\lambda(m_t|x,a)} \left[\frac{n}{n-k} \frac{q\left(m_t|x\right)}{(1-\lambda)q\left(m_t|x\right) + \lambda q\left(m_t|a\right)} \left\| \frac{m_t - \gamma_t x}{\sigma_t} - \epsilon_\theta(m_t, t) \right\|_2^2 - (1+s)\frac{k}{n-k} \frac{q\left(m_t|a\right)}{(1-\lambda)q\left(m_t|x\right) + \lambda q\left(m_t|a\right)} \left\| \frac{m_t - \gamma_t a}{\sigma_t} - \epsilon_\theta(m_t, t) \right\|_2^2 \right]. \tag{8}$$

and is studied for stability and interpretability in Appendix A.1.

Despite Lemma 1, we find distinct SISS behavior for $\lambda = 0$ and $\lambda = 1$ that emulates naive deletion and NegGrad, respectively. We speculate that this discrepency is due to high gradient variance. Namely, for $\lambda = 0$, the SISS only selects noisy samples from $q(\cdot|x)$ that are high unlikely to come from $q(\cdot|a)$. As a result, the first term of the SISS loss dominates, resulting in naive deletion-like behavior. Similarly, for $\lambda = 1$, the second term of the SISS loss will dominate which matches NegGrad. Thus, in practice, we choose $\lambda = 0.5$ to ensure the beneficial properties of both naive deletion and NegGrad in SISS.

5 Experiments

We evaluate our SISS method, its ablations, EraseDiff, NegGrad, and naive deletion through unlearning experiments on CelebA-HQ, MNIST T-Shirt, and Stable Diffusion. The SISS ablations are defined as follows:

Setting $\lambda = 0$ and $\lambda = 1$. Using $\lambda = 0$ or $\lambda = 1$ effectively disables the use of the mixture distribution and can be viewed as using only importance sampling.

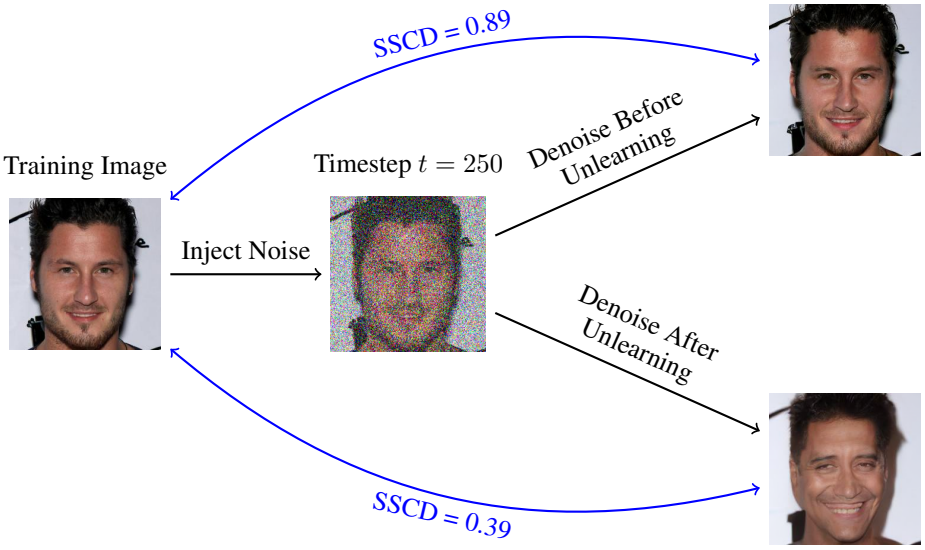


Figure 2: CelebA-HQ SSCD Metric Calculation. The process begins by taking the training face to be unlearned and injecting noise as part of the DDPM's forward noising process. Prior to unlearning, denoising the noise-injected face will result in a high similarity to the original training face. After unlearning, we desire for the denoised face to be significantly less similar to the training face.

SISS (No IS). The loss function defined in Eq. 4 after manipulating $L_{X \setminus A}(\theta)$ disables the use of importance sampling. Note, however, that it requires two forward passes through the denoising network ϵ_θ and is thus *doubly more expensive in compute and memory*.

Our model quality metrics are standard and given by the Frechet Inception Distance (FID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016), and CLIP-IQA (Wang et al., 2023). To evaluate the strength of unlearning, we employ the SSCD (Pizzi et al., 2022) to measure the similarity between celebrity faces before and after unlearning as in Figure 2. On MNIST T-Shirt and Stable Diffusion, we analyze the decay in the proportion of T-shirts and memorized images through sampling the model. Moreover, we also use the exact likelihood computation (Song et al., 2021b) that allows us to directly calculate the negative log likelihood (NLL) of the datapoint to unlearn. More details on the experimental setup and resources used are provided in Appendix B.

The objective across all 3 sets of experiments is to establish the Pareto frontier between model quality and unlearning strength.

To ensure stability of our SISS method, we adjust the superfactor s in Eq. 8 so that the gradient norm of the second NegGrad term responsible for unlearning is fixed to around 10% of the gradient norm of the first naive deletion term responsible for ensuring the model retains $X \setminus A$. This helps to control the second term's magnitude which can suffer from an exploding gradient. In Appendix A.3, we prove that for small step sizes, this gradient clipping adjustment reduces the naive deletion loss, thereby preserving model quality.

5.1 CelebA-HQ

The CelebA-HQ dataset consists of 30000 high-quality celebrity faces (Karras et al., 2018). We use the unconditional DDPM trained by Ho et al. (2020) as our pretrained model. 6 celebrity faces were randomly selected to separately unlearn across all 7 methods. We found that injecting noise to timestep $t = 250$ on the celebrity face to be unlearned and denoising both before and after unlearning had a significant difference in similarity to the original face. Figure 2 illustrates this procedure and the quantification of similarity through the SSCD metric. The decrease in SSCD by over 50% is observed visually in Figure 3 where SISS ($\lambda = 0.5$) and SISS (No IS) guide the model away from the celebrity to be unlearned. Guiding the model away from generating the celebrity face even with

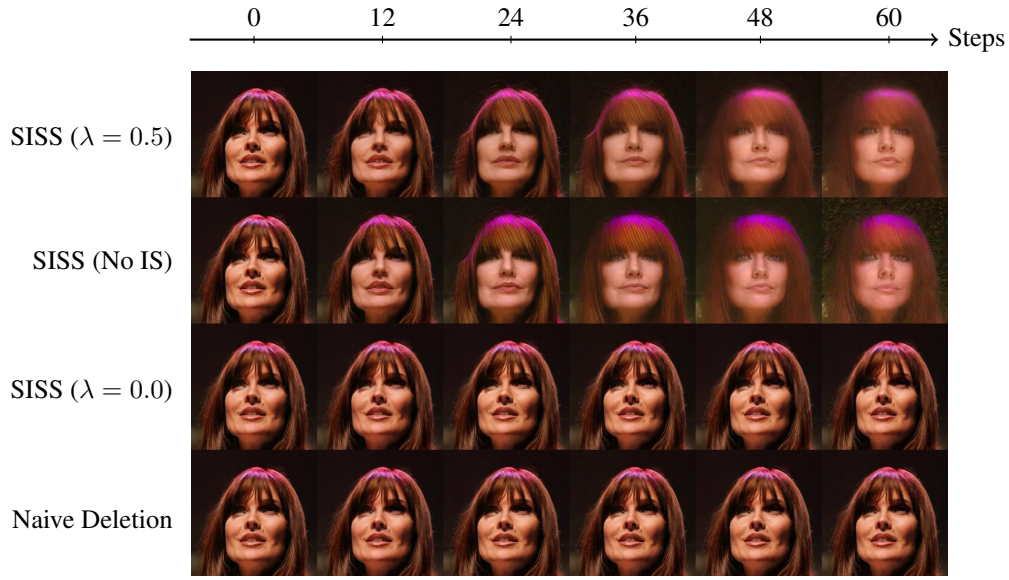


Figure 3: Visualization of celebrity unlearning over fine-tuning steps on quality-preserving methods. The images shown are made by applying noise to the original face and denoising as explained in Figure 2. Only SISS ($\lambda = 0.5$) and SISS (No IS) demonstrate the ability to guide the model away from generating the celebrity face.

Table 1: CelebA-HQ Unlearning Metrics. All methods were run for 60 fine-tuning steps on 6 separate celebrity faces. Methods in blue preserve the model quality, while the methods in red significantly decrease model quality. Only SISS ($\lambda = 0.5$) and SISS (No IS) are able to preserve model quality and unlearn the celebrity faces simultaneously relative to the pretrained model.

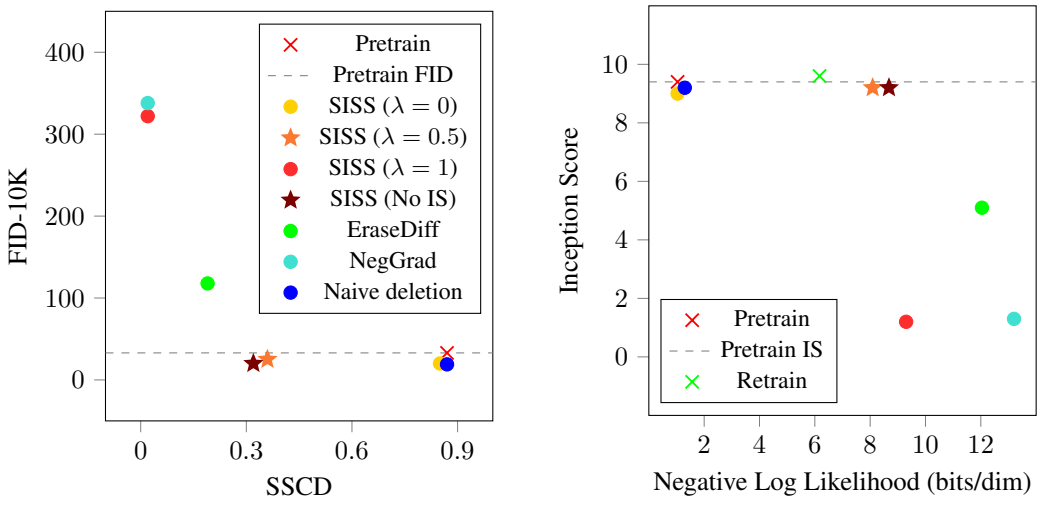
Method	**FID** ↓	**NLL** ↑	**SSCD** ↓
Pre-trained	30.3	1.257	0.87
Naive deletion	19.6	1.240	0.87
SISS ($\lambda = 0.0$)	20.1	1.241	0.87
SISS ($\boldsymbol{\lambda = 0.5}$)	25.1	1.442	0.36
SISS (No IS)	20.1	1.592	0.32
EraseDiff	117.8	4.445	0.19
SISS ($\lambda = 1.0$)	327.8	6.182	0.02
NegGrad	334.3	6.844	0.02

strong signal from timestep $t = 250$ indicates that the model is no longer incentivized to generate the face, especially at inference-time which starts from pure noise.

The SSCD, NLL, and FID unlearning and quality metrics averaged across faces are provided in Table 1. Figure 4a highlights the Pareto optimality of SISS ($\lambda = 0.5$) and SISS (No IS) along the FID and SSCD dimensions. All other methods either significantly increased the model FID or left the SSCD unaffected. In addition, we found that SISS ($\lambda = 0.5$) maintained high quality when unlearning 50 celebrity faces sequentially for 60 steps each with a final FID of 20.3, suggesting model stability over time.

5.2 MNIST T-Shirt

We train an unconditional DDPM on the MNIST dataset augmented with a T-shirt from Fashion-MNIST at a rate of 1% (Xiao et al., 2017). This experiment serves as a toy setting of analyzing the unlearning behavior of a single data point. After training, we found that the model generated the T-shirt at a rate $p = 0.74\%$ with a 95% confidence interval of $(0.68\%, 0.81\%)$. Table 2 highlights



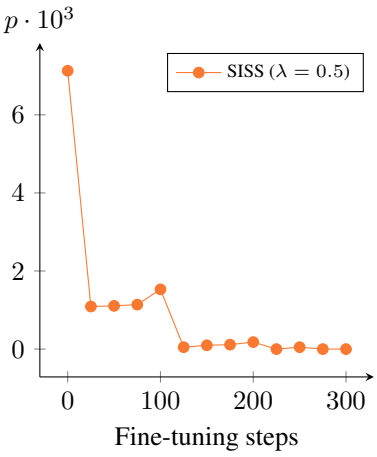
(a) Quality vs. Unlearning on CelebA-HQ.

(b) Quality vs. Unlearning on MNIST T-Shirt.

Figure 4: On both datasets, the only Pareto improvements over the pretrained model are given by SISS ($\lambda = 0.5$) and SISS (No IS). Remarkably, on MNIST T-Shirt, the two methods are Pareto improvements over the retrained model as well.

Table 2: MNIST T-Shirt Unlearning Metrics. p represents the fraction of T-shirts observed from sampling 30720 images. Methods in blue preserve model quality, while methods in red significantly decrease model quality. All numbers are averaged across 5 seeds for each method. We illustrate the decay in p as unlearning progresses for SISS ($\lambda = 0.5$).

Method	**Steps**	**IS** ↑	**NLL** ↑	p ↓
Pre-trained	117500	9.6	1.00	0.74%
Retrain	117500	9.6	6.17	0%
Naive deletion	300	9.5	1.19	0.04%
SISS ($\lambda = 0.0$)	300	9.4	1.04	0.003%
SISS ($\lambda = 0.5$)	300	9.2	8.09	0%
SISS (No IS)	300	9.2	8.68	0%
EraseDiff	100	5.1	12.04	N/A
SISS ($\lambda = 1.0$)	100	1.2	9.30	N/A
NegGrad	100	1.3	15.79	N/A

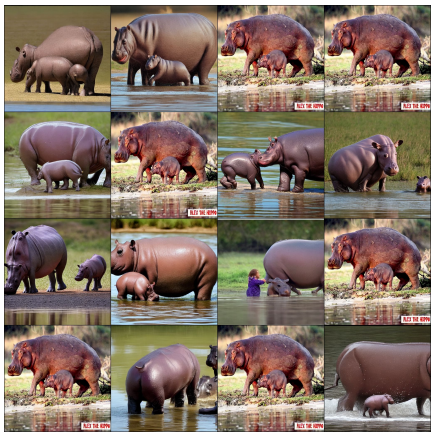


the rate of T-shirts after unlearning, showing that while naive deletion and SISS ($\lambda = 0$) significantly reduce the rate, only SISS ($\lambda = 0.5$) and SISS (No IS) are able to reach 0%.

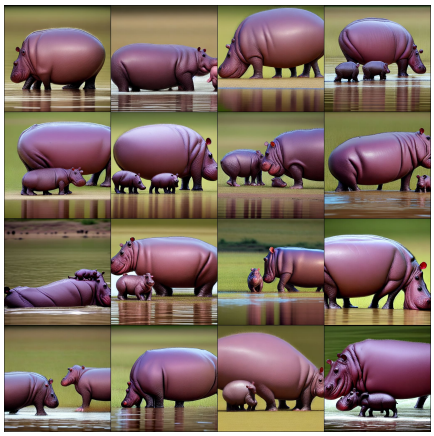
Furthermore, Figure 4b highlights the Pareto optimality of SISS ($\lambda = 0.5$) and SISS (No IS) with respect to Inception Score and NLL even when including retraining. Much like the CelebA-HQ results, all other methods either significantly decreased the Inception Score or did not change the T-shirt's NLL except, of course, retraining.

5.3 Stable Diffusion

We curate a set of 45 prompts that induce memorization on Stable Diffusion v1.4 drawn from Webster (2023). The objective is to unlearn the memorized training image from LAION corresponding to



Before Unlearning: 6 memorized samples.



After Unlearning: 0 memorized samples.

Figure 5: Visualization of memorization mitigation on Stable Diffusion v1.4 using SISS ($\lambda = 0.5$). The number of memorized samples decreases from 6 to 0 on the partially-memorized prompt "Mothers influence on "her young hippo." Note the two apostrophes in red were purposefully inserted to turn the fully-memorized prompt into a partially-memorized prompt (see Section 5.3 for details).

each prompt. Stable Diffusion is a text-conditional model; however, by keeping the prompt fixed, it can be treated as an unconditional model which is where we perform unlearning.

SISS requires a set of training examples that include both unlearning set and non-unlearning set members. Applying it to memorization mitigation on Stable Diffusion requires addressing two fundamental issues: the lack of relevant training examples and the strong memorization of prompts.

Lack of training examples. For a given memorized prompt, there is only one corresponding training LAION image. However, our method relies on having access to a dataset $X \setminus A$. We instead synthetically generate this dataset by sampling 128 images for each prompt and using a k-means classifier for labelling each image as memorized (A) or not ($X \setminus A$).

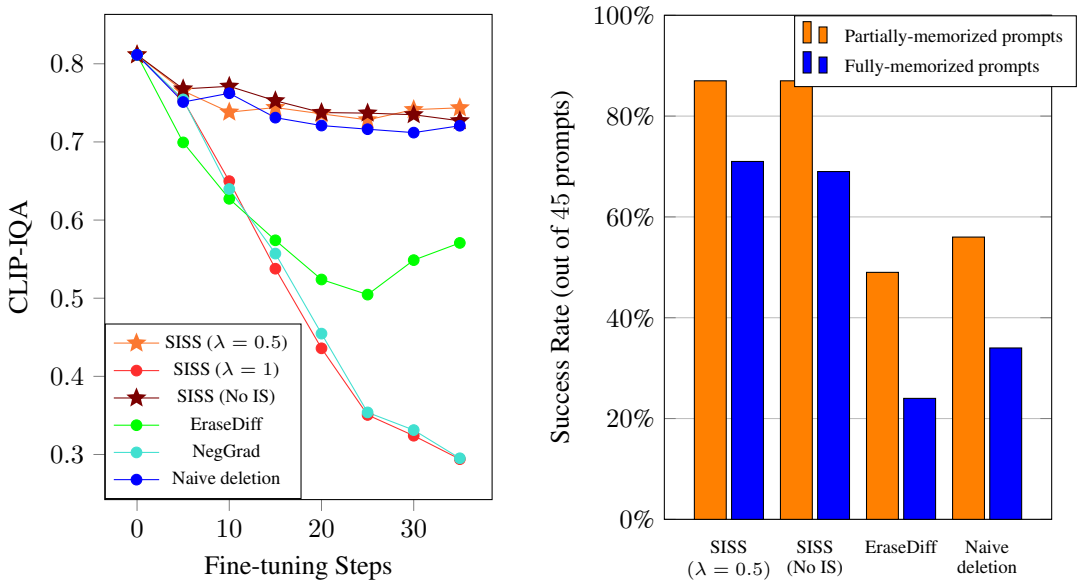
Strong memorization of prompts. Many of the prompts sourced from Webster (2023) are *fully-memorized*, i.e. our synthetically-generated dataset from this prompt would exclusively contain examples of memorized images. Inspired by the prompt modification results of Somepalli et al. (2023b), we manually delete and add tokens to obtain a *partially-memorized* prompt that generates a greater frequency of non-memorized images on each of our 45 prompts without fully mitigating memorization. The caption of Figure 5 shows an example of a partially-memorized prompt where two apostrophes were inserted to introduce sample diversity. For each prompt, we perform deletion fine-tuning on the synthetic dataset of its modified version.

We note that SISS ($\lambda = 0$) had numerical instability issues on Stable Diffusion because the NegGrad term is often very small, causing extremely large scaling factors. Thus, we excluded it from this experiment since it would be equivalent to naive deletion if scaling were disabled. Figure 6a shows that only SISS ($\lambda = 0.5$), SISS (No IS), and naive deletion are able to maintain high model quality as deletion fine-tuning occurs.

With respect to unlearning strength, Figure 6b illustrates that SISS ($\lambda = 0.5$) and SISS (No IS) are more successful in unlearning on the partially-memorized prompts than EraseDiff and naive deletion where success is the combination of reaching 0 out of 16 memorized samples and maintaining a CLIP-IQA of at least 0.35 throughout deletion fine-tuning. In addition, the two SISS methods exhibit better unlearning generalization to the fully-memorized prompts, suggesting that the model updates done with the partially-memorized prompt extend to the fully-memorized prompt in latent space.

6 Conclusion

Prior methods in diffusion unlearning have been focused on class and concept unlearning. We introduce SISS, a novel method for data unlearning in diffusion models that utilizes importance



(a) Quality as unlearning progresses.

(b) Success rate among quality-preserving methods.

Figure 6: Stable Diffusion Quality and Unlearning Results. SISS ($\lambda = 0.5$) and SISS (No IS) preserve model quality as strongly as naive deletion. EraseDiff has a moderately negative impact on quality, while the other methods significantly degrade quality. Naive deletion and EraseDiff have noticably poorer success when compared to our SISS methods and do not generalize well to the fully-memorized prompts. Success is defined as a run having no memorized image outputted in 16 samples at the end, and the average CLIP-IQA score being at least 0.35 throughout the run.

sampling for computational efficiency. Our method is able to effectively unlearn training datapoints while maintaining model quality. It exhibits Pareto optimality on multiple datasets across the quality and unlearning dimensions as well as strong memorization mitigation performance on text-conditional models such as Stable Diffusion. In the future, we hope to analyze the unlearning generalization across prompts in more detail and find a way around the "prompt modification" step that is hard to automate. Additional future directions include analyzing data unlearning on other data modalities such as audio and video. To finish, we remark that while our method may be successful in unlearning, it may not be enough legally for copyrighted data since that data is a part of the unlearning process itself.

Acknowledgments

We would like to thank Dongjun Kim for his detailed review of this manuscript and insightful discussions on the behavior of SISS.

References

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019. URL https://doi.org/10.1109/SP40001.2021.00019.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*, pp. 463–480. IEEE Computer Society, 2015. doi: 10.1109/SP.2015.35. URL https://doi.org/10.1109/SP.2015.35.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In Joseph A. Calandrino and Carmela Troncoso (eds.), *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pp. 5253–5270. USENIX Association, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Thomas Cilloni, Charles Fleming, and Charles Walter. Privacy threats in stable diffusion models, 2023. URL https://arxiv.org/abs/2311.09355.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gn0mIhQGNM.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2426–2436. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00230. URL https://doi.org/10.1109/ICCV51070.2023.00230.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 5099–5108. IEEE, 2024. doi: 10.1109/WACV57701.2024.00503. URL https://doi.org/10.1109/WACV57701.2024.00503.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3513–3526, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cb79f8fa58b91d3af6c9c991f63962d3-Abstract.html.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9301–9309. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00932. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Golatkar_Eternal_Sunshine_of_the_Spotless_Net_Selective_Forgetting_in_Deep_CVPR_2020_paper.html.

Aditya Golatkar, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Training data protection with compositional diffusion models, 2024. URL https://arxiv.org/abs/2308.01937.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/376276a95781fa17c177b1ccdd0a03ac-Abstract-Conference.html.

Tim Hesterberg. Weighted Average Importance Sampling and Defensive Mixture Distributions. *Technometrics*, 37(2):185–194, 1995. ISSN 0040-1706. doi: 10.2307/1269620. URL https://www.jstor.org/stable/1269620. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. Who's in and who's out? a case study of multimodal clip-filtering in datacomp, 2024. URL https://arxiv.org/abs/2405.08209.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2008–2016. PMLR, 2021. URL http://proceedings.mlr.press/v130/izzo21a.html.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Zhifeng Kong and Scott Alfeld. Approximate data deletion in generative models. In Kobi Gal, Ann Nowé, Grzegorz J. Nalepa, Roy Fairstein, and Roxana Radulescu (eds.), *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pp. 1288–1323. IOS Press, 2023. doi: 10.3233/FAIA230407. URL https://doi.org/10.3233/FAIA230407.

Zhifeng Kong and Kamalika Chaudhuri. Data redaction from conditional generative models. In *IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2024, Toronto, ON, Canada, April 9-11, 2024*, pp. 569–591. IEEE, 2024. doi: 10.1109/SATML59370.2024.00035. URL https://doi.org/10.1109/SaTML59370.2024.00035.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 22634–22645. IEEE, 2023. doi: 10.1109/ICCV51070.2023.02074. URL https://doi.org/10.1109/ICCV51070.2023.02074.

Guihong Li, Hsiang Hsu, Chun-Fu Chen, and Radu Marculescu. Machine unlearning for image-to-image generative models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=9hjVoPWPnh.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models, 2024. URL https://arxiv.org/abs/2402.08787.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b8a6550662b363eb34145965d64d0cfb-Abstract.html.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 14512–14522. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01413. URL https://doi.org/10.1109/CVPR52688.2022.01413.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 22522–22531. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02157. URL https://doi.org/10.1109/CVPR52729.2023.02157.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy, 2024. URL https://arxiv.org/abs/2305.06360.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 6048–6058. IEEE, 2023a. doi: 10.1109/CVPR52729.2023.00586. URL https://doi.org/10.1109/CVPR52729.2023.00586.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/9521b6e7f33e039e7d92e23f5e37bbf4-Abstract-Conference.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Ayush Kumar Tarun, Vikram Singh Chundawat, Murari Mandal, and Mohan S. Kankanhalli. Deep regression unlearning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33921–33939. PMLR, 2023. URL https://proceedings.mlr.press/v202/tarun23a.html.

David Thiel. Identifying and eliminating csam in generative ml training data and models. Technical report, Technical Report. Stanford University, Palo Alto, CA. https://purl. stanford . . . , 2023.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 2555–2563. AAAI Press, 2023. doi: 10.1609/AAAI.V37I2.25353. URL https://doi.org/10.1609/aaai.v37i2.25353.

Ryan Webster. A reproducible extraction of training images from diffusion models, 2023. URL https://arxiv.org/abs/2305.08694.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=84n3UwkH7b.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasediff: Erasing data influence in diffusion models, 2024. URL https://arxiv.org/abs/2401.05779.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning, 2024. URL https://arxiv.org/abs/2310.10683.

Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2303.17591.

Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, and Sijia Liu. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models, 2024. URL https://arxiv.org/abs/2402.11846.

A SISS MATH

A.1 STABILITY ANALYSIS AND INTERPRETATION OF SISS

Recall that the weighted loss $\ell_{s,\lambda}(\theta)$ is

$$\mathbb{E}_{p_X(x)}\mathbb{E}_{p_A(a)}\mathbb{E}_{q_\lambda(m_t|x,a)}\left[\frac{n}{n-k}\frac{q\left(m_t|x\right)}{(1-\lambda)q\left(m_t|x\right)+\lambda q\left(m_t|a\right)}\left\|\frac{m_t-\gamma_t x}{\sigma_t}-\epsilon_\theta(m_t,t)\right\|_2^2\right.$$
$$\left.-(1+s)\frac{k}{n-k}\frac{q\left(m_t|a\right)}{(1-\lambda)q\left(m_t|x\right)+\lambda q\left(m_t|a\right)}\left\|\frac{m_t-\gamma_t a}{\sigma_t}-\epsilon_\theta(m_t,t)\right\|_2^2\right].$$

An advantage of sampling from the defensive mixture distribution q_λ is that the importance weights of the noise norms are bounded for $0 < \lambda < 1$

$$0 \leq \frac{q\left(m_t|x\right)}{(1-\lambda)q\left(m_t|x\right)+\lambda q\left(m_t|a\right)} \leq \frac{1}{1-\lambda} \tag{9}$$

$$0 \leq \frac{q\left(m_t|a\right)}{(1-\lambda)q\left(m_t|x\right)+\lambda q\left(m_t|a\right)} \leq \frac{1}{\lambda}, \tag{10}$$

ensuring greater numerical stability during deletion fine-tuning. The choice of writing the superfactor as $1+s$ allows us to rewrite the outermost expectation to sample from $p_{X\setminus A}$ instead of p_X:

$$\begin{aligned}
\ell_{s,\lambda}(\theta) &= \sum_{x\in X}\frac{1}{n-k}\mathbb{E}_{q(x_t|x)}\left\|\epsilon-\epsilon_\theta(x_t,t)\right\|_2^2 && (11)\\
&\quad-(1+s)\sum_{a\in A}\frac{1}{n-k}\mathbb{E}_{q(a_t|a)}\left\|\epsilon-\epsilon_\theta(a_t,t)\right\|_2^2\\
&= \sum_{x\in X\setminus A}\frac{1}{n-k}\mathbb{E}_{q(x_t|x)}\left\|\epsilon-\epsilon_\theta(x_t,t)\right\|_2^2 && (12)\\
&\quad-s\sum_{a\in A}\frac{1}{n-k}\mathbb{E}_{q(a_t|a)}\left\|\epsilon-\epsilon_\theta(a_t,t)\right\|_2^2\\
&= \mathbb{E}_{p_{X\setminus A}(x)}\mathbb{E}_{q(x_t|x)}\left\|\epsilon-\epsilon_\theta(x_t,t)\right\|_2^2 && (13)\\
&\quad-s\frac{k}{n-k}\mathbb{E}_{p_A(a)}\mathbb{E}_{q(a_t|a)}\left\|\epsilon-\epsilon_\theta(a_t,t)\right\|_2^2\\
&= \mathbb{E}_{p_{X\setminus A}(x)}\mathbb{E}_{q(x_t|x)}\left\|\frac{x_t-\gamma_t x}{\sigma_t}-\epsilon_\theta(x_t,t)\right\|_2^2 && (14)\\
&\quad-s\frac{k}{n-k}\mathbb{E}_{p_A(a)}\mathbb{E}_{q(a_t|a)}\left\|\frac{a_t-\gamma_t a}{\sigma_t}-\epsilon_\theta(a_t,t)\right\|_2^2\\
&= \mathbb{E}_{p_{X\setminus A}(x)}\mathbb{E}_{p_A(a)}\mathbb{E}_{q_\lambda(m_t|x,a)} && (15)\\
&\quad\left[\frac{q\left(m_t|x\right)}{(1-\lambda)q\left(m_t|x\right)+\lambda q\left(m_t|a\right)}\left\|\frac{m_t-\gamma_t x}{\sigma_t}-\epsilon_\theta(m_t,t)\right\|_2^2\right.\\
&\quad\left.-s\frac{k}{n-k}\frac{q\left(m_t|a\right)}{(1-\lambda)q\left(m_t|x\right)+\lambda q\left(m_t|a\right)}\left\|\frac{m_t-\gamma_t a}{\sigma_t}-\epsilon_\theta(m_t,t)\right\|_2^2\right].
\end{aligned}$$

From Eq. 12, we see that the weighted loss $\ell_{s,\lambda}(\theta)$ has two separate terms: the naive deletion loss $L_{X\setminus A}(\theta)$ and a term proportional to the NegGrad loss that discourages the generation of A's members. Thus, Eq. 12 is equivalent to

$$L_{X\setminus A}(\theta) - s\frac{k}{n-k}L_A(\theta)$$

where the superfactor s controls the weight of the NegGrad term L_A. As a result, we can directly view $\ell_{s,\lambda}$ as interpolating between naive deletion and NegGrad with s controlling the strength of NegGrad.

Notice that if t is small then $q(m_t|x) \gg q(m_t|a)$ if m_t is sampled from $q(\,\cdot\,|x)$ and $q(m_t|x) \ll q(m_t|a)$ if m_t is sampled from $q(\,\cdot\,|a)$. If $\lambda = 0$, we know $q(m_t|x) \gg q(m_t|a)$, making the importance ratio of the first term in the SISS loss equal to 1, and the second importance ratio equal to 0. Hence, SISS with $\lambda = 0$ will be equivalent to naive deletion. Similarly, if $\lambda = 1$, $q(m_t|x) \ll q(m_t|a)$ and the second importance ratio will dominate, which is equivalent to NegGrad.

A.2 PROOF OF LEMMA 1

Lemma 1 Restated. Gradient estimators of $\ell_\lambda(\theta)$ and $L_{X\setminus A}(\theta)$ are the same in expectation. That is, in expectation, Monte Carlo estimates of

$$\begin{aligned}\nabla_\theta \ell_\lambda(\theta) = &\mathbb{E}_{p_X(x)}\mathbb{E}_{p_A(a)}\mathbb{E}_{q_\lambda(m_t|x,a)} \\ &\left[\nabla_\theta\left(\frac{n}{n-k}\frac{q\left(m_t|x\right)}{(1-\lambda)q\left(m_t|x\right)+\lambda q\left(m_t|a\right)}\left\|\frac{m_t-\gamma_t x}{\sigma_t}-\epsilon_\theta(m_t,t)\right\|_2^2\right.\right. \\ &\left.\left.-\frac{k}{n-k}\frac{q\left(m_t|a\right)}{(1-\lambda)q\left(m_t|x\right)+\lambda q\left(m_t|a\right)}\left\|\frac{m_t-\gamma_t a}{\sigma_t}-\epsilon_\theta(m_t,t)\right\|_2^2\right)\right]\end{aligned} \tag{16}$$

and Monte Carlo estimates of

$$\nabla_\theta L_{X\setminus A}(\theta) = \mathbb{E}_{p_{X\setminus A}(x)}\mathbb{E}_{q(x_t|x)}\left[\nabla_\theta \left\|\epsilon-\epsilon_\theta(x_t,t)\right\|_2^2\right] \tag{17}$$

are equal.

Proof. Note that Eqs. 16 and 17 are direct consequences of the linearity of expectation which allows us to take the gradient with respect to θ inside the expectation operations. The equivalence of the two loss functions (Eq. 7) implies that

$$\nabla_\theta \ell_\lambda(\theta) = \nabla_\theta L_{X\setminus A}(\theta). \tag{18}$$

When combined with Eqs. 16 and 17, we see that the Monte Carlo gradient estimators are the same in expectation. □

A.3 GRADIENT CLIPPING

Recall that we adjust the superfactor s in Eq. 8 so that the gradient norm of the second NegGrad term responsible for unlearning is fixed to around 10% of the gradient norm of the first naive deletion term responsible for ensuring the model retains $X \setminus A$. We show that for small step sizes this gradient clipping adjustment reduces the naive deletion loss, thus preserving the quality of the model. To prove this, we start with a variant of the classic descent lemma.

Descent lemma under small perturbations. Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, and that its gradient is Lipschitz continuous with constant $L > 0$, i.e., we have that $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ for any x, y. Then, one step of gradient descent from $x \in \mathbb{R}^n$ with step size t and perturbed update

$$x' = x - t(\nabla f(x) + v)$$

satisifes the improvement bound

$$f(x') \leq f(x) - (1 - p - \epsilon)t\|\nabla f(x)\|^2.$$

We assume

$$\epsilon = \frac{1}{2}Lt + Ltp + \frac{1}{2}Ltp^2,$$

and $v \in \mathbb{R}^n$ is an arbitrary vector satisfying

$$\|v\| \leq p\|\nabla f(x)\|$$

for a proportion p.

Proof. ∇f being L-Lipschitz implies

$$f(x') \leq f(x) + \nabla f(x)^\top (x' - x) + \frac{L}{2}\|x' - x\|^2.$$

Plugging in our update equation and repeatedly applying Cauchy-Schwarz gives

$$\begin{aligned} f(x') &\leq f(x) + \nabla f(x)^\top (x' - x) + \frac{L}{2}\|x' - x\|^2 \\ &= f(x) + \nabla f(x)^\top (-t\nabla f(x) - tv) + \frac{L}{2}\| - t\nabla f(x) - tv\|^2 \end{aligned}$$

$$
\begin{aligned}
&= f(x) - t\|\nabla f(x)\|^2 - t\nabla f(x)^\top v + \frac{Lt^2}{2}\langle \nabla f(x) + v, \nabla f(x) + v\rangle \\
&= f(x) - t\|\nabla f(x)\|^2 - t\nabla f(x)^\top v + \frac{Lt^2}{2}\left(\|\nabla f(x)\|^2 + 2\nabla f(x)^\top v + \|v\|^2\right) \\
&\leq f(x) - t\|\nabla f(x)\|^2 + t\|\nabla f(x)\|\|v\| + \frac{Lt^2}{2}\left(\|\nabla f(x)\|^2 + 2\|\nabla f(x)\|\|v\| + \|v\|^2\right) \\
&\leq f(x) - t\|\nabla f(x)\|^2 + tp\|\nabla f(x)\|^2 + \frac{Lt^2}{2}\left(\|\nabla f(x)\|^2 + 2p\|\nabla f(x)\|^2 + p^2\|\nabla f(x)\|^2\right) \\
&= f(x) - t\left(1 - p - \frac{Lt}{2} - Ltp - \frac{Ltp^2}{2}\right)\|\nabla f(x)\|^2.
\end{aligned}
$$

Hence, setting

$$\epsilon = \frac{1}{2}Lt + Ltp + \frac{1}{2}Ltp^2$$

gives the desired result where we note that $t \to 0$ implies $\epsilon \to 0$. Thus, ϵ can be made arbitrary small by selecting a small step size t. □

Notice that the classic descent lemma corresponds to the special case of $p = 0$. Practically, as long as $p < 1$, we can choose t small so that we are guaranteed improvement on each step of gradient descent. Standard results show that this implies gradient descent will eventually converge to a point where $\|\nabla f(x)\| < \delta$ in $\mathcal{O}(\frac{1}{\delta})$ iterations.

In the data unlearning context, set f to be the naive deletion objective and pick v to be a rescaled version of the NegGrad objective's gradient. Our perturbed descent lemma shows that for appropriate step size we will be no worse off in naive deletion performance, implying that model quality will be preserved. It is not theoretically clear that choosing v to be NegGrad's gradient will result in unlearning. However, our results in Section 5 found this to be empirically true.

B Experimental Setup

All diffusion models were trained and fine-tuned using the Hugging Face `diffusers` package along with the Adam optimizer (Kingma & Ba, 2015). YAML configuration files with all run settings can be found in the `config/` directory of our codebase.

The CelebA-HQ experiments used a pretrained checkpoint from Ho et al. (2020) hosted at `https://huggingface.co/google/ddpm-celebahq-256`. Our pretrain and retrain unconditional MNIST T-Shirt DDPMs were trained for 250 epochs with a batch size of 128 images and a learning rate of $1e-4$ with cosine decay. Both models used the same DDPM sampler at inference with 50 backwards steps. For the Stable Diffusion experiments, we used version 1.4 hosted at `https://huggingface.co/CompVis/stable-diffusion-v1-4` as our pretrained checkpoint with 50-step DDIM as the sampler (Song et al., 2021a). The models for all 3 sets of experiments use a U-Net backbone.

Deletion fine-tuning experiments were run starting from the EMA versions of the trained MNIST T-Shirt DDPMs as well as the pre-trained Stable Diffusion model. For MNIST T-Shirt, the same hyperparameters were kept from pretraining to run fine-tuning. In the case of CelebA-HQ and Stable Diffusion, we did not perform the pretraining and chose a batch size of 64 and 16 images with a learning rate of $5e-6$ and $1e-5$, respectively.

While most individual experiments were not very computationally expensive (roughly half an hour on average), sweeping across all different baselines and and mixture parameters λ totaled to over 500 runs. To streamline this process, a cluster of 8 NVIDIA H100 GPUs were used to execute large numbers of runs in parallel. In addition, an g5.xlarge instance with an NVIDIA A10G GPU on AWS, a personal home computer with an NVIDIA RTX 3090, and a cluster of 3 NVIDIA A4000 GPUs were the primary code development environments.