EFFICIENT GLOBAL DATA ATTRIBUTION FOR DIFFUSION MODELS

Chris Lin,* **Mingyu Lu**,* **Su-In Lee**
Paul G. Allen School of Computer Science & Engineering
University of Washington
{clin25,mingyulu,suinlee}@cs.washington.edu

ABSTRACT

With the widespread usage of diffusion models, effective data attribution is needed to ensure fair acknowledgment for contributors of high-quality training samples, and to identify potential sources of harmful content. In this early work, we introduce a novel framework tailored to removal-based data attribution for diffusion models, leveraging *sparsified unlearning*. This approach significantly improves the computational scalability and effectiveness of removal-based data attribution. In our experiments, we attribute diffusion model FID back to CIFAR-10 training images with datamodel attributions, showing better linear datamodeling score (LDS) than datamodel attributions based on naive retraining.

1 INTRODUCTION

Diffusion models have demonstrated impressive performance on image generation (Ho et al., 2020; Song et al., 2020b), with models such as Dall·E 2 (Ramesh et al., 2022) and Stable Diffusion (Rombach et al., 2022) showing powerful utilities and enabling downstream applications via customization (Hu et al., 2021; Ruiz et al., 2023). The training data for large-scale diffusion models are often scraped from the internet (Schuhmann et al., 2022), raising concerns such as copyright attribution, harmful content generation (Birhane et al., 2021), and bias propagation (Luccioni et al., 2023). *Data attribution*, which aims to trace machine learning model behaviors back to training data, has the potential to address these issues. Indeed, in the context of supervised learning, data attribution methods have already been used to valuate data (Ghorbani & Zou, 2019), identify adversarial samples (Koh & Liang, 2017), and discover similar data points (Ilyas et al., 2022).

Some recent work has developed data attribution methods for diffusion models (Dai & Gifford, 2023; Wang et al., 2023; Georgiev et al., 2023; Zheng et al., 2023). These methods focus on *local* model behaviors related to a given generated image. However, some use cases require understanding data's influence on *global* model behaviors, which are related to the overall generative distribution of a diffusion model. For example, understanding which data have a negative impact on the Fréchet Inception Distance (FID) (Heusel et al., 2017) can identify noisy samples. As another example, the demographic diversity of generated images can be considered a global behavior, and attributing this behavior to training data can help us find the sources of social biases in diffusion models (Luccioni et al., 2023). Hence, here we focus on global data attribution for diffusion models.

Many data attribution methods can be categorized as (1) *gradient-based* methods that compute loss gradients with training samples as inputs (Koh & Liang, 2017; Yeh et al., 2018), (2) *removal-based* methods that retrain models on training subsets to estimate each sample's influence (Ghorbani & Zou, 2019; Ilyas et al., 2022), or (3) a combination of both (Park et al., 2023). Global model behaviors may not be differentiable for diffusion models (e.g., proportion of generated images corresponding to a demographic group). Therefore, methods that require gradient computations are not applicable, while removal-based methods are. However, removal-based data attribution requires models be retrained many times with different training subsets. Although it has been shown possible to retrain tens of thousands of models to measure data importance for image classifiers (Ilyas et al., 2022), diffusion models generally require longer training time and make naive retraining computationally infeasible. In this work, we propose to make global data attribution computationally efficient

*Equal contribution.

for diffusion models, by approximating retraining with *sparsified unlearning*. Specifically, we adapt the "prune first, then unlearn" paradigm (Jia et al., 2023) to approximate diffusion model retraining and improve the performance of datamodel attributions by 10% linear modeling score (LDS) (Ilyas et al., 2022), compared to datamodel attributions with naive retraining.

2 Related Work

Diffusion models. Generally, diffusion models are trained to approximate a data distribution $q(\mathbf{x}_0)$. To perform learning, a training sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ is sequentially corrupted by additive noise (Ho et al., 2020). This procedure is called the *forward process* and defined by $q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I})$, for $t = 1, ..., T$, where $\{\beta_t\}_{t=1}^T$ corresponds to a variance schedule. Notably, the forward process allows sampling of $\mathbf{x}_t$ at any time step t from $\mathbf{x}_0$, with the closed form $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I})$, where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$. Then, a diffusion model learns to denoise $\mathbf{x}_{1:T}$, following the *reverse process* defined by $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2 \boldsymbol{I})$, where $\theta \in \mathbb{R}^d$ is the model parameters, and σ_t corresponds to some sampling schedule (Karras et al., 2022).

Instead of modeling the conditional means μ_θ, it is standard to predict the added noises with a neural network ϵ_θ using the reparameterization trick. Once a diffusion model has been trained, a new image can be generated by sampling an initial noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and iteratively applying ϵ_θ at each step $t = T, ..., 1$ for denoising. In practice, there are different design choices of the denoising process such as the inference time steps and output scaling (Song et al., 2020a;b; Karras et al., 2022).

Data attribution. The goal of data attribution is to identify important training data for a model's behaviors. Some methods are computationally efficient, requiring only the loss gradients computed at the sample to be explained and at the training samples (Koh & Liang, 2017; Yeh et al., 2018). In contrast, methods that rely on retraining models with subsets of the training data have been shown to perform better, at the cost of expensive computation required for retraining models (Ghorbani & Zou, 2019; Ilyas et al., 2022). Park et al. (2023) combine gradient-based attribution with a handful of retrained models to achieve both efficiency and performance. Because training diffusion models is resource-intensive, existing data attribution methods for diffusion models either focus on methods with gradient computations (Georgiev et al., 2023; Zheng et al., 2023) or rely on non-standard training procedures (Dai & Gifford, 2023; Wang et al., 2023). In addition, these data attribution methods focus on local model behaviors specific to a given generated image.

Machine unlearning. One goal of *machine unlearning* is to produce models that behave as if certain training data have been removed (Mercuri et al., 2022). Exact unlearning methods such as SISA apply to models that have already been trained on data subsets (Bourtoule et al., 2021), whereas approximate unlearning methods are applicable to any models (Golatkar et al., 2020; Thudi et al., 2022). For diffusion models, Dai & Gifford (2023) apply SISA to design an ensembled diffusion model that allow exact unlearning for data attribution. Recently, Jia et al. (2023) show that pruning a supervised model first before unlearning can improve unlearning efficiency and efficacy. We adapt this paradigm of sparsified unlearning to efficiently approximate diffusion model retraining.

Contributions. Our contributions include the following. (1) To our knowledge, we are the first to study global data attribution for diffusion models. (2) We propose a general framework to efficiently estimate removal-based global data attributions via sparsified unlearning. (3) We demonstrate that our framework can outperform removal-based methods that rely on naive retraining.

3 Efficient Global Data Attribution via Sparsified Unlearning

To compute global data attributions for a given diffusion model ϵ[1], we first need to define a global model behavior $\gamma : \mathcal{E} \to \mathbb{R}$, where $\mathcal{E}$ denotes the set of all possible diffusion models with the same architecture. Considering the Fréchet Inception Distance (FID) (Heusel et al., 2017) as an example, γ is a wrapper function that generates multiple images using ϵ and calculates the FID between the generated images and reference images. With a model behavior γ specified, removal-based

[1]We consider the common practice of training the noise predictor ϵ as training a diffusion model. We also drop the dependency on the model parameters θ for notational ease.

attribution methods such as Data Shapley (Ghorbani & Zou, 2019) and datamodeling (Ilyas et al., 2022) estimate the influence of training data, by computing γ with models $\{\epsilon_{S_i}\}_i$ trained on different data subsets S_i. In the datamodeling framework particularly, a linear model $g_\beta : \{0,1\}^N \to \mathbb{R}$ is fitted to predict $\{\gamma(\epsilon_{S_i})\}_i$ from training data subsets $\{S_i\}_i$ sampled from a distribution $\mathcal{D}$. Here, N denotes the total number of training data. Specifically,

$$\beta = \underset{\omega \in \mathbb{R}^{N+1}}{\arg\min}\, \mathbb{E}_{S_i \sim \mathcal{D}} \left[\mathcal{L} \left(g_\omega(\mathbf{1}_{S_i}), \gamma(\epsilon_{S_i}) \right) \right], \tag{1}$$

where $\mathbf{1}_{S_i}$ is a characteristic vector indicating the presence of each training sample in S_i, such that

$$(\mathbf{1}_{S_i})_j = \begin{cases} 1 & \text{if the } j\text{th training sample is in } S_i, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here, $\mathcal{L}$ is a regression objective (e.g., mean squared error), which can also include regularization. The fitted parameters $\beta_{1:N}$, without the intercept term β_0, are considered the attribution scores.

Attributing data contributions to diffusion models using a removal-based approach has been considered impractical, because model retraining time increases linearly with the number of sampled subsets. We propose a novel framework of **removal through sparsified unlearning**, to significantly speed up the attribution procedure. Specifically, our approach utilizes *approximate gradient unlearning* (Golatkar et al., 2020; Thudi et al., 2022), which aims to approximate a model retrained with a subset S_i. We also combine approximate unlearning with the "prune first, then unlearn" paradigm (Jia et al., 2023), leveraging model sparsity to make approximate unlearning even faster. Overall, given a trained diffusion model ϵ and instantiating with datamodel attribution, our framework consists of four steps:

1. **(Pruning)** Prune ϵ to obtain a sparser and similarly performant model $\tilde{\epsilon}$.
2. **(Unlearning with subsets)** Sample a training data subset $S_i \sim \mathcal{D}$. Obtain an unlearned model $\tilde{\epsilon}_{S_i}$ by gradient descent on the model parameters of $\tilde{\epsilon}$, with S_i coupled with the original objective used to train ϵ. Repeat this M times.
3. **(Computing model behaviors)** Compute $\gamma(\tilde{\epsilon}_{S_1}), \gamma(\tilde{\epsilon}_{S_2}), ..., \gamma(\tilde{\epsilon}_{S_M})$.
4. **(Datamodel fitting)** Fit a datamodel g_β with $\{\mathbf{1}_{S_i}\}_{i=1}^M$ as inputs and $\{\gamma(\tilde{\epsilon}_{S_i})\}_{i=1}^M$ as outputs.

4 Experiments

Diffusion model setup. We trained Denoising Diffusion Probability Models (DDPMs) following the original implementation of Ho et al. (2020) on CIFAR-10 (Krizhevsky et al., 2009) for 200,000 gradient steps, with $T = 1{,}000$. The DDIM linear scheduler with 100 time steps and $\beta_1 = 10^{-4}, \beta_T = 0.02$ was used for image generation (Song et al., 2020a).

Pruning and unlearning setup. We pruned the DDPM trained on the entire CIFAR-10 dataset using magnitude-based pruning (Han et al., 2015) for its computational efficiency, following Fang et al. (2023) with a pruning ratio of 0.3 and threshold of 0.05. Gradient descent with $4{,}000$ gradient steps was used for unlearning. All models were trained with NVIDIA RTX6000 GPUs.

Evaluating data attribution performance. Data attribution performance is measured using the linear datamodeling score (LDS) Ilyas et al. (2022), which evaluates an attribution method by comparing linearly predicted model behaviors against actual retrained model behaviors. Let $S_1, ..., S_K$ be K randomly sampled subsets of the training set, each of size $\alpha \cdot N$ for some $\alpha \in (0, 1)$. We set $\alpha = 0.5$ for our experiments. The LDS for a set of data attribution scores $\tau \in \mathbb{R}^N$ is defined as

$$LDS := \rho(\{\gamma(\epsilon_{S_i})\}_{i=1}^K, \{\mathbf{1}_{S_i}^\top \tau\}_{i=1}^K), \tag{3}$$

where ρ is the Spearman rank correlation (Spearman, 1961), and ϵ_{S_i} denotes a model retrained from scratch with the training subset S_i. In our experiments we consider the FID as the global model behavior γ.

Results. To validate our proposed framework, we first compare the similarity between 96 pairs of sparsified unlearned models and models retrained from scratch with the same training subsets. As

Table 1: Performance of diffusion models on CIFAR-10 measured by FID and computational time. FID_1 compares to the CIFAR-10 training data. FID_2 compares to generated images from models retrained from scratch.

Method	**FID_1**	**FID_2**	**Train Steps**	**Train Time (min)**
Original training	6.79	-	200k	-
Magnitude-based pruning	7.38	1.0	200k	-
Retraining from scratch (exact unlearning)	-	-	200k	1207
Sparsified unlearning (our framework)	-	1.6	4k	**18.9**

shown in Table 1, sparsified unlearning is faster than retraining from scratch by a factor of nearly 64, while maintaining a similar generative distribution (FID = 1.9). This efficiency gain suggests the potential to increase the number of sampled subsets by up to 64-fold for any removal-based data attribution method.

Subsequently, we evaluate the LDS of datamodel attributions by retraining versus sparsified unlearning under various computation budgets. Here, we consider the computationl time for one instance of retraining from scratch as a reference unit. Our results demonstrate that sparsified unlearning outperforms naive retraining with much less computing time (Fig. 1).

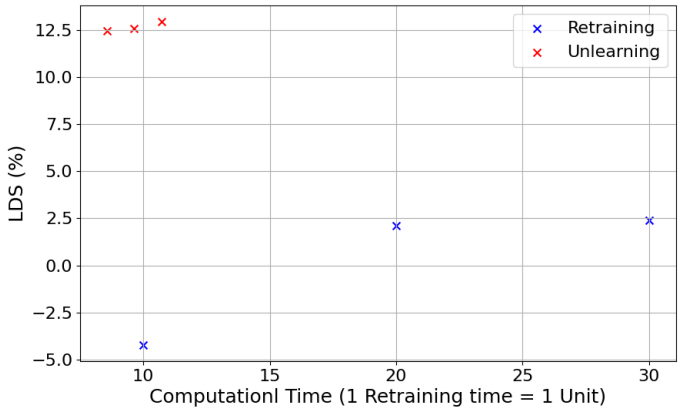

Figure 1: Comparison of retraining-based vs. unlearning-based datamodel attributions by LDS (%) and computational time.

5 Discussion

In this early work, we introduce a novel framework to efficiently estimate global data attributions for diffusion models, using sparsified unlearning as an approach to speed up removal-based data attribution. We apply our framework to attribute diffusion model FID back to training images in CIFAR-10 with datamodel attributions, showing better LDS performance than datamodel attributions based on naive retraining. There are several directions for potential improvements and future work. First, while we only focus on datamodel attributions for this work, the proposed framework can extend to other removal-based methods, such as Data Shapley (Ghorbani & Zou, 2019) and Data Banzhaf (Wang & Jia, 2023). Beyond gradient unlearning, there are opportunities to apply unlearning methods tailored for diffusion models (Heng & Soh, 2023; Gandikota et al., 2023). Another promising aspect of our framework is its flexibility for incorporating customized model behaviors of diffusion models. This is crucial for assessing the impact of individual data points on the influence of foundation generative models, with particular relevance to high-stakes scenarios concerning safety and copyright (Birhane et al., 2021).

REFERENCES

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

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

Zheng Dai and David K Gifford. Training data attribution for diffusion models. *arXiv preprint arXiv:2306.02174*, 2023.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. *arXiv preprint arXiv:2305.10924*, 2023.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023.

Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *arXiv preprint arXiv:2305.10120*, 2023.

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. *Advances in neural information processing systems*, 30, 2017.

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

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

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In *Annual Conference on Neural Information Processing Systems*, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

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

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

Salvatore Mercuri, Raad Khraishi, Ramin Okhrati, Devesh Batra, Conor Hamill, Taha Ghasempour, and Andrew Nowlan. An introduction to machine unlearning. *arXiv preprint arXiv:2209.00939*, 2022.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Charles Spearman. The proof and measurement of association between two things. 1961.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Jiachen T Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 6388–6421. PMLR, 2023.

Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. *arXiv preprint arXiv:2306.09345*, 2023.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Xiaosen Zheng, Tianyu Pang, Chao Du, Jing Jiang, and Min Lin. Intriguing properties of data attribution on diffusion models. *arXiv preprint arXiv:2311.00500*, 2023.