Understanding Data Replication in Diffusion Models

Gowthami Somepalli [1] **Vasu Singla** [1] **Micah Goldblum** [2] **Jonas Geiping** [1] **Tom Goldstein** [1]

Abstract

Images generated by diffusion models like Stable Diffusion are increasingly widespread. Recent works and even lawsuits have shown that these models are prone to replicating their training data, unbeknownst to the user. In this paper, we first analyze this memorization problem in text-to-image diffusion models. Contrary to the prevailing belief attributing content replication solely to duplicated images in the training set, our findings highlight the equally significant role of text conditioning in this phenomenon. Specifically, we observe that the combination of image and caption duplication contributes to the memorization of training data, while the sole duplication of images either fails to contribute or even diminishes the occurrence of memorization in the examined cases.

1. Introduction

A major hazard of diffusion models is their ability to produce images that replicate their training data, often without warning to the user (Somepalli et al., 2022; Carlini et al., 2023). Despite their risk of breaching privacy, data ownership, and copyright laws, diffusion models have been deployed at the commercial scale by subscription-based companies like *midjourney*, and more recently as offerings within search engines like *bing* and *bard*. Currently, a number of ongoing lawsuits (Saveri and Butterick, 2023) are attempting to determine in what sense companies providing image generation systems can be held liable for replications of existing images.

In this work, we take a deep dive into the causes of memorization for modern diffusion models. Prior work has largely focused on the role of duplicate images in the training set. While this certainly plays a role, we find that image duplication alone cannot explain all of the replication behavior we see at test time. Our experiments reveal that text conditioning plays a major role in data replication. We also observe that the joint duplication of both images and captions plays a significantly greater role in inducing training data memorization compared to solely duplicating images. These findings emphasize the critical influence of image-caption synergy in facilitating memorization within the training process.

[1]University of Maryland, College Park [2]NYU. Correspondence to: Gowthami Somepalli <gowthami@cs.umd.edu>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

2. Related work

Memorization in generative models. Most insights on the memorization capabilities of generative models are so far empirical, as in studies by Webster et al. (2021) for GANs and a number of studies for generative language models (Carlini et al., 2022; Jagielski et al., 2022; Lee et al., 2022).

Recently, Somepalli et al. (2022) investigated data replication behaviors in modern diffusion models, finding 0.5-2% of generated images to be partial object-level duplicates of training data, findings also mirrored in Carlini et al. (2023). Yet, what mechanisms lead to memorization in diffusion models, and how they could be inhibited, remains so far uncertain aside from recent theoretical frameworks rigorously studying copyright issues for image duplication in Vyas et al. (2023).

Removing concepts from diffusion models. Mitigations deployed so far in diffusion models have focused on input filtering. For example, Stable Diffusion includes detectors that are trained to detect inappropriate generations. These detectors can also be re-purposed to prevent the generation of known copyrighted data, such as done recently in *midjourney*, which has banned its users from generating photography by artist Steve McCurry, due to copyright concerns (Chess, 2022). However, such simple filters can be easily circumvented (Rando et al., 2022; Wen et al., 2023), and these band-aid solutions do not mitigate copying behavior at large. A more promising approach deletes entire concepts from the model as in Schramowski et al. (2023) and Kumari et al. (2023), yet such approaches require a list of all concepts to be erased, and are impractical for protecting datasets with billions of diverse images covering many concepts.

3. Experimental Setup

A thorough study of replication behavior requires training many diffusion models. To keep costs tractable, we focus on experiments in which large pre-trained models are finetuned on smaller datasets. This process reflects the training of Stable Diffusion models, which are pre-trained on LAION and then finetuned in several stages on much smaller and more curated datasets, like the LAION Aesthetics split.

Datasets: We use Imagenette[1], which consists of 10 classes from Imagenet (Deng et al., 2009) as well as two randomly sampled subsets of $10,000$ images from LAION-2B (Schuhmann et al., 2022) for our experiments. The LAION subsets, which we denote as LAION-10k and LAION-100k, include captions, while the Imagenette data is uncaptioned. For some experiments, we use BLIP v1 (Li et al., 2022) to generate captions for images when needed.

Architecture & Training: We use the StableDiffusion-v2.1 checkpoint as a starting point for all experiments. Unless otherwise noted, only the U-Net (Ronneberger et al., 2015) part of the pipeline is finetuned (the text and auto-encoder/decoder components are frozen) as in the original training run, and we finetune for 100k iterations with a constant LR of $5e-6$ and 10k steps of warmup. All models are trained with batch size 16 and image resolution 256.

Metrics: We use the following metrics to study the memorization and generation quality of finetuned models. **Frechet Inception Distance (FID)** (Heusel et al., 2017) evaluates the quality and diversity of model outputs. FID measures the similarity between the distribution of generated images and the distribution of the training set using features extracted by an Inception-v3 network. A *lower* FID score indicates better image quality and diversity.

Somepalli et al. (2022) quantify copying in diffusion models using a similarity score derived from the dot product of SSCD representations (Pizzi et al., 2022) of a generated image and its top-1 match in the training data. They observe that generations with similarity scores greater than 0.5 exhibit strong visual similarities with their top-1 image and are likely to be partial object-level copies of training data.

Given a set of generated images, we define its **dataset similarity score** as the 95-percentile of its image-level similarity score distribution. Note that measuring average similarity scores over the whole set of generated images is uninformative, as we are only interested in the prevalence of replicated images, which is diluted by non-replicated samples. For this reason, we focus only on the similarity of the 5% of images with the highest scores.

[1] https://github.com/fastai/imagenette

4. Data Duplication is Not the Whole Story

Existing research hypothesizes that replication at inference time is mainly caused by duplicated data in the training set (Somepalli et al., 2022; Carlini et al., 2023; Webster et al., 2023). Meanwhile, data replication has been observed in newer models trained on de-duplicated data sets (Nichol, 2022; Mostaque, 2022). Our goal here is to quantify the extent to which images are duplicated in the LAION dataset, and understand how this impacts replication at inference time. We will see that data duplication plays a role in replication, but it cannot explain much of the replication behavior we see.

4.1. Controlled Experiments with Data Duplication

We train diffusion models with various levels of `data duplication factor (ddf)`, which represents the factor by which duplicate samples are more likely to be sampled during training. We train each model for 100k iterations and evaluate similarity and FID scores on 4000 generated samples.

Figure 1 contains results for LAION-10k and Imagenette. We observe that increased duplication in the training data tends to yield increased replication during inference. The relationship between data duplication and similarity scores is not straightforward for LAION-10k . As the duplication factor increases, similarity scores rise again until reaching a data duplication factor `ddf` of 10, after which they decrease. Regarding FID scores, we find that a certain level of data duplication contributes to improving the scores for models trained on both datasets, possibly because FID is lowest when the dataset is perfectly memorized.

Other Observations from the Literature. Somepalli et al. (2022) found that unconditional diffusion models can exhibit strong replication when datasets are small, despite these training sets containing *no* duplicated images. Clearly, *replication can happen in the absence of duplication*. As the training set sizes grow ($\sim$ 30k), the replication behaviors seen in Somepalli et al. (2022) vanish, and dataset similarity drops, even when the number of epochs is kept constant. One might expect this for large enough datasets. However, they found replication in SD v1.4 which is trained on the *much larger* LAION-2B dataset. We will see below that the trend of replication in diffusion models depends strongly on additional factors, related especially to their *conditioning*.

4.2. The Effect of Model Conditioning

To understand the interplay between model conditioning and replication, we consider four types of text conditioning:

- **Fixed caption:** All data points are assigned the same caption, `An image`.

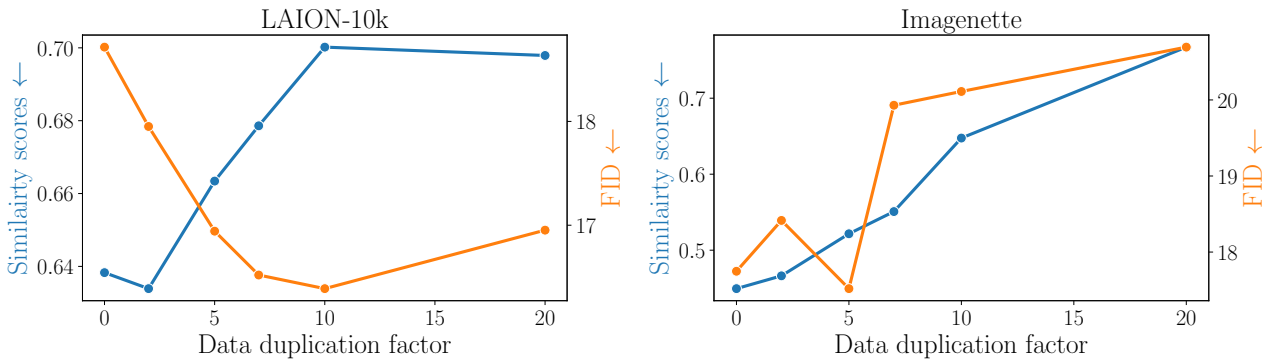

Figure 1. How does data duplication affect memorization? All models are trained with captions. On both datasets, dataset similarity increases proportionally to duplication in training data. FID score are unaffected by light duplication, but increase on higher levels as image diversity reduces.

- **Class captions:** Images are assigned a class-wise caption using the template `An image of <class-name>`.
- **Blip/Original captions:** Each point is trained on the original caption from the dataset (LAION-10k) or a new caption is generated for each image using BLIP (Li et al., 2022) (Imagenette).
- **Random captions:** A random sequence of 6 tokens is sampled from the vocabulary used to uniquely caption each image.

By varying caption scenarios, we transition from no diversity in captions ("fixed caption") case to completely unique captions with no meaningful overlap ("random caption") case. We again finetune on the Imagenette dataset, now using these caption types. As a baseline, we consider the pretrained Stable Diffusion model without any finetuning. Figure 2 (left) shows the dataset similarity among the baseline and models finetuned using the different caption types.

We observe that the finetuned models exhibit higher similarity scores compared to the baseline model. Furthermore, the level of model memorization is influenced by the type of text conditioning. The "fixed caption" models exhibit the lowest amount of memorization, while the "blip caption" models exhibit the highest. This indicates that the model is more likely to memorize images when the captions are more diverse. However, the model does not exhibit the highest level of memorization when using "random captions", meaning that captions should be correlated with image content in order to maximally help the model retrieve an image from its memory.

Training the text encoder. So far, the text encoder was frozen during finetuning. We can amplify the impact of conditioning on replication by training the text encoder during finetuning. In Figure 2 (right), we observe a notable increase in similarity scores across all conditioning cases when the text encoder is trained. This increase is particularly prominent in the cases of "blip captioning" and "random captioning". These findings support our hypothesis that the model is more inclined to remember instances when the captions associated with them are highly specific, or even unique, keys.

4.3. The Impact of Caption vs. Image Duplication

In this section, we control separately for duplication of images and duplication of their captions to better understand how the two interact.

In the case of **full duplication**, both the image and its caption are replicated multiple times in the training data. On the other hand, **partial duplication** involves duplicating the image multiple times while using different captions for each duplicate(although the captions may be semantically similar). To study the partial duplication scenario, we generate 20 captions for each image using the BLIP model for both Imagenette and LAION-10k datasets. For the full-duplication case in the LAION-10k experiments, we keep the original caption from the dataset.

We present the results on LAION-10k and Imagenette in Figure 3. We investigate how dataset similarity changes for both full and partial image-caption duplication at varying levels of duplication. Overall, our findings demonstrate that partial duplication consistently leads to lower levels of memorization compared to full duplication scenarios.

In Figure 3 (left), we compare the similarity scores of several models: the pretrained checkpoint, a model finetuned without any data duplication, a model finetuned with full duplication (`ddf=5`), and a model finetuned with partial duplication (`ddf=5`). We include dashed horizontal lines representing the background self-similarity computed between the dataset and itself. In the Imagenette case, models trained without duplication and with partial duplication ex-

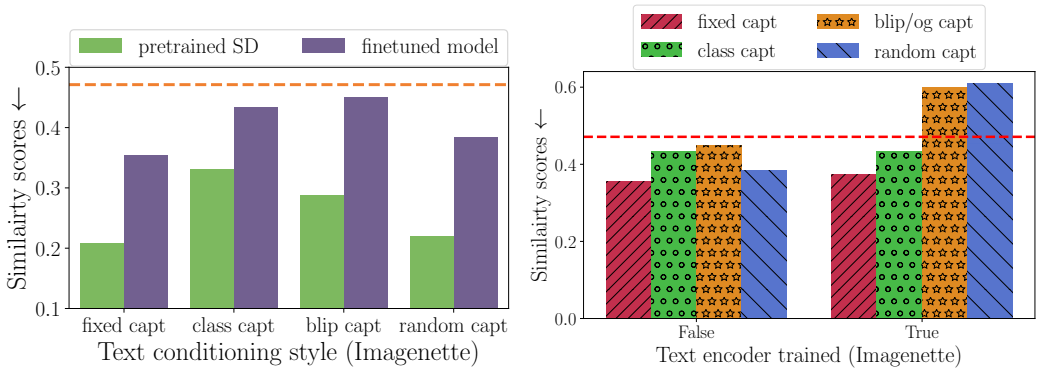

Figure 2. **Left:** Diffusion models finetuned on Imagenette with different styles of conditioning. FID scores of finetuned models are as follows (in order) $40.6, 47.4, 17.74, 39.8$. **Right:** We show the effects of training the text encoder on similarity scores with different types of conditioning

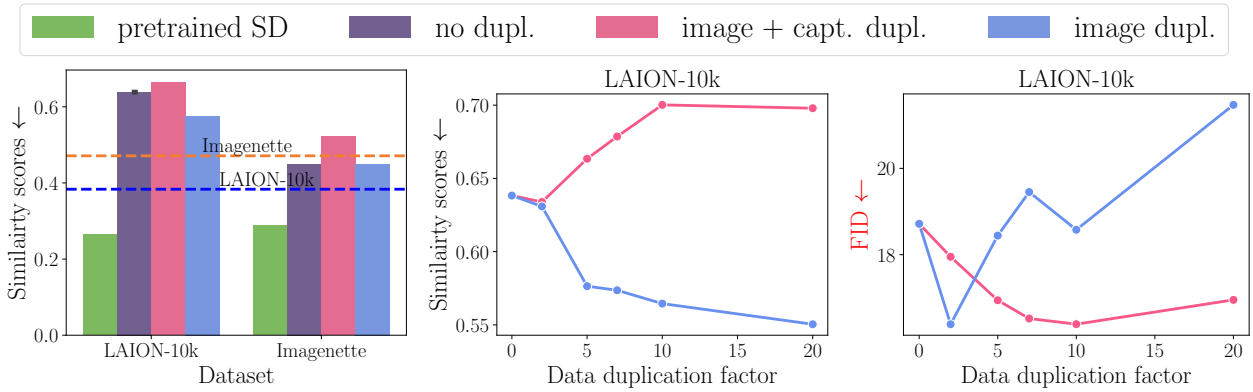

Figure 3. Models trained with different levels duplication and duplication settings. **Left:** Dataset similarity between models trained with no duplication, with partial duplication, and full duplication. Dashed lines show dataset similarity of each training distribution. **Middle, Right:** Dataset similarity and FID for full duplication vs partial duplication for different `data duplication factors`.

hibit dataset similarity below the baseline value, indicating a lower level of memorization. In contrast, the model trained with full duplication demonstrates higher levels of memorization compared to both the baseline and other cases. In the LAION-10k experiments, the model trained without duplication surpasses the training data similarity baseline. This observation raises the question of whether the memorization is inherited from the pretrained model, considering it is also trained on the larger LAION-2B dataset. However, when we compute the similarity scores on the pretrained checkpoint, we observe significantly lower values, indicating that the observed memorization is acquired during the fine-tuning process.

In Figure 3 (middle, right), we analyze how the similarity scores and FID scores vary at different `data duplication factors (ddf)` for full and partial data duplication. As the `ddf` increases, we observe an increase in memorization for models trained with full duplication. However, for partial duplication, dataset similarity actually *decreases* with increased duplication. In our previous analogy, we now have multiple captions, i.e. keys for each duplicated image, which inhibits the memorization capabilities of the model. However, this memorization reduction comes at the cost of a moderate increase in FID at higher duplication levels.

5. Conclusion

In our research, we make significant contributions to the understanding of dataset memorization in large text-conditional models. We demonstrate that beyond the conventional focus on de-duplicating image data, the conditioning and diversity of captions also play pivotal roles. Our findings, as illustrated in Figure 3, reveal that the phenomenon of memorization can be mitigated even as duplication levels increase, provided that captions exhibit sufficient diversity. These insights shed light on the contrasting behavior observed between large text-conditional models like Stable Diffusion and class-conditioned models on Imagenet, as reported in the study by Somepalli et al. (Somepalli et al., 2022).

References

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying Memorization Across Neural Language Models. *arxiv:2202.07646[cs]*, February 2022. doi: 10.48550/arXiv.2202.07646. URL http://arxiv.org/abs/2202.07646.

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. *arxiv:2301.13188[cs]*, January 2023. doi: 10.48550/arXiv.2301.13188. URL http://arxiv.org/abs/2301.13188.

David Chess. Some light infringement?, December 2022. URL https://ceoln.wordpress.com/2022/12/16/some-light-infringement/.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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.

Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring Forgetting of Memorized Training Examples. *arxiv:2207.00099[cs]*, June 2022. doi: 10.48550/arXiv.2207.00099. URL http://arxiv.org/abs/2207.00099.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating Concepts in Text-to-Image Diffusion Models. March 2023. URL https://arxiv.org/abs/2303.13516v2.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. Do Language Models Plagiarize? *arxiv:2203.07618[cs]*, March 2022. doi: 10.48550/arXiv.2203.07618. URL http://arxiv.org/abs/2203.07618.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

Emad Mostaque. Some more from first batch. Lots of optimisation to do, took about an hour of playing about. Prompts here: https://t.co/4soGak9op0 negative important for 2.0 given how we flatten distribution of latents with dedupe etc Embeddings will make it easier out of the box https://t.co/sZxhrk1v8J, November 2022. URL https://twitter.com/EMostaque/status/1596907328548139008.

Alex Nichol. Dall·e 2 pre-training mitigations, 2022. URL https://openai.com/research/dall-e-2-pre-training-mitigations. Accessed: 2023-05-05.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A Self-Supervised Descriptor for Image Copy Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Pizzi_A_Self-Supervised_Descriptor_for_Image_Copy_Detection_CVPR_2022_paper.html.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-Teaming the Stable Diffusion Safety Filter. *arxiv:2210.04610[cs]*, October 2022. doi: 10.48550/arXiv.2210.04610. URL http://arxiv.org/abs/2210.04610.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

Joseph Saveri and Matthew Butterick. Stable Diffusion Litigation, 2023. URL https://stablediffusionlitigation.com/.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. *arxiv:2211.05105[cs]*, April 2023. doi: 10.48550/arXiv.2211.05105. URL http://arxiv.org/abs/2211.05105.

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. *arXiv preprint arXiv:2210.08402*, 2022.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. *arxiv:2212.03860[cs]*, December 2022. doi: 10.

48550/arXiv.2212.03860. URL http://arxiv.org/abs/2212.03860.

Nikhil Vyas, Sham Kakade, and Boaz Barak. Provable Copyright Protection for Generative Models. *arxiv:2302.10870[cs, stat]*, February 2023. doi: 10.48550/arXiv.2302.10870. URL http://arxiv.org/abs/2302.10870.

Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. This Person (Probably) Exists. Identity Membership Attacks Against GAN Generated Faces. *arxiv:2107.06018[cs]*, July 2021. doi: 10.48550/arXiv.2107.06018. URL http://arxiv.org/abs/2107.06018.

Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. On the de-duplication of laion-2b. *arXiv preprint arXiv:2303.12733*, 2023.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard Prompts Made Easy: Gradient-Based Discrete Optimization for Prompt Tuning and Discovery. February 2023. URL https://arxiv.org/abs/2302.03668v1.