Diffusion Art or Digital Forgery? Investigating Data Replication in Stable Diffusion

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

1. Introduction

The rapid rise of diffusion models has led to new generative tools with the potential to be used for commercial art and graphic design. The power of the diffusion paradigm stems in large part from its reliance on simple denoising networks that maintain their stability when trained on huge web-scale datasets containing billions of image-caption pairs. These mega-datasets have the power to forge commercial models like *DALL·E* (Ramesh et al., 2022) and *Stable Diffusion* (Rombach et al., 2022), but also bring with them a number of legal and ethical risks (Birhane et al., 2021). Because these datasets are too large for careful human curation, the origins and intellectual property rights of the data sources are largely unknown. This fact, combined with the ability of large models to memorize their training data (Feldman and Zhang, 2020; Carlini et al., 2021; 2022), raises questions about the originality of diffusion outputs. There is a risk that diffusion models might, without notice, reproduce data from the training set directly, or present a collage of multiple training images.

We informally refer to the reproduction of training images, either in part or in whole, as *content replication*. In principle, replicating partial or complete information from the training data has implications for the ethical and legal use of diffusion models in terms of attributions to artists and photographers. Replicants are either a benefit or a hazard; there may be situations where content replication is acceptable, desirable, or fair use, and others where it is "stealing." While these ethical boundaries are unclear at this time, we focus on the scientific question of *whether replication actually happens with modern state-of-the-art diffusion models, and to what degree*.

In this study we find the dataset replication in *Stable Diffusion* model which is trained on millions of images (Figure 1). Furthermore, we believe that the rate of content replication we identify in *Stable Diffusion* likely underestimates the true rate because the model is trained on a 2B image split of LAION, but we only search for matches in the smaller 12M "Aesthetics v2 6+" subset.

The level of image similarity required for something to count as "replication" is subjective and may depend on both the amount of diversity within the image's class as well as the observer. Some replication behaviors we uncover are unambiguous, while in other instances they fall into a gray area. Rather than choosing an arbitrary definition, we focus on presenting the results to the reader, leaving each person to draw their own conclusions based on their role and stake in the process of generative AI. Project page: `https://somepago.github.io/diffrep.html`

2. Training data replication in Stable Diffusion

In this section, we evaluate *Stable Diffusion* v.1.4 (Rombach et al., 2022), which was trained on the publicly available LAION (Schuhmann et al., 2022) dataset. Since it is computationally expensive to store and search 2 billion+ images, we narrow our search scope to the smaller LAION Aesthetics v2 6+ dataset which has 12M images and is a **subset** of images that were used for the final rounds of training. We load the model and the checkpoints via HuggingFace [1].

In the first experiment, we randomly sample 9000 images, which we call *source images*, from LAION Aesthetics 12M and retrieve the corresponding captions. Then, we generate synthetic images by passing these captions into *Stable Diffusion*. We study the top-1 matches, which we call *match images*, for each generated sample.

We attempt to answer the following questions in this analysis. 1) Is there copying in the generations? 2) If yes, what kind of copying? 3) What triggers replication? 4) Can the model copy style?

We evaluated many self-supervised and instance retrieval, and copy detection models as feature descriptors to find the most similar-looking images to a given generation. Qualitatively we observed that SSCD performs best compared to other backbones. We constructed visualizations in this section by choosing from images with an SSCD similarity

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

Workshop on Generative AI and Law, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]`huggingface.co/CompVis/stable-diffusion-v1-4`

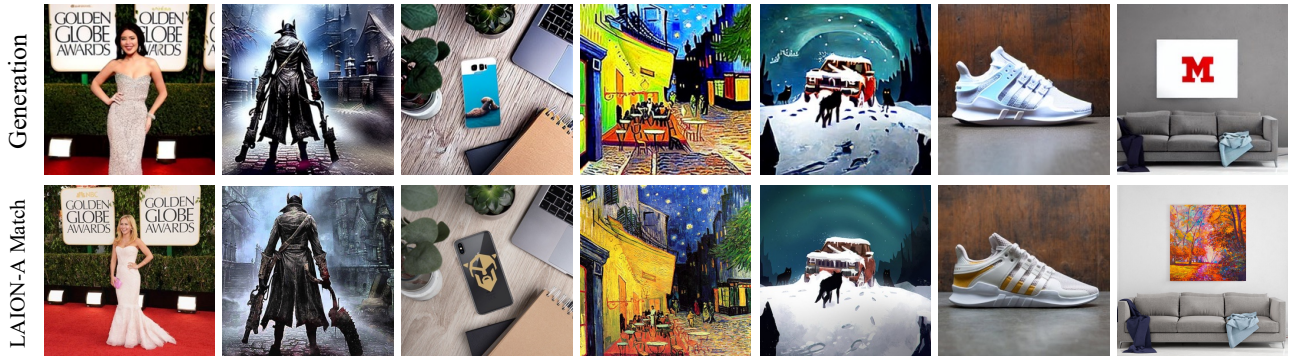

Figure 1. Top row: generated images. Bottom row: closest matches in the LAION-Aesthetics v2 6+ set. Sometimes source and match prompts are quite similar, and sometimes they are quite different.

> 0.5.

Figure 2. Including the phrase highlighted in red into a random prompt for *Stable Diffusion* leads to exact replications of the sofa (top row) and wave shape (bottom row).

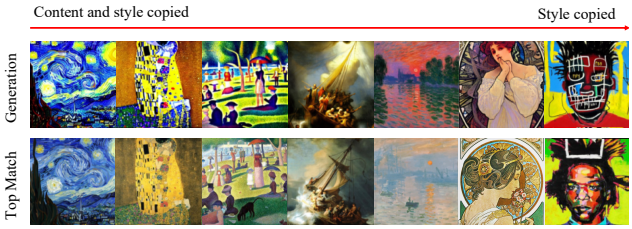

Figure 3. Stable Diffusion replicates pixel-level details, structures, and styles of well known paintings.

Observations. In Figure 1, we visualize a few instances of copying found in samples generated by *Stable Diffusion*. We choose them from a small set of points (≈ 170 images) whose top-1 similarity scores are > 0.5 (top 1.88 percentile). Above this 0.5 threshold, we observe a significant amount of copying. First row shows the generation and the second row shows the top-1 match based on SSCD. In many cases, we see verbatim replication of an object and background. And in a few cases, only the background is recycled from the training set.

While all synthetic images were generated using captions sourced from LAION, none of the generations match their respective source image. In fact, sometimes the caption of the source image is not representative of the source image content, and the generation is quite different from the source.

In those 170 images, we find instances where replication behavior is highly dependent on key phrases in the caption. We show two examples in Figure 2 and highlight the key phrase in red. For the first row, the presence of the text `Canvas Wall Art Print` frequently ($\approx$ 20% of the time) results in generations containing a particular sofa from LAION (also see Fig 1). Similarly, the second row shows various generations by tweaking the prompt `A painting of the Great Wave off Kanagawa by Katsushika Hokusai`. We gradually remove words until only `painting` and `wave` remain. All of the generations have a wave structure that resembles the original painting. We also notice instances of generations where style is copied rather than content. This can be explicitly seen when the name of an artist is used in the generation prompt. We generate many paintings with the prompt style "`<Name of the painting>` by `<Name of the artist>`". We tried several classical and contemporary artists, and we observe that the generations frequently reproduce known paintings with varying degrees of accuracy. In Figure 3, as we go from left to right, we see that content copying is reduced, however, style copying is still prevalent.

In conclusion, *Stable Diffusion* is capable of reproducing training data, creating images by piecing together foreground and background objects that it has memorized. Furthermore, the system sometimes exhibits *reconstructive* memory, in which recalled objects are semantically equivalent to their source object without being pixel-wise identical. We showed cases of content and style copying. The presence of such replicated images raises questions about the nature of data memorization and the ownership of diffusion images.

References

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arxiv:2110.01963[cs]*, October 2021. doi: 10.48550/arXiv.2110.01963. URL http://arxiv.org/abs/2110.01963.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

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.

Vitaly Feldman and Chiyuan Zhang. What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation. *arxiv:2008.03703[cs, stat]*, August 2020. doi: 10.48550/arXiv.2008.03703. URL http://arxiv.org/abs/2008.03703.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 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*, pages 10684–10695, 2022.

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.