





Eric Wallace @Eric_Wallace_22h

Models such as Stable Diffusion are trained on copyrighted, trademarked, private, and sensitive images.

Yet, our new paper shows that diffusion models memorize images from their training data and emit them at generation time.

Paper: arxiv.org/abs/2301.13188

👉 [1/9]

Training Set	Generated Image
	
<p><i>Caption: Living in the light with Ann Graham Lotz</i></p>	<p><i>Prompt: Ann Graham Lotz</i></p>

Jan 31, 2023 · 3:52 PM UTC · Twitter Web App

💬 119 ↗️ 1,596 🗨️ 305 ❤️ 7,674



Eric Wallace @Eric_Wallace_22h

Diffusion models are trained to denoise images from the web. These images are often vulgar or malicious, and many are potentially risky to use (e.g., copyrighted).

Moreover, many ongoing projects apply diffusion models to private applications such as medical imagery. [2/9]



💬 1 ↻ 19 🗨️ 1 ❤️ 463



Eric Wallace @Eric_Wallace_ 22h

We thus study if diffusion models “memorize” training examples, which we define as generating a near-identical copy of any image.

We propose to extract memorized images by generating many times with the same prompt and flagging cases where many of the generations are the same.

💬 1 ↻ 21 🗨️ ❤️ 479



Eric Wallace @Eric_Wallace_ 22h

Applying our method to Stable Diffusion and Google’s Imagen, we extract hundreds of images, and do so with high precision.

Many of these images are copyright or licensed, and some are photos of individuals. [4/9]

Original:



Generated:



💬 4 ↻ 76 🗨️ 6 ❤️ 834

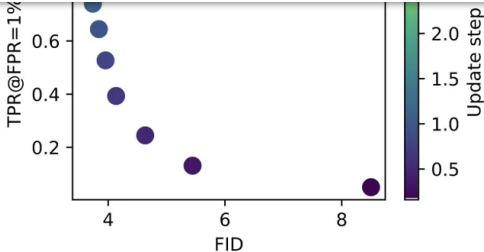


Eric Wallace @Eric_Wallace_ 22h

We also train hundreds of our own diffusion models to study the impact of various factors. Some highlights:

- Diffusion models memorize more than GANs
- Outlier images are memorized more
- Existing privacy-preserving methods largely fail

GANs	StyleGAN-ADA [43]	150	2.9
	DiffBigGAN [82]	57	4.6
	E2GAN [69]	95	11.3
	NDA [63]	70	12.6
	WGAN-ALP [68]	49	13.0
DDPMs	OpenAI-DDPM [52]	301	2.9
	DDPM [33]	232	3.2



7 32 2 530



Eric Wallace @Eric_Wallace_

22h

See our paper for a lot more technical details and results.

Speaking personally, I have many thoughts on this paper. First, everyone should de-duplicate their data as it reduces memorization. However, we can still extract non-duplicated images in rare cases! [6/9]



3 13 4 444



Eric Wallace @Eric_Wallace_

22h

Second, Stable Diffusion is small relative to its training set (2GB of weights and many TB of data). So, while memorization is rare by design, future (larger) diffusion models will memorize more.

Third, don't apply today's diffusion models to privacy sensitive domains. [7/9]

1 23 4 477



Eric Wallace @Eric_Wallace_

22h


Finally, there are open questions about the impact of our work on ongoing lawsuits against StabilityAI, OpenAI, GitHub, etc. Specifically, models that memorize some of their training points may be viewed differently under statutes like GDPR, US trademark + copyright, and more.

3 26 2 452



industry (Google + Deepmind) and academia (Berkeley + Princeton + ETH Zurich).

See our paper for more details arxiv.org/abs/2301.13188 and I am happy to take questions and comments! [9/9]



Extracting Training Data from Diffusion Models

Image diffusion models such as DALL-E 2, Imagen, and Stable Diffusion have attracted significant attention due to their ability to generate high-quality synthetic images. In this work, we show...

arxiv.org

💬 23 ↗️ 30 🗣️ 1 ❤️ 484



David [@hcetamd](#) 13h

Replying to [@Eric_Wallace_](#)

Now imagine that applied to source code or anything else that's not public domain.

💬 1 ↗️ 🗣️ ❤️ 11



Eric Wallace [@Eric_Wallace_](#) 9h

Yes exactly! We have also studied similar questions of memorization for language models in a past paper arxiv.org/abs/2012.07805



Extracting Training Data from Large Language Models

It has become common to publish large (billion parameter) language models that have been trained on private datasets. This paper demonstrates that in such settings, an adversary can perform a...

arxiv.org

💬 ↗️ 2 🗣️ 1 ❤️ 25

Load more

