GenAI Confessions:
Black-box Membership Inference for Generative Image Models

Matyas Bohacek
Stanford University
maty@stanford.edu

Hany Farid
University of California, Berkeley
hfarid@berkeley.edu

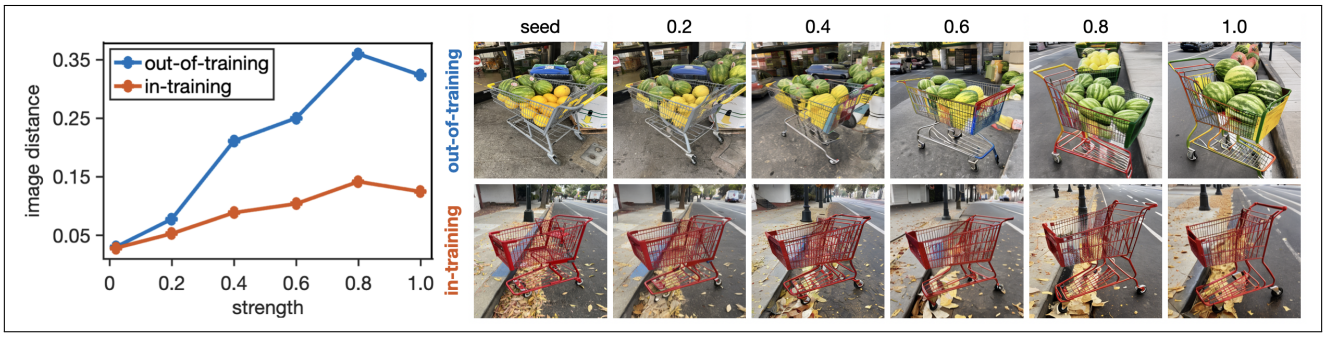

Figure 1. **Overview of Our Membership Inference Method.** Images are generated using an image-to-image pipeline, seeded with an in-training image (bottom) and an out-of-training image (top), across increasing strengths. A larger strength corresponds to reduced emphasis on the seed image. The plot on the left shows the similarity between each generated image and its corresponding seed image (the DreamSim image similarity metric), where smaller values indicate greater similarity. The in-training seed consistently yields more similar outputs than the out-of-training seed. Our membership inference method exploits this phenomenon.

Abstract

From a simple text prompt, generative-AI image models can create stunningly realistic and creative images bounded, it seems, by only our imagination. These models have achieved this remarkable feat thanks, in part, to the ingestion of billions of images collected from nearly every corner of the internet. Many creators have understandably expressed concern over how their intellectual property has been ingested without their permission or a mechanism to opt out of training. As a result, questions of fair use and copyright infringement have quickly emerged. We describe a method that allows us to determine if a model was trained on a specific image or set of images. This method is computationally efficient and assumes no explicit knowledge of the model architecture or weights (so-called black-box membership inference). We anticipate that this method will be crucial for auditing existing models and, looking ahead, ensuring the fairer development and deployment of generative AI models.

Project Page: genai-confessions.github.io
Data: hf.co/datasets/faridlab/stroll

1. Introduction

Most agree that – despite occasional hallucinations, extra fingers and toes, and gravity-defying motion – AI-powered systems are now capable of creating human-like prose, image, and video from a simple prompt. Most also agree that these systems, in the form of large-language [17], image [19], and video [15] models, are only possible thanks to the ingestion of massive amounts of human-generated content. Here, however, is where disagreements begin [6].

As the courts – and the court of public opinion – adjudicate these matters in the coming years, the question of whether a model was trained on a specific dataset will be critical. This so-called question of *membership inference* is challenging for a number of reasons. First, training datasets are massive and often created through large-scale web scraping without careful record keeping [25]. Second, once trained, the models are opaque, making a post-hoc inference challenging. And third, given the competitive landscape and lack of clear laws, there is currently little incentive for rule following, with even some former tech CEOs encouraging young entrepreneurs to steal intellectual property and later hire lawyers to "clean up the mess" [10].

The study of membership inference emerged as deep-learning applications started to be trained on large datasets [12]. Early work focused on membership inference targeting classification (as compared to generative) models [23]. More recently, attention has turned to membership inference for generative models, including large language models (LLMs) [4, 14], generative adversarial networks (GANs) [9, 11], and diffusion-based text-to-image models [3, 13, 26–28].

Diffusion-based models are the leading contenders in producing photorealistic images and video and hence the focus of our effort. Previous membership inference methods for these models either assume full or partial knowledge of the generative model architecture and trained weights (white-box or gray-box), are applicable to only one model architecture, require massive computing power to operationalize, or only work for analyzing membership for an entire dataset, as compared to a single image.

By contrast, our membership inference assumes no explicit knowledge of the model details (black-box), generalizes to different model architectures, is computationally efficient, and can operate on both a dataset of images as well as a single image.

The contributions of this paper can be summarized as follows:

- A computationally efficient and easy to implement, black-box membership inference method for generative-AI image models (Figure 1);
- A dataset (**STROLL**) of semantically matched image pairs for evaluating membership inference;
- Empirical analysis of membership inference and memorization across different model architectures.

2. Methods

2.1. Data

We compiled three datasets consisting of images generated by Stable Diffusion (v1.4, v2.1, v3.0), Midjourney (v6), and DALL-E (v2). As described below, these datasets consist of paired in-training and out-of-training images used to evaluate our membership inference technique. Each pair of images is constructed to be semantically similar in terms of content so as to ensure that any observed differences between in-training and out-of-training seed images is not due to semantic differences.

2.1.1. STROLL

This dataset contains 100 in-training and out-of-training image pairs of outdoor city objects and scenes recorded on a smartphone in the San Francisco Bay area over the course of two days in July 2024. Shown in the top panel of Figure 2 are representative image pairs (first two rows).

Figure 2. Representative examples of in-training and out-of-training images from the STROLL (top), Carlini (middle), and Midjourney (bottom) datasets.

Prompted with "Provide a detailed, 15 word long caption of this image," ChatGPT-4o [1] was used to generate a detailed caption for each image. These captions were used for the in-training/out-of-training experiment. For a second, in-training (alt caption)/out-of-training experiment, a new caption was generated for the in-training image (with the out-of-training captions remaining the same). This alternate caption was generated with the prompt "Provide a detailed, 15 word long caption of this image that is distinctly different from [previous caption]".

2.1.2. Carlini

This dataset contains 74 images that appear to have been memorized [3] by Stable Diffusion (v1.4) [20]. Shown in the middle panel of Figure 2 are representative images. These images are a tiny fraction of the LAION-5B [22] dataset used to train Stable Diffusion (v1.4). Each in-training image in this dataset is accompanied by its original caption from LAION-5B, which was also used to generate matching out-of-training images using DALL-E (v2) [18] (middle panel of Figure 2).

2.1.3. Midjourney

This dataset contains 10 images that appear to have been memorized [24] by Midjourney (v6) [2]. Shown in the bottom panel of Figure 2 are representative images. Each in-training image in this dataset is accompanied by its original caption from LAION-5B, which was used to generate matching out-of-training images using Stable Diffusion (v3) [7] (Figure 2). Unlike the previous two datasets in which the control images were generated using DALL-E, here we use Stable Diffusion (v3) because DALL-E would not generate many of the images in this dataset consisting of recognizable celebrities.

2.2. Models

2.2.1. STROLL/Stable Diffusion (v2.1)

We created a custom derivative of the Stable Diffusion (v2.1) variant image model by fine-tuning the 2.1 model weights[1] on the in-training portion of the STROLL dataset. The official training script[2] was used to fine-tune the model's UNet module while keeping the CLIP encoder and variational autoencoder (VAE) weights frozen. The learning rate was set to 10^{-5}, and the maximum-steps parameter was set to 100[3], while the remaining parameters were left at their recommended default values: image resolution of 512×512, mixed precision turned off, and a random horizontal flip augmentation.

Given a seed image and text prompt as input, the image-to-image feature of this fine-tuned model, with default parameters and varying strengths, was used to power our membership inference. The strength $s_i \in [0, 1]$ controls the influence of the text prompt relative to the seed image, where a value of 0 yields a generated image that is identical to the seed image, and a value of 1 generates an image guided fully by the text prompt, effectively ignoring the seed image.

In order to determine the impact of the number of training steps, we created a second fine-tuned model in which the training steps was increased from 100 to 1,000.

2.2.2. Carlini/Stable Diffusion (v1.4)

We used the off-the-shelf Stable Diffusion (v1.4) model[4] and invoked its image-to-image pipeline. All image generation parameters were set to the default values with image strength $s_i \in [0, 1]$.

2.2.3. Midjourney/Midjourney (v6)

We used the commercial Midjourney model[5] and manually invoked its image-to-image pipeline through their Discord interface. All parameters were set to the default values, except for the image strength (termed weight in Midjourney), ranging from 0 (yielding a generated image that ignores the seed image) to 3 (yielding a generated image identical to the seed image). Note that this strength parameter is reversed as compared to Stable Diffusion.

2.3. Membership Inference

Our membership inference method predicts whether a model M was trained on an image I with caption C. This method does not access any explicit information about M's architecture or trained weights. This method only requires access to the image-to-image inference engine for generating an image from a descriptive prompt, seed image, and variable strength parameter that controls the deviation between the seed image and generated image. Intuitively, this approach exploits a – perhaps unintended – property of image-to-image generation that produces less variation for an in-training seed image as compared to an out-of-training seed image.

Our method involves three steps: (1) image-to-image inference with varying strengths, (2) measurement of perceptual similarity between a generated and seed image I; and (3) membership inference prediction quantifying the likelihood that model M was trained on image I.

In the first step, the image-to-image pipeline of model M is invoked with a seed image I, its descriptive caption C, and strength parameters s_i, where $i = 1, 2, \ldots, m$. For each strength s_i, the image generation is repeated n times, resulting in a set of output images $\hat{I}_{i,j}$, where $j = 1, 2, \ldots, n$.

[1] www.huggingface.co/stabilityai/stable-diffusion-2-1

[2] https://github.com/huggingface/diffusers/blob/main/examples/text_to_image/train_text_to_image.py

[3] For reference, fine-tuning SD v2.1 from v2.0 took 210,000 steps.

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

[5] docs.midjourney.com/docs/image-prompts

In the second step, the distance $d_{i,j}$ between the seed image I and each generated image $\hat{I}_{i,j}$ is calculated using DreamSim [8]. This perceptual metric of image similarity computes a distance $d_{i,j} \in [0, 1]$ where a value of 0 is maximally similar and a value of 1 is maximally different. For each strength s_i, the minimum distance across n generated images is retained, yielding a m-D vector of distances for each strength value: $\vec{d} = \begin{pmatrix} d_1 & d_2 & \dots & d_m \end{pmatrix}$.

We use a simple logistic-regression model to distinguish in-training from out-of-training images based on the distances $\vec{d}$.

Stable Diffusion and Midjourney afford a different parametrization of the strength variable s_i: For Stable Diffusion (v1.4 and v2.1), $s_i \in [0.02, 0.2, 0.4, 0.6, 0.8, 1.0]$; for Midjourney (v6), $s_i \in [0, 1, 2, 3]$. For Stable Diffusion, we used a minimum strength of 0.02 because a strength of 0.0 simply returned the seed image. Throughout, $n = 10$ were generated at each strength parameter.

3. Results

As described in detail in Section 2, our membership inference method predicts whether a model M was trained on an image I with caption C. This method involves three steps: (1) image-to-image inference with varying strengths, (2) measurement of perceptual similarity between a generated image and seed image I with caption C; and (3) membership inference prediction quantifying the likelihood that model M was trained on image I. Intuitively, this method exploits an emergent property in which image-to-image generation produces less variation for an in-training seed image as compared to an out-of-training seed image.

3.1. STROLL

Shown in Figure 3(a) is a Gaussian fitted log-probability density function to the DreamSim distance between the in-training seed images and the result of image-to-image generation under Stable Diffusion 2.1. Each curve corresponds to a different image-to-image strength parameter s_i (see Section 2) where, as strength increases, the seed image has increasingly less impact on the generated image. As expected, for strength parameters close to 0 (thinnest curve), the DreamSim distance is relatively small, and as the strength increases (thicker curves), the distance increases proportionally, meaning that the generated images are increasingly more distinct from the seed image.

Shown in Figure 3(d) are the same density functions but for the out-of-training seed images. Here we see the same trend, where small strength parameters lead to more similarity as compared to larger strength parameters. However, the mean of these densities as a function of strength s_i is larger for these out-of-training images. In particular, notice

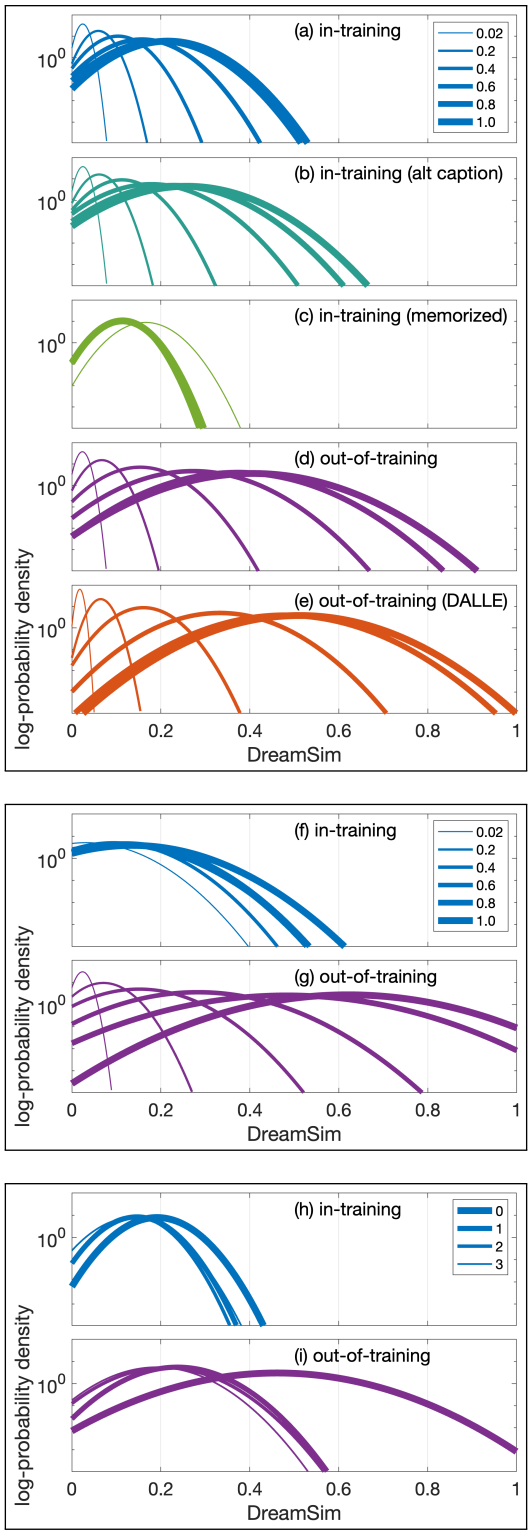

Figure 3. The log probability densities for image similarity (DreamSim) between a seed image and a generated image with varying strength for the STROLL (top), Carlini (middle), and Midjourney (bottom) datasets. In all cases, in-training seed images lead to generated images that are more perceptually similar to the seed (smaller DreamSim distance).

that the mean of the densities for strengths greater than 0.6 are significantly larger for out-of-training as compared to in-training images. That is, the images generated with an out-of-training seed are more distinct than those generated with an in-training seed.

Shown in the top panel of Figure 4 is an in-training seed image (far left) and the resulting image-to-image generation for strengths $s_i \in [0.02, 0.2, 0.4, 0.6, 0.8, 1.0]$. Consistent with the DreamSim distance (Figure 3(a)), all of the generated images are perceptually similar to the seed image. By comparison, also shown in Figure 4, is an out-of-training seed image and the resulting image-to-image generation for varying strengths. Again, consistent with the DreamSim distance (Figure 3(d)), the generated images deviate from the seed starting at a strength of 0.6.

Independent-samples, two-sided t-tests reveal a significant difference between the DreamSim distributions for the in-training vs out-of-training data at a strength $s_i \geq 0.4$, with an increasing effect size (Cohen's D) with increasing strength:

- 0.02 ($t(198) = 0.5,\ p = 0.6,\ D = 0.07$)
- 0.2 ($t(198) = 2.3,\ p = 0.02,\ D = 0.32$)
- 0.4 ($t(198) = 7.3,\ p < 10^{-8},\ D = 1.0$)
- 0.6 ($t(198) = 11.2,\ p < 10^{-8},\ D = 1.6$)
- 0.8 ($t(198) = 14.5,\ p < 10^{-8},\ D = 2.1$)
- 1.0 ($t(198) = 14.8,\ p < 10^{-8},\ D = 2.1$)

3.1.1. Alternate Caption

In the above analysis, we assume that the image-to-image generation is provided with the same image caption used in training of the image-generation model. To test the sensitivity of this assumption, alternate captions were generated for each image (see Section 2). Shown in Figure 3(b) are the same Gaussian fitted log-probability density functions to the DreamSim distances. As compared to the in-training images with the original caption (Figure 3(a)), the generated images with alternate captions are less similar to their seed images, but still distinct from the out-of-training images (Figure 3(d)).

An independent-samples, two-sided t-tests again reveals a significant difference between the DreamSim distributions for the in-training vs out-of-training data at a strength $s_i \geq 0.4$:

- 0.02 ($t(198) = 0.7,\ p = 0.5, D = 0.1$)
- 0.2 ($t(198) = 1.6,\ p = 0.1, D = 0.2$)
- 0.4 ($t(198) = 5.7,\ p < 10^{-8},\ D = 0.8$)
- 0.6 ($t(198) = 8.5,\ p < 10^{-8},\ D = 1.2$)
- 0.8 ($t(198) = 11.1,\ p < 10^{-8},\ D = 1.6$)
- 1.0 ($t(198) = 10.5,\ p < 10^{-8},\ D = 1.5$)

3.1.2. Effect Size

We observe qualitative differences between the distributions for the in-training and out-of-training data increases with strength s_i. We quantify this with Cohen's D, a measure of effect size: for strengths $0.02, 0.2, \ldots, 1.0$, the effect sizes are $0.1, 0.3, 1.0, 1.6, 2.1, 2.1$. A similar pattern emerges for the in-training and out-of-training (alt caption) distributions, with effect sizes of $0.1, 0.2, 0.8, 1.2, 1.6, 1.5$. With a Cohen's D of 0.8 considered large, we see large effects with strength parameters greater than 0.4

3.1.3. Classifier

The distributions in the top panel Figure 3 show a population-level difference between in-training and out-of-training images. To predict if an individual image belongs to the in-training or out-of-training set, we trained a logistic regression on the 6-D distance vectors $\vec{d}$ corresponding to the 100 in-training (original caption), 100 in-training (alternate caption), 100 out-of-training and 100 out-of-training (DALL-E) images. These DALL-E images, generated to match the content of the out-of-training images (see Section 2), balanced the data set for model training. As shown in Figure 3(e), the distributions for these images are similar to the out-of-training images in panel (d).

A logistic regression was trained on a random subset of 80% of this data and evaluated on the remaining 20%. From 100 random training/testing splits, the average testing accuracy (measured as equal error rate) is 85% with a variance of 0.17. For a fixed false positive rate (misclassifying an out-of-training image as in-training) of 1%, the average true positive rate (correctly classifying an in-training image) is 74% with a variance of 0.36.

3.1.4. Memorization

In the results described above, the fine-tuned Stable Diffusion model was trained with 100 steps. We next consider the impact of increasing the training steps to 1000 as described in Section 2.

Shown in Figure 3(c) is the Gaussian-fitted log probability densities for this model in which we can see a qualitatively different pattern than before, Figure 3(a). Here, almost regardless of strength, the generated images are uniformly similar to the seed image. This, we posit, is because the prolonged learning caused the model to effectively memorize the training images and associated captions. This is consistent with the results described next.

3.2. Carlini

In the previous section, we showed that when the pre-trained Stable Diffusion (v2.1) model is fine-tuned on a set of 100 images of our creation, we can determine that the model was trained on these images. Because this is a fairly

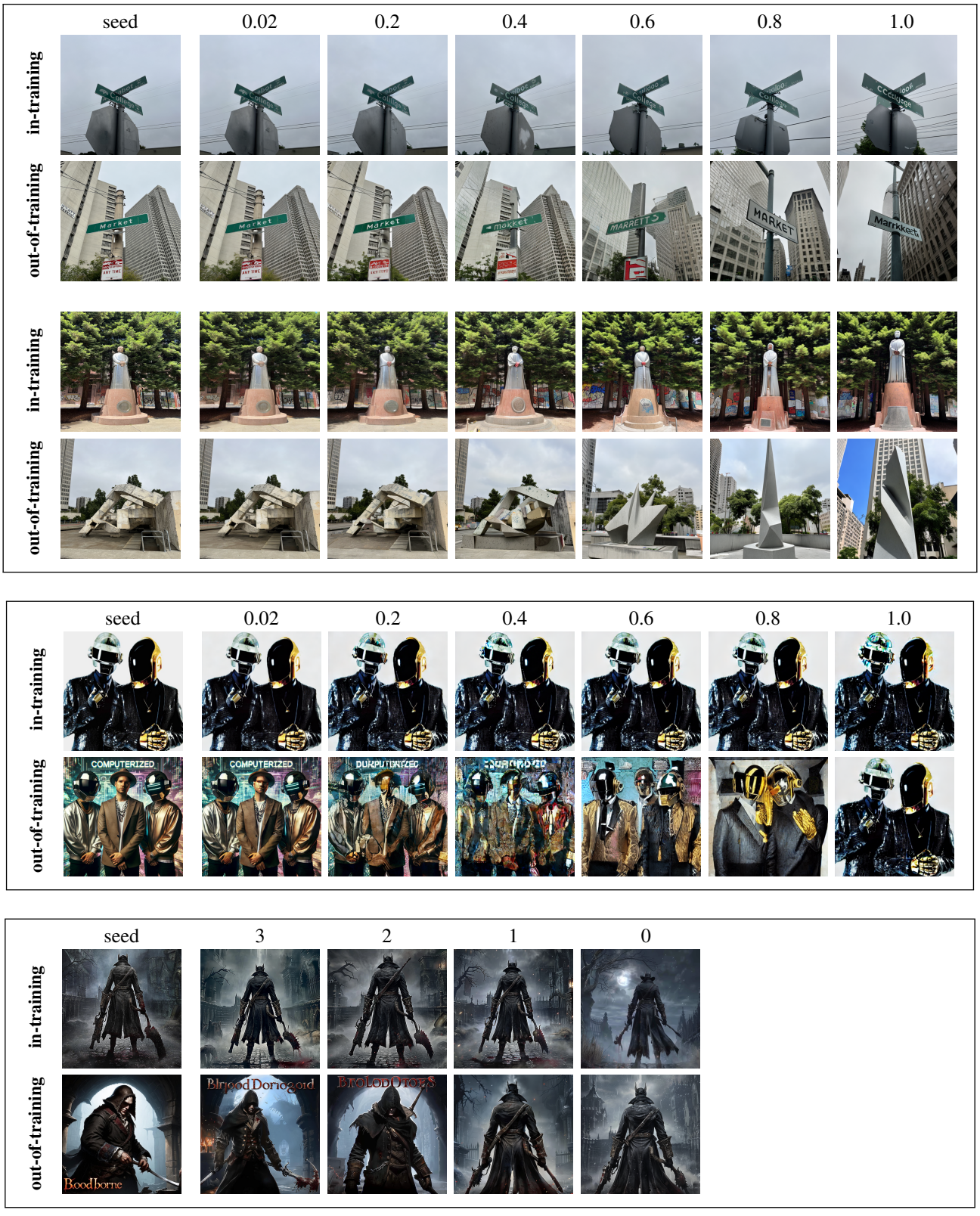

Figure 4. Images generated using the STROLL dataset and Stable Diffusion (v2.1) image-to-image pipeline (top), the Carlini dataset and Stable Diffusion (v1.4) model (middle), and Midjourney (v6) dataset and image-to-image pipeline. The seed image is shown in the left-most column. The remaining columns correspond to generated images with increasing strength, where moving rightward corresponds to less emphasis placed on the seed image. In all three cases, the in-training seed image leads to perceptually more similar images than the out-of-training seed images.

constrained experiment, we next validate that our membership inference generalizes to a real-world scenario.

As described in Section 2, the Carlini dataset consists of 74 images used to train Stable Diffusion (v1.4) and, as shown in [3], this model can be coaxed to produce images that are nearly indistinguishable from these training images, Figure 2.

Shown in Figure 3(f)-(g) are the DreamSim distances for the 74 in-training images and 74 semantically matched out-of-training images (see Section 2). Here, we see the same pattern as with the STROLL results described above: the generated images are perceptually more similar to the in-training seed images than the out-of-training seed images.

Shown in the middle panel of Figure 4 is an in-training seed image (far left) and the resulting image-to-image generation for varying strengths. As with the previous STROLL results, the generated images are perceptually similar to the seed image. By comparison, the out-of-training seed image yields generated images that deviate from the seed starting at a strength of 0.6. Interestingly, the out-of-training image at strength $s_i = 1.0$ is nearly identical to the in-training seed image. This is because, as shown by Carlini et al. [3], this image was memorized during the original training of the model, and so the original prompt yields the training image at a strength of 1.0 where the seed image is ignored.

3.2.1. Classifier

The average equal error rate for logistic regression trained on the STROLL images and evaluated on the Carlini dataset is 93% with a variance of 0.01, and for a false positive rate of 1%, the average true positive rate is 90% with a variance of 0.01. This accuracy is somewhat better than for the STROLL images because this dataset was not just trained on but was effectively memorized by the image generator, leading to a larger difference between in-training and out-of-training seed images. Here, we see that the classifier trained on a different dataset and version of Stable Diffusion generalizes quite nicely.

3.2.2. Memorization

Note that the in-training distributions, Figure 3(f), are qualitatively similar to the distributions in Figure 3(c) corresponding to the in-training (memorized) results. This, we believe, is because both of these models have memorized some training images and so we see less variation than in the case when the model was simply exposed to these images.

3.3. Midjourney

In the previous two sections, we showed the efficacy of our membership inference on two different versions of Stable Diffusion (v2.1 and v1.4). Here we show that our approach generalizes to different model architectures.

As described in Section 2, the Midjourney dataset consists of 10 images that appear to have been part of the training datatset for Midjourney (v6). In particular, as shown in [24], Midjourney can be coaxed to produce images that are nearly indistinguishable from these well-recognized images, Figure 2.

Shown in Figure 3(h)-(i) are the DreamSim distances for the 10 in-training images and 10 semantically matched out-of-training images (see Section 2). Here, we see the same pattern as with the STROLL and Carlini results: the generated images are perceptually more similar to the in-training seed images than the out-of-training seed images.

Shown in the bottom panel of Figure 4 is an in-training seed image (far left) and the resulting image-to-image generation for varying strengths. As with the previous STROLL and Carlini results, the generated images are perceptually similar to the seed image. By comparison, the out-of-training seed image leads to images that deviate more noticeably.

We again see that the out-of-training image, at a strength $s_i = 0$ where the seed image is ignored, is nearly identical to the in-training seed image. This is because this image was memorized during the original training of the model [24], and so the prompt simply reproduces it.

3.3.1. Memorization

As before, the in-training distributions, Figure 3(h), are qualitatively similar to those in Figure 3(c) and (f) corresponding to the STROLL in-training (memorized) and Carlini in-training results. This, again, is because all three of these models have memorized some training images.

3.3.2. Classifier

Because Midjourney uses a different strength parametrization than Stable Diffusion, we are not able to deploy the logistic regression model on a per-image basis.

3.4. Comparison to Previous Work

Comparing membership inference methods for generative-AI image models remains challenging due to the lack of standardized benchmarks and varying problem definitions/-configurations (including the frequency and intensity with which images are presented to the model during training). Nonetheless, we compare our method to existing membership inference approaches for the latest generative-AI image model architectures (diffusion) in terms of computational demands and overall effectiveness.

The method in [26] assumes access to a confirmed subset of the targeted model's in-training and out-of-training data. Its reported accuracy ranges from 65% to 100% across multiple models and datasets. The method in [27] assumes access to internal representations of the model during diffusion steps. At a fixed false positive rate of 1%, it achieves a

true positive rate ranging from 54% to 68% across multiple datasets. Similarly, the method in [13] also relies on internal representations during diffusion steps. At a fixed false positive rate of 1%, it achieves a true positive rate between 50% and 58% across different datasets. And, the method in [5] is an ensemble of four other methods, whose predictions are analyzed for statistical significance and overall confidence. At a fixed false positive rate of 1%, it achieves a true positive rate ranging from 25% to 100% across multiple datasets.

By contrast, our method does not require access to a confirmed subset of in-training/out-of-training data or internal representations from diffusion steps and achieves an average accuracy between 85% and 93%. At a fixed false positive rate of 1%, our method attains an average true positive rate of between 74% and 90%.

When tested in the wild, many existing methods have been found largely ineffective [4]. We, on the other hand, demonstrate that the DreamSim trends leveraged by our method persist even in an in-the-wild setting with Midjourney (v6) (see Section 3.3).

Most effective membership inference and training data extraction methods are computationally expensive. The method in [3], which assumes a white-box scenario with additional access to the model's internals, achieves a true positive rate of 71% at a false positive rate of 1%. This approach, however, is computationally demanding, requiring training 16 shadow models, each of which computes loss for all known in-training data points at each of $1{,}000$ diffusion steps. By contrast, our method does not require training and only needs to perform six generations with at most 50 diffusion steps.

4. Discussion

We have observed that when seeded with a previously trained image, image-to-image generation produces an image more similar to the seed image as compared to those generated from an out-of-training seed image. This is distinct from pure memorization, where it has previously been shown that, with a sufficient amount of exposure, models can reproduce training images [3]. Our approach applies to both this less common case of memorization as well as the more typical and broader class of training images.

We hypothesize a few different mechanisms that may explain why generative-AI models behave this way. One possibility is that when a seed image is partially corrupted with additive noise (proportional to the user-supplied strength parameter) and placed in the latent space for denoising, because of previous exposure to a training image-caption pair, the previously learned local gradients guide the denoising to a latent representation near a trained image. This explanation would be consistent with the differences seen in Figure 3(a)-(d), where a memorized image/caption yields more self-similar images than an in-training image/caption, which yields more self-similar images than an in-training image/alternate caption. That is, the level of exposure to a specific image/caption pair at training leads to proportionally learned gradients in the denoiser.

Another possibility is that after training, the latent space is non-uniformly structured, and so once an image/caption pair is placed into latent space near an in-training exemplar, it is simply more likely to converge to the in-training image because of this structure. This is more likely to occur with image-to-image generation because the initialization in the latent space is dependent on the seed image and the strength parameter constrains the number of steps that can be taken by the denoiser.

Understanding why models are biased to produce content similar to their training data may provide insights into reducing the likelihood of infringement in the form of reproducing training data, and may provide insights into how a model can be made to forget training exemplars.

An attractive aspect of our membership inference for generative image models is that it does not require access to model architecture details or trained weights, is computationally efficient, and generalizes to multiple different AI models. A drawback of our approach is that it only applies to models that allow for an image-to-image synthesis with a controllable strength parameter, as compared to text-to-image. Depending on the underlying mechanism by which models produce images similar to their in-training data, our method may be adaptable to text-to-image generation.

Many of today's tech leaders have admitted that their generative-AI models would not exist without their training on billions of pieces of content scraped from all corners of the internet [16]. These same leaders have also called for the loosening of fair-use and copyright rules. While it is for the courts to decide on these matters of law [21], we contend that content creators have legitimate concerns for whether and how their content is used to train generative-AI models, some of which are designed to offer services directly competing with these very content creators.

A critical component of adjudicating these issues will be determining if a deployed model was trained on a specific piece of content. Equally important is determining how creators can and should be compensated when their content is used for training, and how models can be made to forget its training on a specific piece of content should this be the wish of the content's creator. We have focused only on the first of these questions, but all of these issues are important to resolve as generative AI continues its impressive and impactful trajectory.

Acknowledgments

This work was funded by funding from the University of California Noyce Initiative. The authors thank Nicholas Carlini for providing the list of extracted instances of memorization from [3], and both Nicholas Carlini and Milad Nasr for helpful discussions.

References

[1] Chatgpt-4o. https://chatgpt.com/, 2024. 3

[2] Midjourney. https://www.midjourney.com/, 2024. 3

[3] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. 2, 3, 7, 8, 9

[4] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? arXiv:2402.07841, 2024. 2, 8

[5] Jan Dubiński, Antoni Kowalczuk, Franziska Boenisch, and Adam Dziedzic. CDI: Copyrighted data identification in diffusion models. arXiv:2411.12858, 2024. 8

[6] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. Art and the science of generative AI. *Science*, 380(6650):1110–1111, 2023. 1

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *41st International Conference on Machine Learning*, 2024. 3

[8] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[9] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership inference attacks against generative models. arXiv:1705.07663, 2017. 2

[10] Alex Heath. Ex-Google CEO says successful AI startups can steal ip and hire lawyers to 'clean up the mess'. The Verge, 2024. 1

[11] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte Carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019. 2

[12] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys*, 54(11s):1–37, 2022. 2

[13] Qiao Li, Xiaomeng Fu, Xi Wang, Jin Liu, Xingyu Gao, Jiao Dai, and Jizhong Han. Unveiling structural memorization: Structural membership inference attack for text-to-image diffusion models. In *ACM International Conference on Multimedia*, pages 10554–10562, 2024. 2, 8

[14] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. arXiv:2305.18462, 2023. 2

[15] Andrew Melnik, Michal Ljubljanac, Cong Lu, Qi Yan, Weiming Ren, and Helge Ritter. Video diffusion models: A survey. arXiv:2405.03150, 2024. 1

[16] Dan Milmo. 'impossible' to create AI tools like ChatGPT without copyrighted material, OpenAI says. The Guardian, 2024. 8

[17] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. arXiv:2402.06196, 2024. 1

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125, 2022. 3

[19] Gaurav Raut and Apoorv Singh. Generative AI in vision: A survey on models, metrics and applications. arXiv:2402.16369, 2024. 1

[20] 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 (CVPR)*, pages 10684–10695, 2022. 3

[21] Pamela Samuelson. Ongoing lawsuits could affect everyone who uses generative AI. *Science*, 381:6654, 2023. 8

[22] 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. 3

[23] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18. IEEE, 2017. 2

[24] Stuart A. Thompson. We asked A.I. to create the Joker. It generated a copyrighted image. *The New York Times*, 2024. 3, 7

[25] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal*, 32 (4):791–813, 2023. 1

[26] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. arXiv:2210.00968, 2022. 2, 7

[27] Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. arXiv:2405.14800, 2024. 7

[28] Minxing Zhang, Ning Yu, Rui Wen, Michael Backes, and Yang Zhang. Generated distributions are all you need for

membership inference attacks against generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4839–4849, 2024. 2