# Measuring the Success of Diffusion Models at Imitating Human Artists

**Stephen Casper** [*1] **Zifan Guo** [*1]
**Shreya Mogulothu** [1] **Zachary Marinov** [1] **Chinmay Deshpande** [2] **Rui-Jie Yew** [1 3] **Zheng Dai** [1]
**Dylan Hadfield-Menell** [1]

## Overview

Modern diffusion models have set the state-of-the-art in AI image generation. Their success is due, in part, to training on Internet-scale data which often includes copyrighted work. This prompts questions about the extent to which these models learn from, imitate, or copy the work of human artists.

This work suggests that questions involving copyright liability should factor in a model's *capacity* to imitate an artist. Tying copyright liability to the capabilities of the model may be useful given the evolving ecosystem of generative models. Specifically, much of the legal analysis of copyright and generative systems focuses on the use of protected data for training (Sag, 2018; Lemley & Casey, 2020). However, generative systems are often the result of multiple training processes. As a result, the connections between data, training, and the system are often obscured.

In our approach, we consider simple image classification techniques to measure a model's ability to imitate specific artists. Specifically, we use Contrastive Language-Image Pretrained (CLIP) (Radford et al., 2021) encoders to classify images in a zero-shot fashion. Our process first prompts a model to imitate a specific artist. Then, we test whether CLIP can be used to reclassify the artist (or the artist's work) from the imitation. If these tests match the imitation back to the original artist, this suggests the model can imitate that artist's expression.

Our approach is simple and quantitative. Furthermore, it uses standard techniques and does not require additional training. We demonstrate our approach with an audit of Stable Diffusion's (Rombach et al., 2022) capacity to imitate 70 professional digital artists with copyrighted work online. When Stable Diffusion is prompted to imitate an artist from this set, we find that the artist can be identified from the imitation with an average accuracy of 81.0%. Finally, we also show that a sample of the artist's work can be matched to these imitation images with a high degree of statistical reliability. Overall, these results suggest that Stable Diffusion is broadly successful at imitating individual human artists. Code is available here.

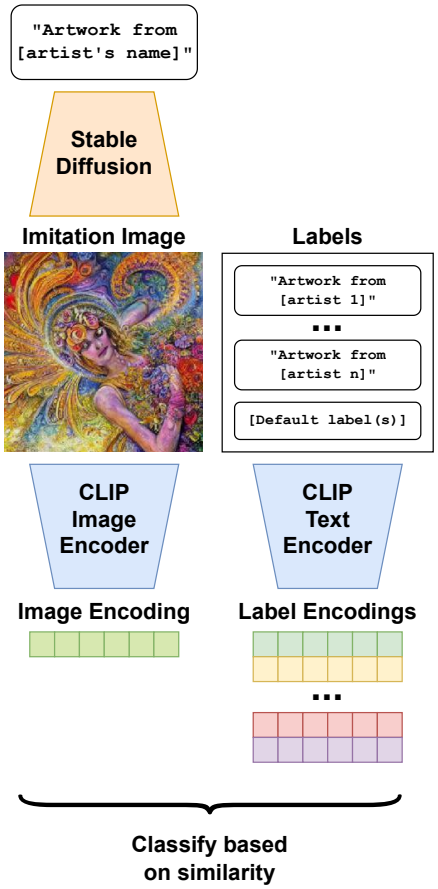


*Figure 1.* **Identifying human artists from Stable Diffusion Imitations.** For each artist, we generate an imitation image from Stable Diffusion with the prompt "Artwork from < artist name >." Next, we encode the image with a CLIP image encoder (Radford et al., 2021). We also encode labels corresponding to $n$ total artists plus one or more 'default' labels with a CLIP text encoder. Finally, we classify the image among all labels using a geometric similarity measure between the encodings. If the label reliably corresponds to the correct artist, we consider the model to have the capability to imitate that artist.

[*]Equal contribution [1]MIT [2]Harvard University [3]Brown University. Correspondence to: Stephen Casper <scasper@mit.edu>.

Accepted to the *1st Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA. 2023. 

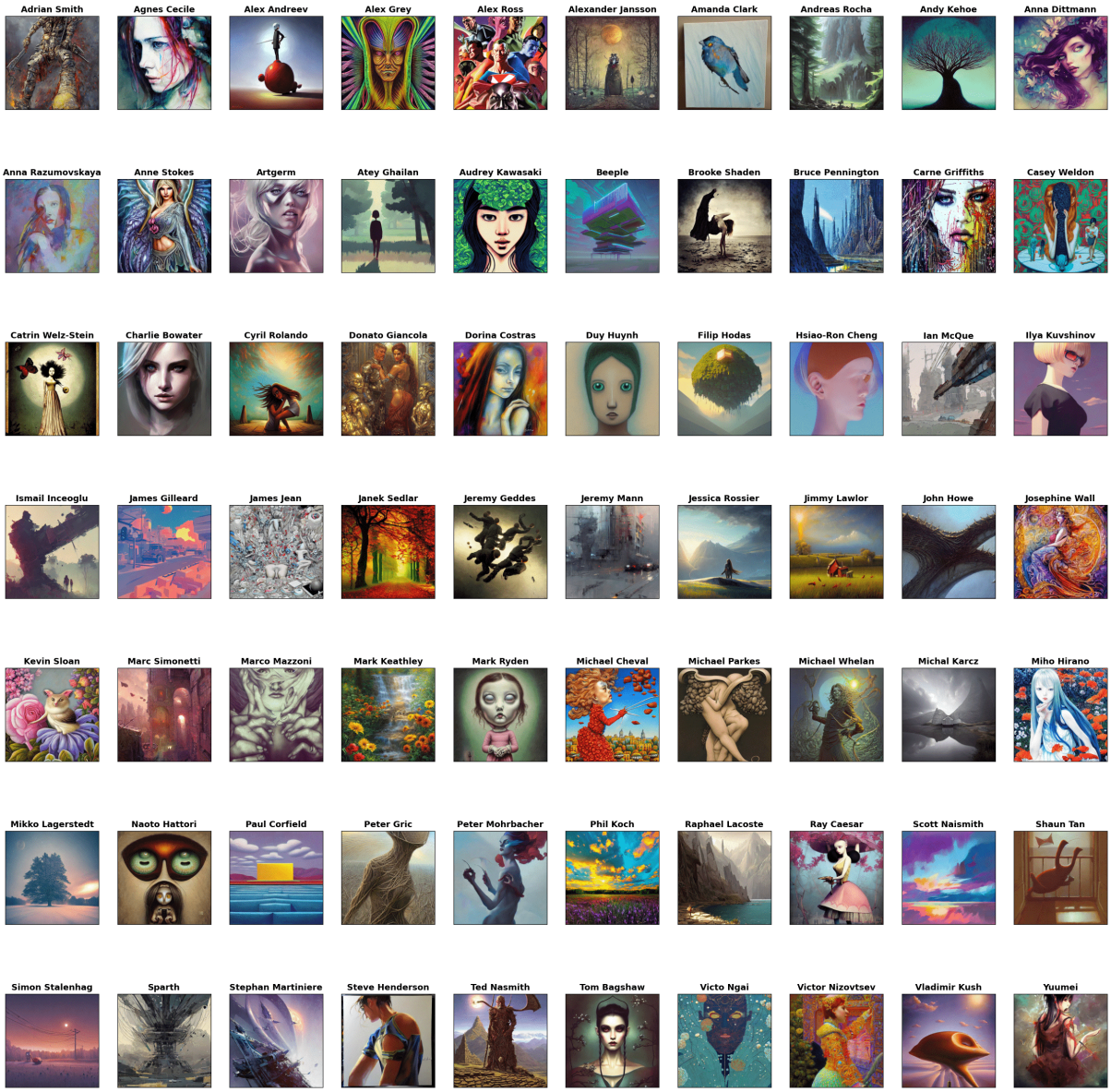


*Figure 2.* **Example images generated by Stable Diffusion from prompts of the form "Artwork from <artist's name>".** Using the method depicted in Figure 1, we show that the artists used in the prompts can often be classified from these imitations of their work.

## 1. Background

**Contrastive Language-Image Pretraining (CLIP):** CLIP (Radford et al., 2021) is a technique for training AI systems that encode images and text into fixed-length vector representations. CLIP image and text encoders are trained to produce similar encodings of image/caption pairs and dissimilar encodings of image/caption non-pairs. The more geometrically distant two encodings of images or captions are, the less related they are according to the encoder, and vice versa. Using this principle, Radford et al. (2021) introduced a method to classify an image among a set of labels based on the distances between encodings. We use this method in our proposed test.

**Diffusion Models:** Diffusion models (Sohl-Dickstein et al., 2015) such as Stable Diffusion (Rombach et al., 2022) and Midjourney (Midjourney, 2022), are capable of generating images from arbitrary, user-specified prompts. Their success has largely been due to training on large amounts of text/image data, often including copyrighted works (Schuhmann et al., 2021). Modern image-generation diffusion models are trained using CLIP-style encoders. When given an encoding of a caption, a diffusion model is trained to generate an image corresponding to the caption (Ramesh et al., 2022). Accordingly, a diffusion model that generates

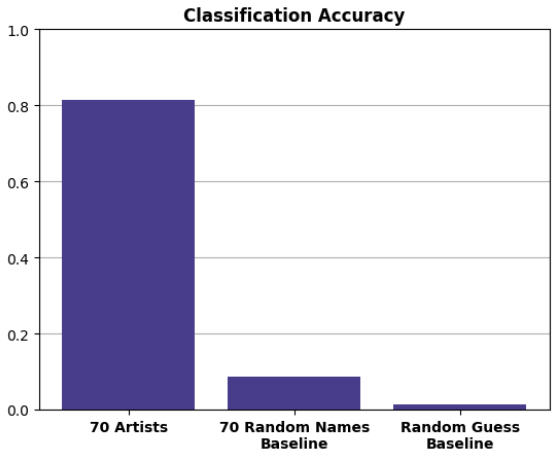


*Figure 3.* **Results from human artists strongly outperform baselines.** Success rates for classifying artists from Stable Diffusion's attempts to imitate them. Professional artists can be classified from imitations over 81.0% of the time on average. This compares to 8.6% for a baseline in which we used random names instead of artists' names and 1.4% for a random guess baseline.

images from these embeddings is trained to be the inverse of a CLIP image encoder.

**Legal Motivation:** In the United States, *Newton v. Diamond*, 388 F.3d 1189, 1195 (9th Cir. 2004) established that copyright infringement "is measured by considering the qualitative and quantitative significance of the copied portion in relation to the plaintiff's work as a whole". However, the subjective nature of these determinations makes practical enforcement complicated. (Balganesh et al., 2014; Kaminski & Rub, 2017; Balagopalan et al., 2023). In evaluating copyright questions involving AI systems, legal analyses have focused on how copyrighted work is used in the system's training data (Sag, 2018; Lemley & Casey, 2020), but such a focus on training data does not connect liability to an AI system's ability to copy an artist. In contrast, we show how standard image classification techniques can be used to help determine how successful AI image generators are at imitating individual human artists. This approach is *consistent*, *quantitative*, and connected to the *capabilities* of the resulting AI system. Our goal, however, is not to automate determinations of infringement but to demonstrate how tried and tested image classification techniques from machine learning can be used to analyze legal claims.

## 2. Experiments

We conduct two complementary experiments to evaluate Stable Diffusion's ability to imitate human artists. First, we classify human artists from imitations of their work, and second, we match real work from human artists to imitations. Both experiments suggest that Stable Diffusion is broadly successful at imitating human artists.

### 2.1. Identifying Artists from Imitations

**Method:** We used CLIP encoders to classify artists from Stable Diffusion's imitations of them. We selected 70 artists from the LAION-aesthetics dataset (Schuhmann et al., 2021), the dataset used to train Stable Diffusion. We selected these 70 as artists who may potentially be harmed by digital imitations using several criteria: each artist is alive, has a presence on digital art platforms (Instagram, DeviantArt, and ArtStation), publishes artwork or sells their artwork (e.g., prints or digital works), and has more than 100 images in the LAION dataset.

Figure 1 outlines our method. We prompted Stable Diffusion (v1.5) to generate images in the style of each artist, using prompts of the form "Artwork from <artist's name>". Example images are in Figure 2. We then used CLIP encoders to classify each image among a set of 73 labels. The 73 labels consisted of each of the 70 artist's prompts ("Artwork from <artist's name>") plus three default labels: "Artwork", "Digital Artwork", and "Artwork from the public domain." These additional labels lend insight into how confident CLIP is that an image imitates a particular artist's style instead of some more generic style. We then classified each imitation image among these labels using the technique from Radford et al. (2021). CLIP-based classification produces a probability of an image matching each label, and we evaluate the model on the correctness of its most-likely prediction and confidence in the correct artists.

**Results:** We repeated the experiment with the 70 artists ten times to reduce the effect of random variation. On average, CLIP correctly classified 81.0% of the generated images as works made by artists whose names were used to generate them. Over the ten trials, 69 of the 70 artists were correctly classified in a plurality of the ten trials. Overall, these results suggest that Stable Diffusion has a broad-ranging ability to imitate the styles of individual artists. We compared these results to two baselines. First, we implemented a random-name baseline by running the same experiment with 70 random names from a random name generator. Since Stable Diffusion was not trained on artists with these names (unless a random name is coincidentally the same as some artist's), this experiment serves as a proxy for how Stable Diffusion would handle artists not in its training data. In this case, only 6 names (8.6%) were guessed correctly. Second, a random guess would only result in a successful classification every 1 in 73 attempts (1.4%) on average. We visualize results from our main experiment alongside the controls in Figure 3.

**Results are Robust to Different Sets of Artists:** To test whether our 70 artists were especially classifiable, we ran the original experiment but with a larger set of indiscriminately-selected artists and found similar results. We selected the 250 artists with the highest number of images in the LAION dataset and found that CLIP correctly classified 81.2% of

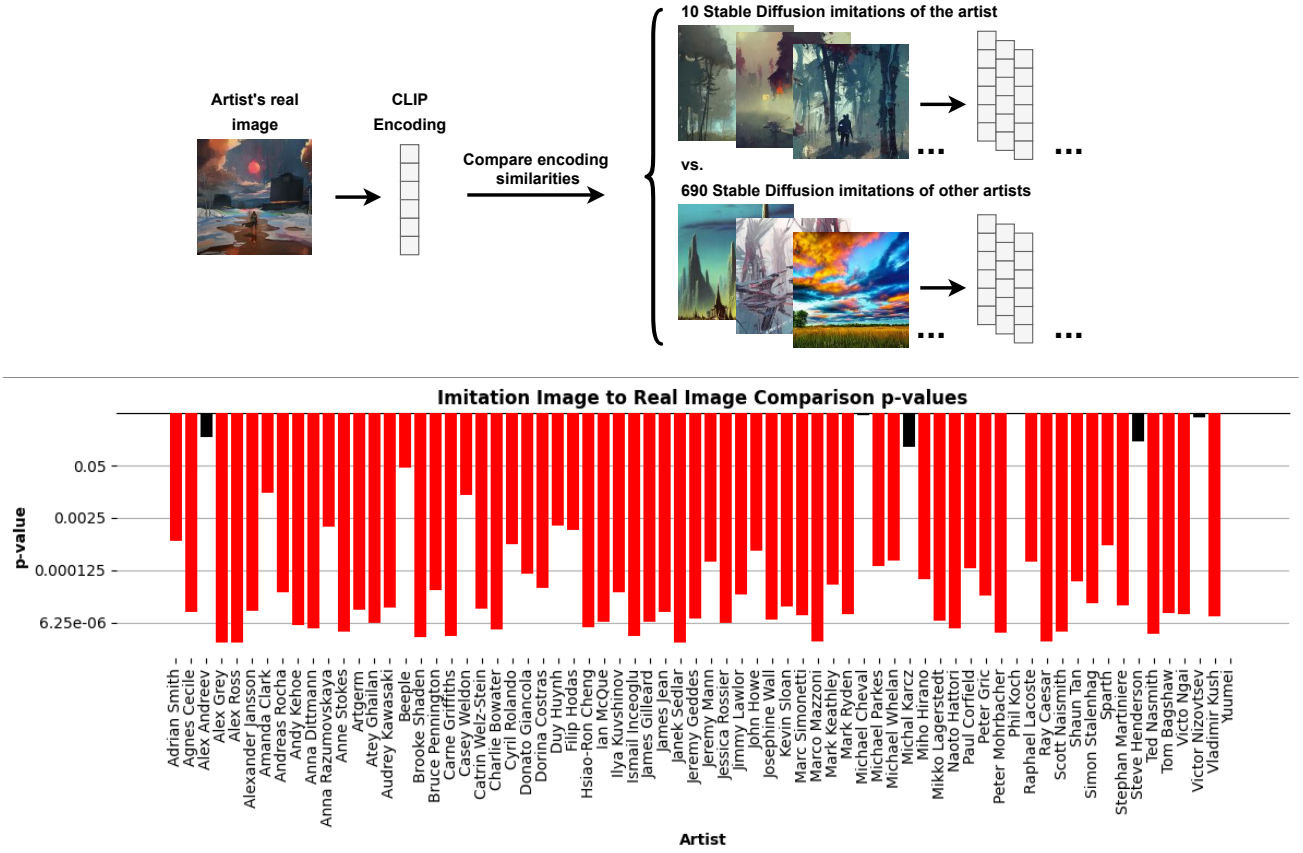


*Figure 4.* **Matching human artwork to imitations from Stable Diffusion:** (Top) Method: We compared the encoding of one real image per artist to 10 imitations of that artist and 690 imitations of the other artists. Then we aggregated these results with a statistical rank sum test. (Bottom) Results: Artwork generated by Stable Diffusion with the prompt "Artwork from <artist's name>" is significantly more similar to real artwork by the artist in question than artwork generated to imitate other artists. Artists for which the experiment resulted in a (Bonferroni-corrected) p-value below 0.05 are highlighted in red. This occurred for 90% (63/70) of the artists.

the images. This demonstrates that successful classification transcends a particular specific set of artists.

### 2.2. Matching Artwork to Imitations

**Method:** Our first experiment tested how easily artists could be identified from diffusion model imitations of them. To provide a complementary perspective, we also directly study the similarity of artists' digital works to Stable Diffusion's imitations of them. For each of the 70 artists, we retrieve the top result obtained by Google Image searching "<artist's name> art." As before, we then use Stable Diffusion to generate 10 images for each artist with the prompt "Artwork from [artist's name]." We then compare the real images and generated images. Distances are measured by first encoding images using the CLIP image encoder and calculating the cosine distance between encodings.

**Results:** For each artist, we calculate whether real images from artists are more similar to imitations of that artist or other artists. The significance was calculated using a rank sum test with a Bonferroni correction factor of 70. Results are in Figure 4. 90% (63/70) of the experiments produce $p$ values less than 0.05. This compares to an average of 22.8% (16/70) for a control experiment using random artist assignments of real images. These results further support that Stable Diffusion is broadly successful at imitating artists.

## 3. Conclusion

We have demonstrated how AI image classification can help to measure the success of diffusion models imitating human artists. We argue that these methods can provide a practical way to tie questions about copyright liability to the *capabilities* of a model instead of its training data alone. By matching imitation images to both artists' names and works, we find that Stable Diffusion is broadly successful at imitating human digital artists. We hope that future work can use image classification to analyze legal claims and to test defenses against AI imitation of copyrighted work.

## Acknowledgements

We thank Taylor Lynn Curtis and Lennart Schulze for feedback.

## References

Balagopalan, A., Madras, D., Yang, D. H., Hadfield-Menell, D., Hadfield, G. K., and Ghassemi, M. Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. *Science Advances*, 9(19):eabq0701, 2023.

Balganesh, S., Manta, I. D., and Wilkinson-Ryan, T. Judging similarity. *Iowa L. Rev.*, 100:267, 2014.

Kaminski, M. E. and Rub, G. A. Copyright's framing problem. *UCLA L. Rev.*, 64:1102, 2017.

Lemley, M. A. and Casey, B. Fair learning. *Tex. L. Rev.*, 99: 743, 2020.

Midjourney. Midjourney, 2022. URL https://www.midjourney.com/.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Sag, M. The new legal landscape for text mining and machine learning. *J. Copyright Soc'y USA*, 66:291, 2018.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

*Newton v. Diamond*, 388 F.3d 1189, 1195 (9th Cir. 2004).