Section 1 - On the topic of alleged copyright infringement occurring when a generative AI model is trained and made available to the public

As both an artist and a Bachelor of Computer Science with a background in Artificial Intelligence, it is my opinion that the training of a generative AI model on publicly accessible data (eg. Images scraped from the internet) does not constitute copyright infringement. Rather, the information contained in these models consists either of uncopyrightable "ideas" or else ought to be considered *de minimis* use of otherwise copyrightable expressive content.

I will argue for the above claim with reference to the popular AI art generator *Stable Diffusion*, as it is the generative AI application with which I am most familiar, but it is very likely that my arguments in defense of Stable Diffusion are also applicable to other forms of generative AI.

1A - The intent of generative AI training; a refutation of claims that all training set images are "stored" in the AI model

The intent of training a generative AI model is not to compress millions of images for the sake of reproducing them at-will. In fact, such a goal is physically impossible, as explained by the Electronic Frontier Foundation in their article on the subject:

> ```
> > The Stable Diffusion model makes four gigabytes of observations regarding more
> than five billion images. That means that its model contains _less than one
> byte_ of information per image analyzed (a byte is just eight bits—a zero or a
> one).
> > The complaint against Stable Diffusion characterizes this as “compressing”
> (and thus storing) the training images, but that’s just wrong. With few
> exceptions, there is no way to recreate the images used in the model based on
> the facts about them that are stored. Even the tiniest image file contains _many
> thousands_ of bytes; most will include _millions_. Mathematically speaking,
> Stable Diffusion _cannot_ be storing copies of all of its training images...
> ```

1B - Applying the Idea/Expression distinction to the contents of generative AI models

Stable Diffusion does not (and cannot) "search" a compressed database of copyrighted images to produce an output; In the overwhelming majority of cases, a generative AI model cannot reproduce its specific training images it at all. Instead, the data which makes up a Generative AI model consists of mathematical functions describing commonalities in relationships between patterns of pixels in an image and the textual metadata associated with that image (in other words, it stores non - copyrightable **ideas** about how visual and artistic concepts tend to be described, rather than the expressive elements of specific, copyrighted images); While the model can be prompted to output works which bear some stylistic similarities to one or more of the works included in the training set, the copyrightable expressive features from the training images will, in almost all cases, not be present in the output.

For more information on this process, please refer to the attached excerpt from the article "Stable Diffusion copyright lawsuits could be a legal earthquake for AI" published on Ars Technica by Timothy B. Lee on 4/3/2023.

1C - A preemptive defense of the above claims; On the topic of Stable Diffusion replicating expressive content in output images

In response to the above claim, a common criticism is that Stable Diffusion *is* capable of reproducing one or more specific images from its training set, and therefore must contain compressed copies of *all* images in its training set.

It is my opinion that this copying, while theoretically possible, is *de minimis* in practice and not sufficient evidence to support the claim that generative AI models are themselves copyright infringing derivative works (nor is it sufficient evidence that the generative AI model merely recombines whole fragments of existing copyrighted images to produce its outputs).

1D - Experimental evidence demonstrates extremely low incidence for direct replication of training images;

To support the above claim, I cite the research paper "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models", in which authors Somepalli *et al* made targeted attempts to induce a Stable Diffusion model to reproduce specific training images by prompting it with those images' associated metadata tags (the same specific tags which would have been used during the training process to "teach" the model about the relationship between the contents of the image and natural language descriptions of the image).

They observed that significant replication of expressive details from the training image occurred in 1.88% of their 9000 generated images. If a Stable Diffusion model contained compressed copies of all the images used train it, then it would be reasonable expect a much higher rate of copying under these conditions.

Furthermore, it is worth noting that:

1. Not all of the reproduced images were protected by copyright (indeed, public domain images are *more likely* to be replicated than copyrighted images, for reasons which are explained in the following section 1E), so the rate of *infringing* replication ought to be much lower than 1.88%.
2. The test conditions described in this paper were specifically optimized to maximize the rate of the rate of training image replication: in practice, the incidence of replication will be much lower for the average Stable Diffusion user.

1E - Description of the technical causes for training image replication

Furthermore, such replications are incidental consequences of the method used to acquire training data and train models: Replication of training images is the result of a phenomenon called *overtraining*, which occurs when redundant copies of a single image are included in a training set, causing the resulting model to build too strong of an association between the specific expressive elements of that image and the text associated with it. An example of this is cited by the EFF:
[Training a Machine Learning Model] is not too different from how babies learn things. For example, a lot of kids basically think all animals are "doggies" until they have enough exposure and correction by adults to distinguish "doggie" from "horsie." Machine learning can make similar mistakes, finding connections that, to humans, are obscure. For example, a cancer classifier can “learn” that an image shows a tumor if that image contains a ruler. The AI learned a shortcut: images of structures that a radiologist has identified as cancerous tumors have pictures with rulers for scale and to track size. The training images of benign growths were from a different set, and they didn’t have rulers.

The important thing to note is that this is considered an *error* in machine learning development, not a desired or intended function. Developers and users of generative AI generally have no interest in merely reproducing a vague approximation of a pre-existing image (if they merely wished to acquire copyrighted images and redistribute them *en masse* via the internet, there are much simpler methods than training a generative AI model).

Rather, generative AI developers have a vested interest in working to *prevent* overtraining (and, by extension, to prevent their models from reproducing significant expressive elements of training set images) for the simple reason that it improves the quality of the models' outputs. Consequently, the already low incidence of training image duplication will only decrease over time as these technologies are further refined.

In my opinion, these factors ought to be sufficient grounds for considering the copyrighted material stored in generative AI models to be *de minimis* usage.

Section 2 - On the topic of alleged copyright infringement occurring when producing AI generated outputs

A perspective of some generative AI critics (apart from or in addition to claims of copyright infringement caused by training a generative AI model) is that the outputs produced by a generative AI model are themselves copyright infringement. The claim seems to be: Because the "contents" of a generative AI model are obscure, and the model is *capable* of producing potentially infringing outputs, there is no way of knowing with certainty whether or not any given output contains significant portions of copyrighted expression. Therefore, it is assumed that *any and all* outputs of the model are copyright infringement unless proven otherwise to their satisfaction (which, owing to the need to prove a negative, tends to be impossible). This leads to a (mis)characterization of AI generators as "automated collage machines", which function by piecing together whole chunks of preexisting copyrighted works to give the illusion of creating a unique image.

2A - A tool's capacity for copyright infringement does not imply that infringement is inherent to the tool

In cases of copyright infringement, there is a requirement on the part of the copyright holder to demonstrate a *significant similarity* between the allegedly infringing image and the alleged original. In most cases of alleged infringement involving AI generated content, there is no attempt made on the part of the copyright holder (or individuals claiming to act on their behalf) to identify the significant expressive elements which they believe to have been copied. Rather, they tend to either cite non-specific, on non-copyrightable stylistic similarities, or else the mere fact that generative AI was used *at all* to produce an image is cited as evidence of copyright infringement

In so doing, critics of generative AI are making a common mistake: They observe that a tool has the capability to do harm, and so assume that this harm is inherent to the tool. However, this is not how copyright infringement has historically been understood: Traditional art instruments such as pencils and paint brushes, too, have the capacity to reproduce copyrighted images, as do numerous pieces of computer software for image manipulation; But it would be absurd to suggests that such properties are inherent features of these tools and demand that every user of pencils, paint or *Adobe Photoshop* prove that they aren't infringing when they produce a work of art. So too it ought to be with AI generated images: If an AI model produces a demonstrably infringing image (by accident or through malicious user action), then that ought to be evidence of ignorant or malicious use of an otherwise legally neutral tool, not evidence that the tool itself is malicious.

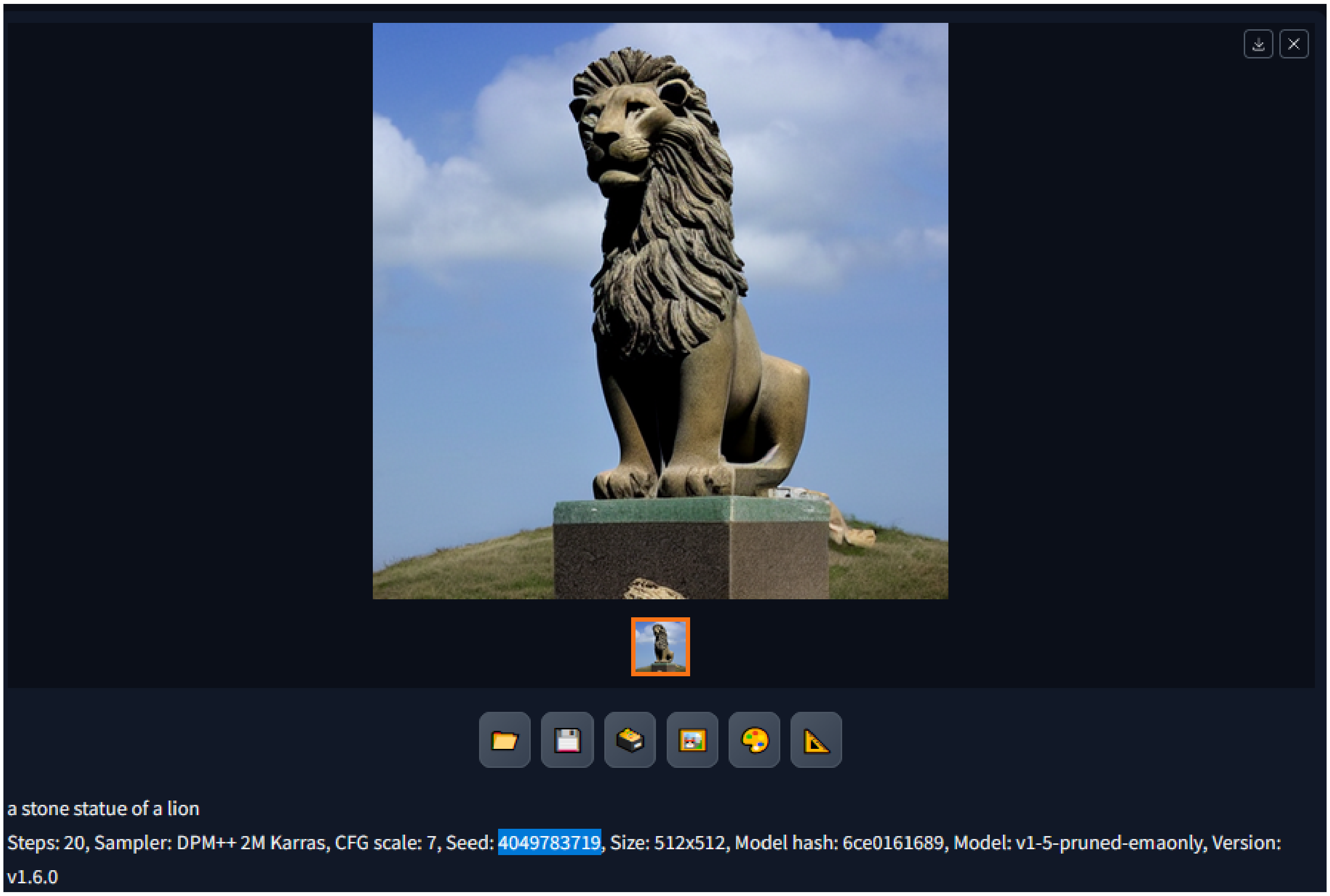
Section 3 - On the topic of generative AI and human authorship

The **Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence** denies copyright to all AI generated works, citing the *Compendium of U.S. Copyright Office Practices* sec. 313.2 (3d ed. 2021) (“*Compendium (Third)*”) which reads that the Office "will not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author."

I consider this ruling to grossly overstate the AI model's capacity for agency and independent decision-making, as well as to grossly understate the role of the human operator's own creative judgment in producing the image. In this section I intend to demonstrate why the operation of generative AI tools is neither random, automatic, nor devoid of creative input or intervention from a human author.

Section 3A - On "Random operation": A demonstration of the effect of prompt changes when randomness is removed from the image generation process

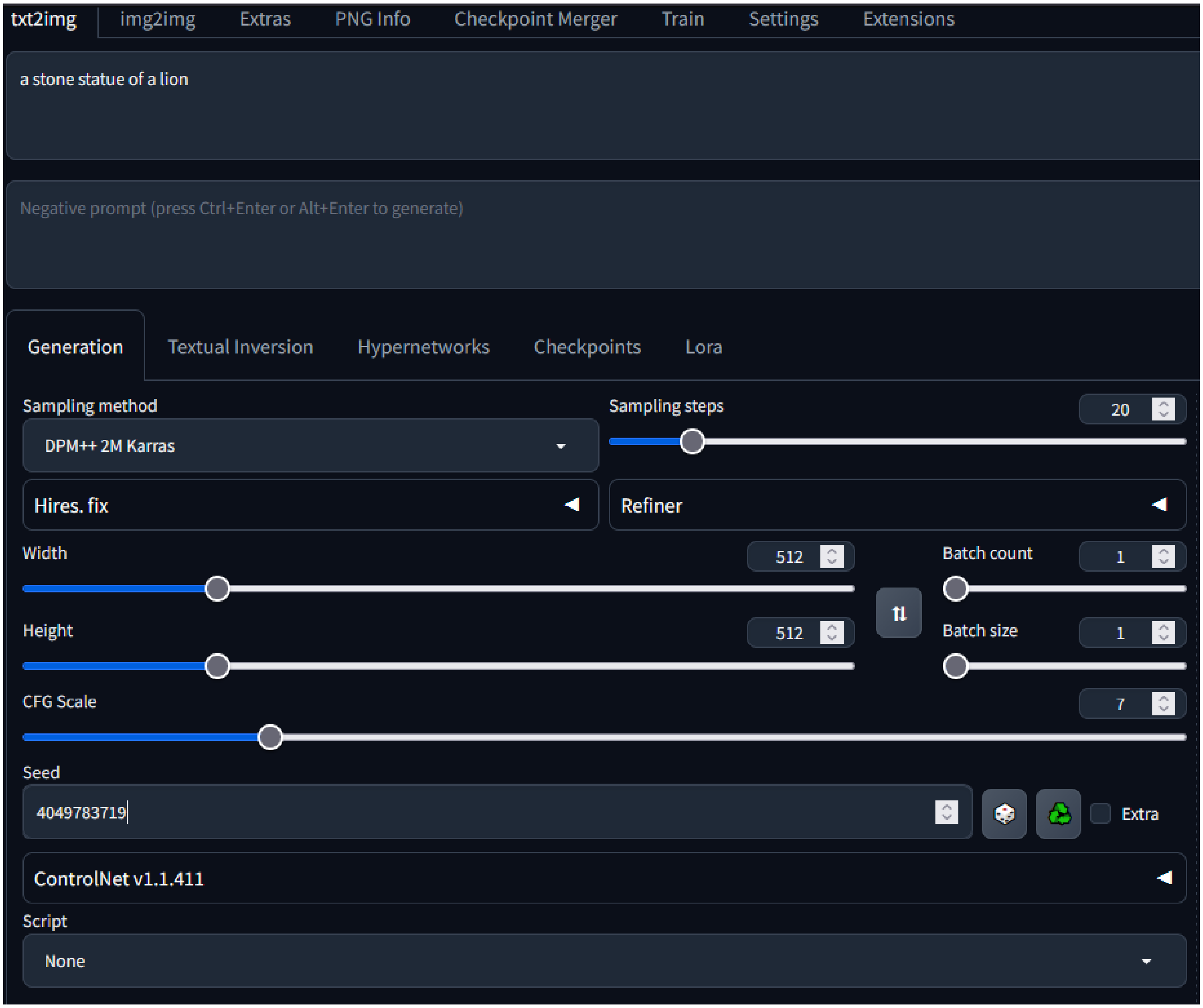
To refute the claim of "random operation", one need only to demonstrate that randomness is not an essential element of the generation process: Stable Diffusion, like other AI image generators, uses a pseudo-random integer value as a "seed" to determine the initial distribution of visual "noise" which is resolved by the model into an image. Stable Diffusion presents this seed number to the user along with the generated image, so that if the user re-enters the same seed value when generating a new image, the resulting image will be identical (assuming that no other changes were made):



a stone statue of a lion

Steps: 20, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 4049783719, Size: 512x512, Model hash: 6ce0161689, Model: v1-5-pruned-emaonly, Version: v1.6.0

> A screenshot of a Stable Diffusion UI displaying the seed value for an image generated using the prompt "a stone statue of a lion"



a stone statue of a lion

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Sampling method: DPM++ 2M Karras

Sampling steps: 20

Width: 512

Height: 512

Batch count: 1

Batch size: 1

CFG Scale: 7

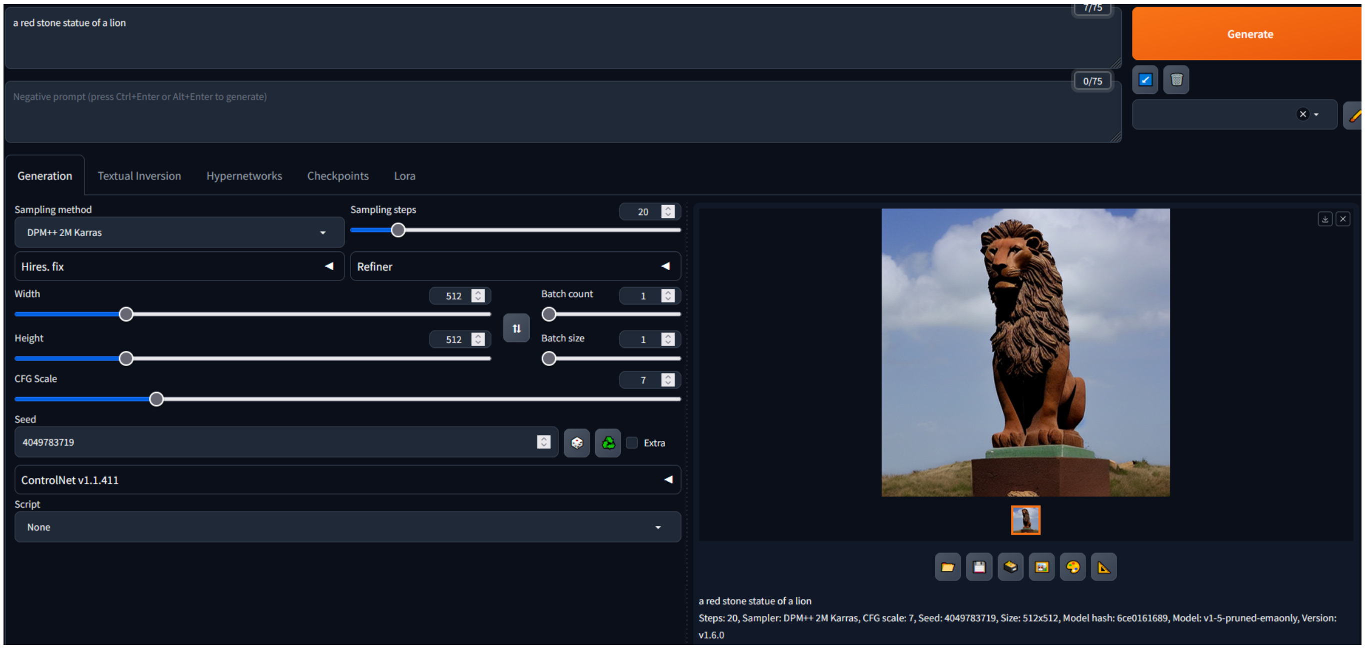
Seed: 4049783719

ControlNet v1.1.411

Script: None

> A screenshot of the same Stable Diffusion UI displaying the field where a seed can be manually entered

In this way, a user can remove the random element from AI image generation altogether. This technique is frequently used to reproduce specific images generated by other users, but it can also be used to demonstrate the non-trivial effect that the content of the user's prompt has on the output. For example:

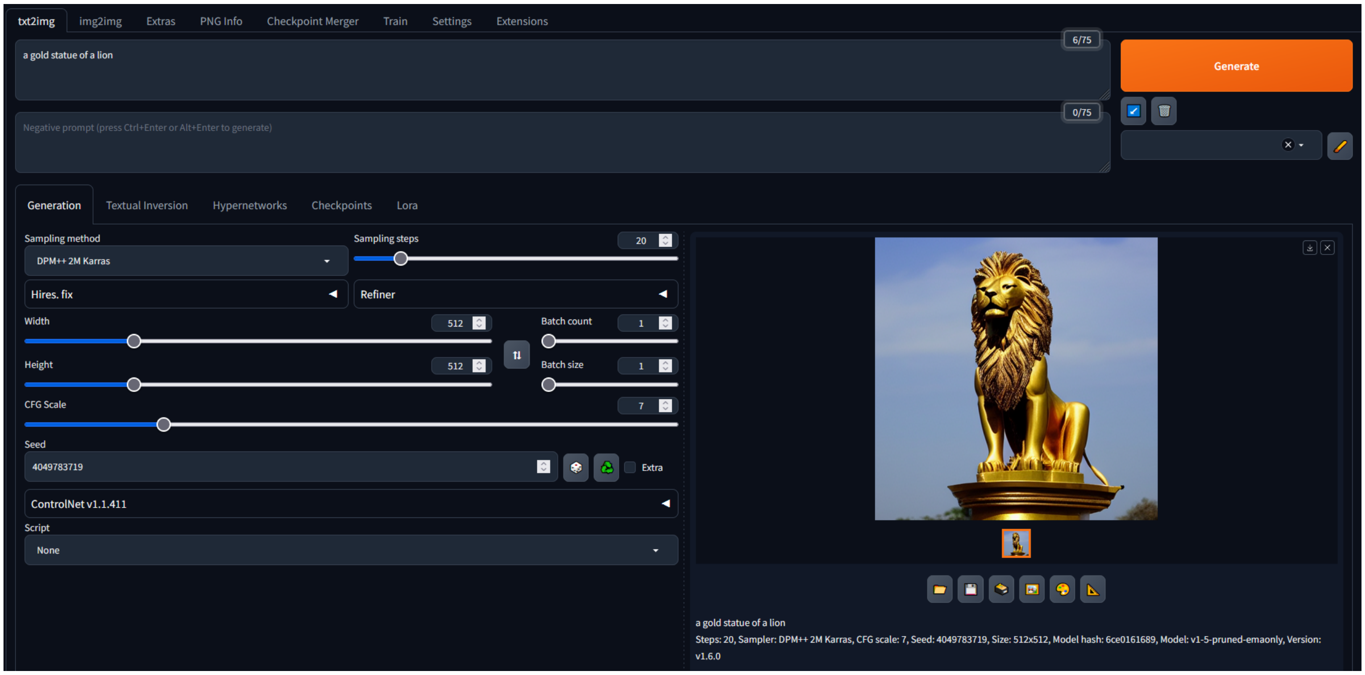


a red stone statue of a lion

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Sampling method: DPM++ 2M Karras
Sampling steps: 20
Width: 512
Height: 512
Batch count: 1
Batch size: 1
CFG Scale: 7
Seed: 4049783719
Script: None

a red stone statue of a lion
Steps: 20, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 4049783719, Size: 512x512, Model hash: 6ce0161689, Model: v1-5-pruned-emaonly, Version: v1.6.0

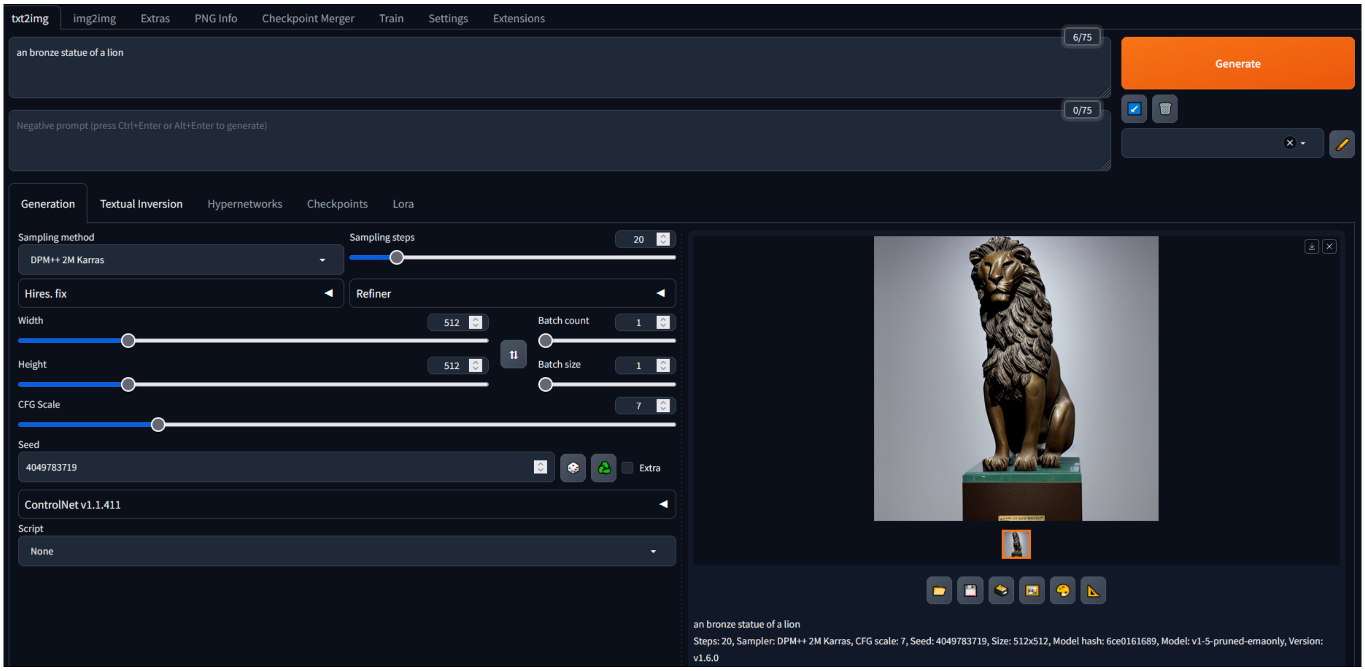


a gold statue of a lion

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Sampling method: DPM++ 2M Karras
Sampling steps: 20
Width: 512
Height: 512
Batch count: 1
Batch size: 1
CFG Scale: 7
Seed: 4049783719
Script: None

a gold statue of a lion
Steps: 20, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 4049783719, Size: 512x512, Model hash: 6ce0161689, Model: v1-5-pruned-emaonly, Version: v1.6.0



an bronze statue of a lion

6/75

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

0/75

Sampling method: DPM++ 2M Karras

Sampling steps: 20

Hires. fix

Refiner

Width: 512

Height: 512

Batch count: 1

Batch size: 1

CFG Scale: 7

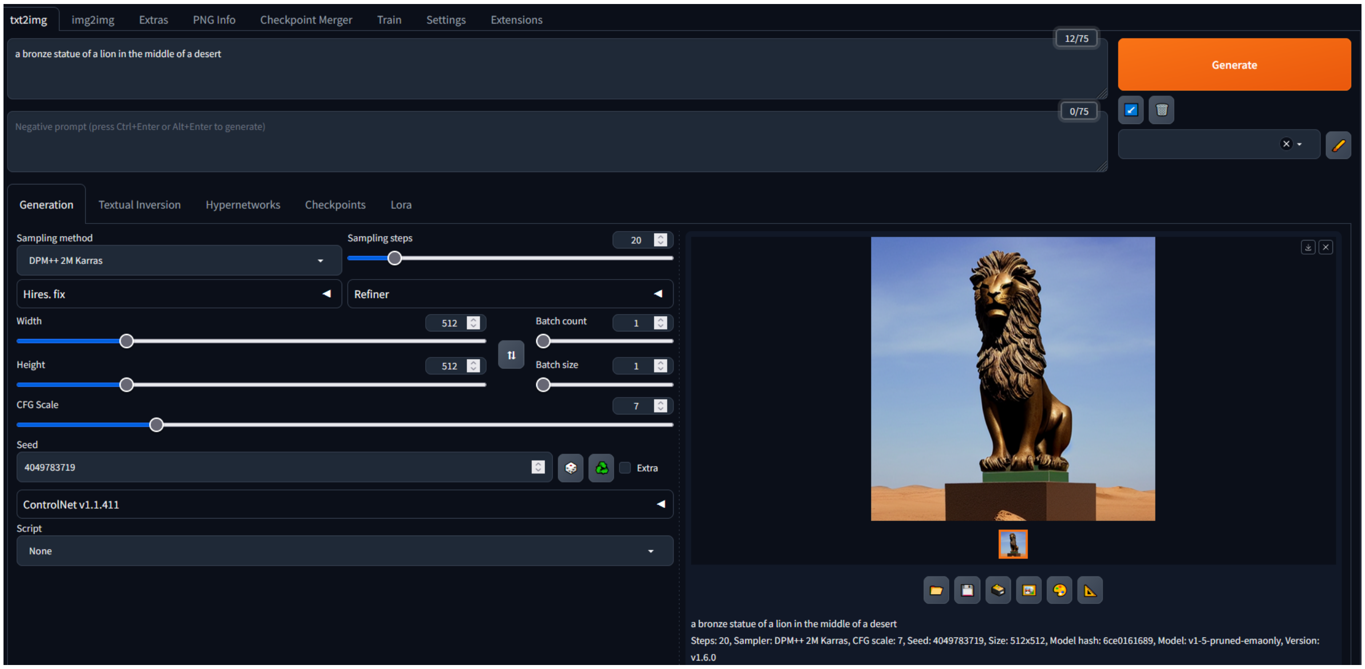
Seed: 4049783719

ControlNet v1.1.411

Script: None

an bronze statue of a lion

Steps: 20, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 4049783719, Size: 512x512, Model hash: 6ce0161689, Model: v1-5-pruned-emaonly, Version: v1.6.0



a bronze statue of a lion in the middle of a desert

12/75

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

0/75

Sampling method: DPM++ 2M Karras

Sampling steps: 20

Hires. fix

Refiner

Width: 512

Height: 512

Batch count: 1

Batch size: 1

CFG Scale: 7

Seed: 4049783719

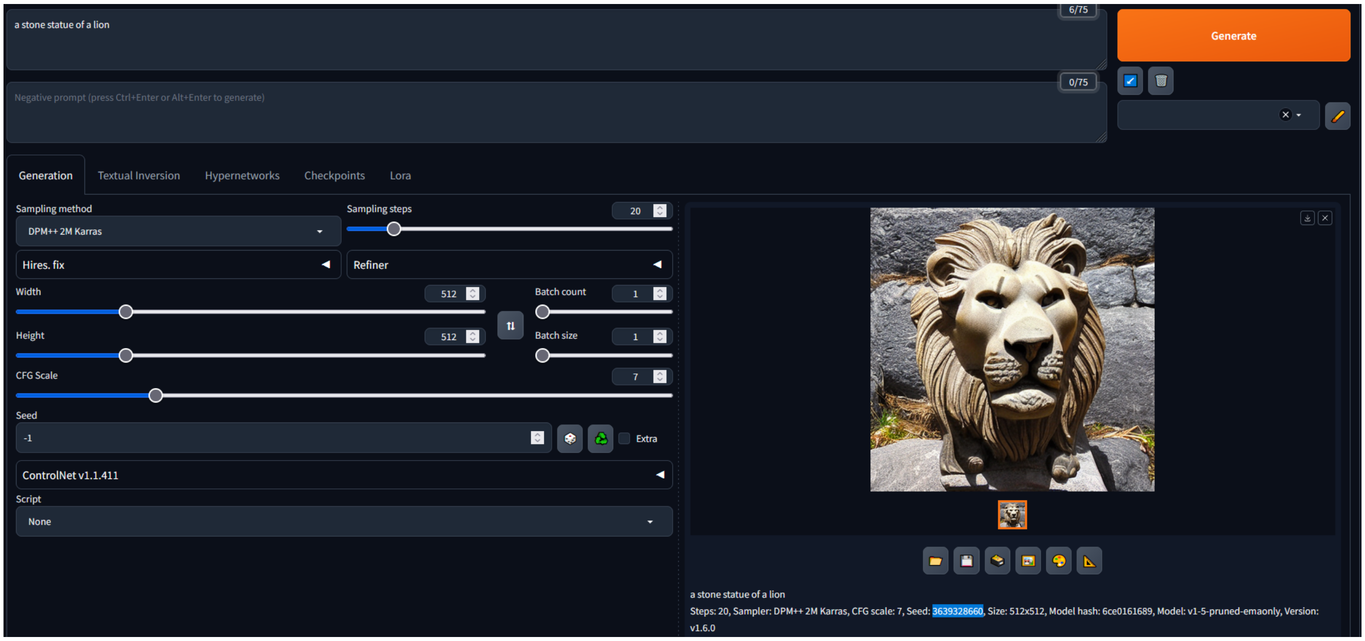
ControlNet v1.1.411

Script: None

a bronze statue of a lion in the middle of a desert

Steps: 20, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 4049783719, Size: 512x512, Model hash: 6ce0161689, Model: v1-5-pruned-emaonly, Version: v1.6.0

> Changing the prompt while keeping the same seed results in variations on the same underlying composition.



a stone statue of a lion

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Sampling method: DPM++ 2M Karras

Sampling steps: 20

Width: 512

Height: 512

Batch count: 1

Batch size: 1

CFG Scale: 7

Seed: -1

Script: None

a stone statue of a lion
Steps: 20, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 3639328660, Size: 512x512, Model hash: 6ce0161689, Model: v1-5-pruned-emaonly, Version: v1.6.0

> Changing the seed without changing the prompt results in a drastically different composition.

In *Burrow-Giles Lithographic Co. v. Sarony*, the court ultimately ruled that "[the] plaintiff made the [photograph]… entirely from his own original mental conception, to which he gave visible form by posing the said Oscar Wilde in front of the camera, selecting and arranging the costume, draperies, and other various accessories in said photograph, arranging and disposing the light and shade, suggesting and evoking the desired expression, and from such disposition, arrangement, or representation, made entirely by plaintiff, he produced the picture in suit."

By using a static seed value, the influence of the operator's prompt can clearly be seen: Far from being a purely random process, the prompt has a significant effect on the contents of the output, making the composition of the prompt functionally equivalent to the "disposition, arrangement, or representation" of visual elements which lead the court to rule in the plaintiff's favour in Burrow-Giles Lithographic Co._ v. _Sarony.

Section 3B: On "Automatic operation": Generative AI models have no independent agency

The requirement to provide a prompt to a generative AI model might, on a superficial level, appear to be analogous to a commissioner providing a work-for-hire artist with a description of their desired end product. However, the analogy is not appropriate for the simple reason that a human artist possesses independent agency, while the generative AI model does not. In the absence of commissioners to provide work, an artist is capable of independently conceiving of new creative projects, can express them in their chosen medium without any outside input, and can independently assert their ownership of the finished work.

By contrast, a generative AI model can take *no action at all* in the absence of a human operator. If no prompter exists to provide an instance of Stable Diffusion (or any other generative AI) with a prompt, then the application will simply do nothing. It cannot act independently to produce works from its own imagination (it wholly lacks an imagination, being essentially a ~4GB block of linear algebra functions), and it has no capacity to assert ownership over the works that it *does* produce (though this has evidently not stopped human operators from confusing the issue by attempting to assert ownership on the model's behalf).

Simply put, the process of generating an image using an AI model is in no sense "automatic": Everything that the model produces is only done so at the behest of a human operator.

Section 3C - On "operation without any creative input or intervention from a human author": Curation of generated images is an act of significant human contribution

As a generative AI model is incapable of independent action, it follows that the model is fundamentally incapable of judging the quality of a generated work: the model has no capacity to determine if any given generated image is "good" or "bad" and make corrections. It will generate the image which corresponds to the seed value and the prompt provided to it, no more and no less.

The Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence quotes the U.S. Copyright Office, *Sixty-Eighth Annual Report of the Register of Copyrights for the Fiscal Year Ending June 30, 1965*, at 5 (1966)) in asking, "whether the ‘work’ is basically one of human authorship, with the computer [or other device] merely being an assisting instrument, or whether the traditional elements of authorship in the work (literary, artistic, or musical expression or elements of selection, arrangement, etc.) were actually conceived and executed not by man but by a machine."

The phrase "elements of selection" clearly stands out here: If selection is considered by the Office to be a "traditional element of authorship", then it is an element which fundamentally *cannot* be "conceived and executed" the AI algorithm. Such assessments of quality are solely the domain of the human operator, whose judgment will be informed by their own understanding of artistic concepts (composition, colour, lighting, etc), their own assessment of the model's accuracy in portraying the desired subject, as well as other subjective criteria. In this way, the curation of generated images must *itself* be an act of human authorship.