Destroying the Previous Original: A Trip Report from Latent Space

Assistant Professor Chad Eby, M.A., M.F.A.
School of Art and Visual Studies, University of Kentucky, Lexington, United States of America
https://chadeby.studio
e-mail: chad.eby@uky.edu

.

Abstract

Stable Diffusion is an open-source generative artificial neural network capable of detailed text-to-image (TTI) and image-to-image (ITI) generation. Publicly distributed models and user interfaces may be run locally on personal computers with dedicated graphic cards, making Stable Diffusion far more accessible and customizable than previous TTI models.

This paper provides a brief history on AI image generation, and a basic primer on how latent diffusion models work, followed by a look at some specific projects employing Stable Diffusion to develop a series of images that, in various ways, probe the TTI black boxes, including *Latent Alphamerics*, *Deep Negatives*, and a catalog of the collective visual imaginary of cold war command and control devices, *Destroy the Previous Original*.

Discussions follow of prompt engineering and graphic intervention, as well as perspectives on how one might negotiate artistic intent and system autonomy when working with deep generative systems.

The paper concludes with a brief roundup of Stable Diffusion resources that will hopefully be useful to other artists and designers who wish to engage with these tools.

A Brief History of AI Image Generation

Artificial Intelligence Generated Content (AIGC) can arguably trace its roots back to the 1960s with the development of statistical models that could be employed generatively, such as Hidden Markov Models (for generating sequences of discrete data) and Gaussian Mixture Models (for generating multivariate data). *Practical* generative AI image creation, as presently understood, would not be realized until after deep-learning algorithms were developed in the 2000s, particularly the Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [1].

Arguably the most significant pivot point in terms of contemporary AI image generation [2], was the breakthrough represented by the original GANs by Ian Goodfellow and his team in 2014. This kicked off an alphabet soup of ever-more-sophisticated generative AI image making systems. [3] The "adversarial" part of a GAN refers to a *discriminator* implementation which iteratively evaluates the fitness of the *generator's* output in relation to natural images; the generator attempts to refine a synthetic image to the point that the discriminator classifies it as natural, and both networks become more sophisticated in their respective roles as they receive feedback from each other.

This method produces detailed images with sometimes uncanny verisimilitude, but early models suffered from low image output resolution, a fixed set of predetermined object categories. This method also suffers generally from issues of "mode collapse," in which the model under-represents some objects from images, and "mode-dropping," in which it omits some objects completely. [4]

Just one year later, *DeepDream*, which initially began as an effort by a team including Alexander Mordvintsev at Google to visualize the inner workings of Convolutional Neural Networks (CNNs) [5], unleashed hordes of psychedelic "slugdogs" [6] onto the internet and introduced generative AI images to the general public. In 2016, Leon Gatys et al. separated the "style" from the "content" of an image with Neural Style Transfer (NST); this split led to being able to (crudely) transform any image to resemble the style of famous artists with distinctive styles like Vincent Van Gogh, Henri Matisse, or Katsushika Hokusai.

In 2017, the Artificial Intelligence Creative Adversarial Network (AICAN) platform, housed at Rutgers University, promised an "artificial intelligence artist and a collaborative creative partner" [7], emphasizing the fine arts potential of generative AI.

Open AI, in 2021, introduced *DALL-E*, a versatile text-to-image (TTI) generative AI system that, after a waitlist period, was opened to limited use by the general public. The next year, *DALL-E 2* followed, increasing the resolution of generated images by a factor of four. [8]. *DALL-E 3* is now (in the fall of 2023) available to Open AI's ChatGPT+ subscribers, as well as through Microsoft's Bing image creator. It promises better fidelity between prompt and image and integrates with *ChatGPT* for interactive prompt generation; *DALL-E 3* also includes the ability for artists to opt-out of its training model.

A less visible but still vitally important component of contemporary AI image generation is the Contrastive Language-Image Pre-Training (CLIP) method, developed by OpenAI. [9] This approach "pre-trains" on descriptive text, predicting which caption goes with which image. This pretraining allows for the use of more natural and varied language for image prompting (since it does not depend on pre-defined categories) on the generation side and provides for "few-shot" learning, which is useful when there is limited image caption data on the training side.

Latent Diffusion Models and *Stable Diffusion*

Diffusion models for generative AI image making depart from the adversarial dance of generator and discriminator and rely instead on iterative statistical de-noising operations—essentially cyclically removing visual noise from an existing natural image or (more commonly) a random distribution of noise, until an image that (to some degree!) satisfies the initial guidance emerges. This general model was shown to overcome many of the shortcomings of GANs at the cost of somewhat less efficient image generation. [10]

Stable Diffusion is a specific open-source implementation of a diffusion model for image generation that incorporates CLIP and VAE along with UNET. In a text-to-image operation, the process of transforming a prompt to an image in Stable Diffusion v1.6 goes something like this: First, the user writes a natural language prompt ("a detailed photograph of a cat," for example). This prompt is then broken down into a series of tokens—combining some words and breaking others apart—and those tokens are then encoded into vectors that the UNET neural network can consume. As the prompt is tokenized, a low resolution (64x64 pixel) noise image is created (based on a seed value) and also converted to vectors. The UNET then takes these two encoded vector inputs, noise and prompt, and using a chosen noise "scheduler," begins to iteratively remove noise from the initial low-resolution image in latent space, based on the contents of the particular training set in use.

After the specified number of steps, the denoised image is passed to the VAE which decodes the latent vectors into standard pixels at a higher (in this case 512 x 512 pixel) resolution—and we see the image of a cat.

The training set (called a "checkpoint" in Stable Diffusion terminology) that serves as the repository of encoded and compressed image/description pairs is created in the reverse way: vast numbers of captioned images, usually scraped from the web, are injected with noise, and then immediately de-noised. The original ("natural") image is then statistically compared to the processed image to determine how closely the two match. The model typically uses this process on billions of text image pairs to train "weights" which are later used by the UNET neural network in the image generation process.

The Library of Babel

When it was announced that the Library contained all books, the first reaction was unbounded joy. All men felt themselves the possessors of an intact and secret treasure.
There was no personal problem, no world problem, whose eloquent solution did not exist—somewhere in some hexagon.

–Jorge Luis Borges, *The Library of Babel* [11]

Much like in the early days of sound synthesis, when it was believed that any natural sound could be produced [12] synthetically, the advent of AI image generation has engendered a similar

idea that any image, in any style, might be pulled from latent space—that it might be a visual Library of Babel. And even though Borges' library is unimaginably large [12], Stable Diffusion's latent space is even larger.

Are *all possible images* available in Stable Diffusion's latent space? No. It seems that Stable Diffusion's VAE has a decidedly finite capacity to decode images, but that number of possible images is still shockingly vast: likely something like 2^{524288} images [13] (this represents a number so large that it would require 61 pages in this book just to type out all the digits in integer form). It is no wonder that people imagine that any possible image might be waiting to be discovered/created there.

Coupled with the idea that nearly image is possible to create, is the fear that specific images (artist's own works, for instance) might be perfectly re-created. Diffusion models generally do not replicate specific images in their training data. If a checkpoint is trained well (not "overfitted" for instance), and reasonably large, it is unlikely that a diffusion model will ever precisely reproduce a specific image that was included in the training data, even if the precise wording of the caption in the image-caption pair is used as the only prompt. However, Somepalli, et. al [14] found that, in specific cases, Stable Diffusion would recreate, pixel-for-pixel, portions of images in the training data within generated images, so there is a legitimate concern—even more so with Low Rank Adaptations (LORAs) which are deliberately trained on a specific set of visuals.

The Question of Autonomy

Why this mountain?
Why this sky?
This long road, this empty room?

—Laurie Anderson

I first experienced producing generative AI images in 2019 with Joel Simon's GAN breeder [15]. This browser-based system used BigGAN on the back-end (trained on 128 x 128 pixel images) to synthesize new images, not by writing a text prompt, but by adding "genes" (object categories) and using a slider to weight each gene's contribution to the final image + a "chaos" modifier. Genes included categories as diverse and infelicitous as "mousetrap," "landscape," "cassette," "underwater creatures," "bikini," "patterns," and "vase." The breeding aspect included an ability to combine a pair of generated images with adjustments to how similar or different the offspring images would be to the parent images.

Even though it was possible to add only one gene at maximum weight to generate an image that had at least a passing resemblance to a member of the chosen category (setting the king crab slider to 1.0 would always produce something with legs and spikes), it was far more interesting to add a number of genes with a variety of weights and just see what emerged.

Rather than tending to create an image of each item (as contemporary TTI systems generally do, GANBreeder would instead generally create a single object or environment that somehow expressed aspects of the selected

categories, but not necessarily in a legible way. This led to images of ambiguous and dreamlike objects and scenes that had a mood, but seldom bore a strong resemblance to anything in reality.

As AI models and training sets have improved, the ability to generate an abstract image that conveys a mood without being littered with literal representations from the prompt has become increasingly difficult.

Prompt engineering, LORAs, Textual Inversions, and specialized training sets tend to reinforce the drive to erase the marks of AI image generation and render the process invisible.

In "Is Writing Prompts Really Making Art?" McCormack, et al, [16] seem to variously argue that AI image generators are both too autonomous and not autonomous enough. This ambivalence is understandable (if not well-explained in the article) since autonomy in these systems is largely a function of perspective and specific use. The question of autonomy is, nonetheless, an important one when considering if (and how) to engage with these systems in an art practice.

Tighter and more extensive prompt engineering tends to provide more control over the generated image, ceding less autonomy to the system at the generation phase, even if the process is still somewhat fraught by the ambiguities of language and the quirks of the training data, (which are largely unknowable). On the other side, huge image generation batches with varying parameters (random or incremented) may be launched so that the system has relatively more autonomy at generation, but an artist will have a greatly expanded pool to curate from and iteratively enhance.

In addition, with current systems, extra-textual graphic interventions such as *inpainting* (visually indicating areas of a generated image to recalculate by "painting" with the mouse—often used to repair faces and hands in images of people), *outpainting* (asking the system to fill in around a desirable part of an image, or by using an existing image (natural or generated) as a nucleus for the system to, to varying degrees, augment with prompt-generated imagery.

Despite McCormack, et al's generally hostile (and, in my estimation, poorly argued) take on AI image generation, an interesting concession is made that these tools will likely be used in unexpected ways by artists to create new work—particularly when employed in a "meta" that embraces the flaws of the method and interrogates the process. With this perspective in mind, I would like to briefly present three projects.

Praxis

In 2021, at this conference [17], I presented Automatic Cities, a spatial soundwalk through the ancient city of Sardinia. The mobile phone interface to the piece included AI-generated illustrations for each of the imaginary places along the walk. As an indicator of how rapidly this technology is progressing, I offer you an image I showed as an example then generated with VQCLIPGAN from the prompt "Life

is the parable of fading shadows:"

Figure 1 "Life is the parable of fading shadows | Chromolithography," 2021, VQCLIPGAN

The same prompt, executed in fall of 2023 using Stable Diffusion XL:

Figure 2 "Life is the parable of fading shadows | chromolithography," 2023, Stable DiffusionXL

The contrast in these two images, using an identical prompt, is visually striking (and emblematic of the increasing concreteness of these systems).

The first current work-in-progress project I'd like to speak about is called *Latent Alphamerics*. This is a series of large-format digital prints that present a somewhat critical response to generative AI image creation. The images in this series are the result of prompting a generative AI text-to-image diffusion system with deeply ambiguous texts—in this case, single letters of the alphabet and single-digit numbers.

Each image in the series is named for its numeric seed, which represents the latent space starting conditions. The lack of context and meaning is no impediment to the system's image generation ability, but the images that surface are like core samples from the trained model's latent space; multiversal fragments of potential people, potential places, and potential events.

It is somewhat unclear (at least to me) if single characters without surrounding context are properly tokenized in Stable Diffusion, but even if they are, practice shows that not much semantic information adheres to them: images generated by these prompts owe far more of their structure to the generative seed used than to the prompt itself (in a reversal of usual situation where seed variations are far less significant than substantive changes to the text prompts).

Here are three example image matrices generated with Automatic1111's X-Y grid feature with a constant seed value and varying single character prompt (A–Z and 0–9) from left to right:

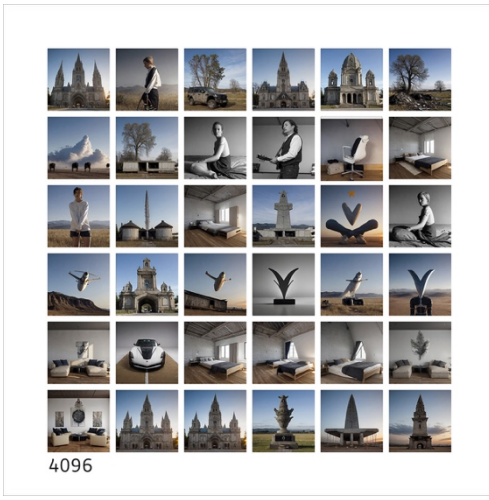

Figure 3, Seed = 4096

Note the persistent color palette and overall compositional similarities, but semantically drifting imagery.

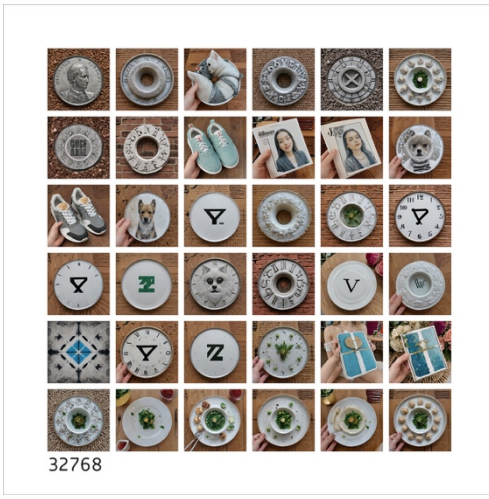

Figure 4, Seed = 32768

Here, the seed strongly favors a centered circular composition, though images periodically deviate toward a still central but slightly rotated rectilinear element.

Figure 5, Seed = 1066

Notice too the emergence of human figures in each of the three series, but in locations that don't correspond to the same characters.

These matrices of usually banal images offer glimpses into the otherwise overdetermined secret gardens of latent spaces.

Next is *Deep Negative (the Monstrous Feminine)* which offers a different kind of probe into a training model's latent space, as well as some insight into preferences and biases of an anticipated typical user. This series came about due to experiments with placing a general-purpose negative prompt (DeepNegative 1.6) as a "textual inversion" (TI) into the *positive* prompt area of the Automatic1111 interface.

A TI is another black box (or at least a very dark grey one) that embeds a new "vocabulary word" for Stable Diffusion prompts derived from training on a small set of images.

If there is a particular coffee mug, for example, that you would like to see represented in AI generated images, but that was not included in the training data for the checkpoint in use, you could train on a handful of images of that mug and give it a designation “M*.” Then, in a prompt, you could ask for “a photo of M*” or even a “photo of a broken M* full of red roses.”

Besides using the TIs to add specific concepts to a generated image, they may also be used in the negative prompt to suppress specific concepts. In Automatic 1111, I was curious what would happen if I populated the *positive* prompt box with the DeepNegative TI instead of the negative one as expected.

Figure 6, One of the few images from Deep Negatives *I felt comfortable including in this paper; the vast majority of these images feature full frontal female nudity*

Clicking on the render button immediately produced a wild-haired, green-skinned scaly young woman—completely nude—gleefully showing me her heavily deformed hands. Every image generated, seed after seed and batch after batch was similar: a never-ending parade of corpse-tinged women, each unique, but all united in nakedness, grotesquery, and most exhibiting a kind of feral joy. A sisterhood of monsters. It is telling that the negative prompt *always* produces unambiguously female figures with virtually any Stable Diffusion 1.6 checkpoint—only the art style and color palette seem to vary.

I am currently exploring making work with and around these images with my colleague (and talented performance artist!) at the University of Kentucky, Professor Rae Goodwin.

Destroy the Previous Original is the final piece I would like to present. This is also a work in progress; one that I am deeply ambivalent about. The initial concept was to use Stable Diffusion to explore the collective imaginary of cold war Command and Control systems. In working through the piece, the designer in me was frustrated by the lack of precision the images exhibited in the form of distorted perspective, warped geometries, and indistinct and melted looking knobs, dials and sliders. Specific training data for mechanical controls seemed to be seemed to be even less represented than well-formed hands. As the prompt text ballooned to over 85 tokens, in an attempt to reform these images towards more aesthetic directions, I steadily undercut the original intent of the project.

At the moment, I am back to favoring sparser prompts such as simply “Cold War weapons control panel.”

Figure 7, An image from Destroy the Previous Original *series; see the artworks section in this volume for more examples*

Conlusion

While I can't agree that AI TTI is "parasitic" and skill-less or somehow uniquely misrepresented by its promoters as an art-making tool (ignoring the long history of "paint-by-number" kits and "draw Tippy the Turtle") or even that TTI is art "forgery" or fast food art, (celebrating the marks of a tool or obscuring them are both valid contemporary artistic strategies), I do believe that it is something of a dancing bear when "used as directed" and, presently, may best be engaged with as a sort of weird mirror held up to the specific collective imagery of the internet and the assumed intent of its users.

Resources

Figure 8
Automatic1111, a web-based GUI for local execution of Stable Diffusion:
https://github.com/AUTOMATIC1111/stable-diffusion-webui
An excellent first tool for a locally-hosted installation of Stable Diffusion.

Figure 9
ComfyUI, a node-type web-based GUI for local execution of Stable Diffusion:
https://github.com/comfyanonymous/ComfyUI
ComfyUI is neither comfortable nor exactly a UI! It is a modular "boxes and wires" type tool, and it is more complex to maintain and use than Automatic 1111, but it lends itself to customized workflows and experimentation.

LAION crawl is an aesthetic subset image database browser for the Large-scale Artificial Intelligence Open Network (LAION) image/text pair training set:
https://laion-aesthetic.datasette.io/laion-aesthetic-6pls/images
Here you may search or browse a subset of the specific images and associated text data used in training Stable Diffusion. Entries may be sorted by different criteria including dimensions, caption (alt tag), URL, and aesthetic score.

Magic Prompt is a web-based tool to help build better TTI prompts, in this case for Stable Diffusion:
https://huggingface.co/spaces/Gustavosta/MagicPrompt-Stable-Diffusion
This tool may provide insights into what goes into a particular model as well as how it is expected to be addressed.

References

[1] Cao, Yihan, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt." *arXiv preprint arXiv:2303.04226* (2023). 111:4

[2] Cetinic, Eva, and James She. "Understanding and creating art with AI: Review and outlook." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, no. 2 (2022): 1-22. 66:9.

[3] Durgadevi, M. "Generative adversarial network (gan): a general review on different variants of gan and applications." In *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1-8. IEEE, 2021.

[4] Bau, David, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. "Seeing what a gan cannot generate." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4502-4511. 2019.

[5] "Inceptionism: Going Deeper into Neural Networks." 2015. Blog.research.google. June 17, 2015. https://blog.research.google/2015/06/inceptionism-going-deeper-into-neural.html.

[6] Avery, Grace. 2016. "Deepdreaming without the Slugdogs | Graceavery." July 1, 2016. https://graceavery.com/deepdreaming-without-the-slugdogs/.

[7] "AICAN." n.d. AICAN. https://www.aican.io/.

[8] OpenAI. 2023. "DALL·E 2." OpenAI. 2023. https://openai.com/dall-e-2.

[9] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *International conference on machine learning*, pp. 8748-8763. PMLR, 2021.

[10] Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." *Advances in neural information processing systems* 34 (2021): 8780-8794.

[11] Borges, Jorge Luis. "The library of Babel." *Collected fictions* (1998).

[12] Théberge, Paul. *Any sound you can imagine: Making music/consuming technology*. Wesleyan University Press, 1997. p. 76.

[13] Bloch, William Goldbloom. *The unimaginable mathematics of Borges' Library of Babel*. Oxford University Press, 2008.

[14] Wiskkey. 2022. "I'll Restrict My Answer…." October 23, 2022. https://www.reddit.com/r/StableDiffusion/comments/ya4te3/comment/iths4d0/.

[15] Somepalli, Gowthami, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. "Diffusion art or digital forgery? investigating data replication in diffusion models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048-6058. 2023.

[16] McCormack, Jon, Camilo Cruz Gambardella, Nina Rajcic, Stephen James Krol, Maria Teresa Llano, and Meng Yang. "Is Writing Prompts Really Making Art?." In International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar),

pp. 196-211. Cham: Springer Nature Switzerland, 2023.

[17] Simon, Joel. 2022. “Ganbreeder.” GitHub. September 28, 2022. https://github.com/joel-simon/ganbreeder.

[18] Eby, Chad Michael. “Among Black Boxes and Maze-building Rats: Reflections on Art Making with Autonomous Rules Systems” Generative Art 2021 Proceedings of XXIV GA Conference (November 9, 2021).