Self-Cannibalizing AI

Artistic Research on Recursive Dynamics in Generative Image Models

Ting-Chun Liu, Leon-Etienne Kühr
Academy of Media Arts Cologne, Offenbach University of Art and Design
t.liu@khm.de, kuehr@hfg-offenbach.de

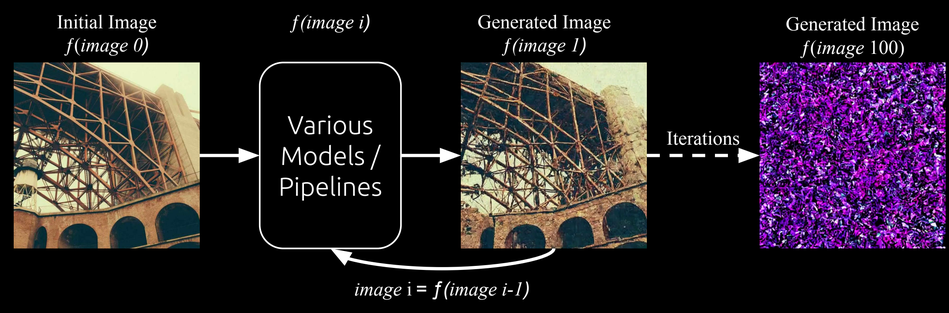
ABSTRACT

In the rapidly evolving landscape of diffusion-based artificial intelligence (AI), feedback processes within generative models like Stable Diffusion (SD) [1] are becoming increasingly significant. This research examines the phenomenon of "self-cannibalizing"-feedback loops, where the recursive use of model outputs as subsequent inputs for the model leads to unique emergent behaviors, biases and hints at a potential model collapse [2].

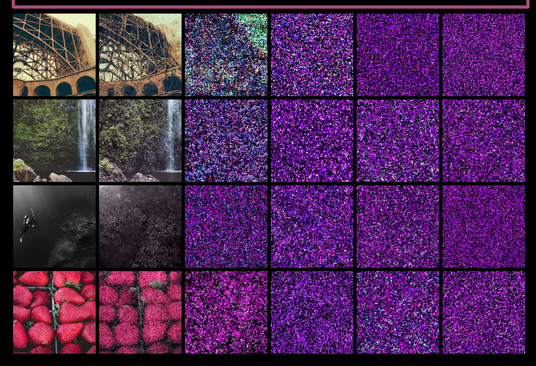
The focus is on the Stable Diffusion pipelines, examining how feedback causes images to transform into limited patterns and concepts after only a few iterations. Instead of focusing on the impact of training data, we explore the algorithmic contributions of various components within the diffusion pipeline. The inner workings of diffusion models embody cybernetic principles, with feedback loops and self-regulating mechanisms shaping generative outcomes. This recursive process reveals the model's tendencies and uncovers potential structures. We investigate the aesthetic and medium specificity within diffusion model-generated imagery, offering insights into its unique visual language and underlying computational frameworks.

METHOD

We apply recursive methods to various models and components of the latent diffusion pipeline, using a generated image as the input for subsequent generation. Repeatedly applying a model accentuates patterns introduced by the process that are nearly invisible when the model is used only once.



Since various input images and seeds result in similar patterns, this experiment focuses on using the same initial image.



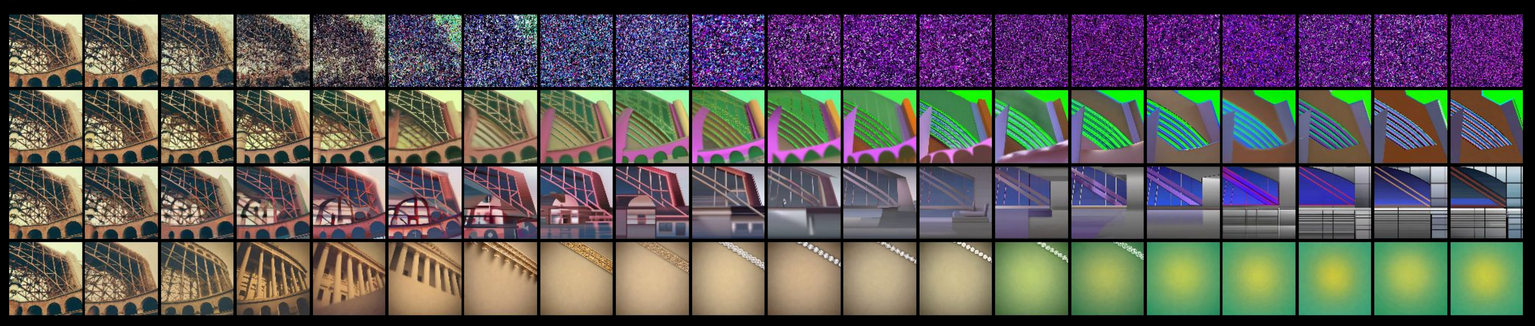
The initial images were sourced from Lorem Picsum [3]

EXPERIMENTS

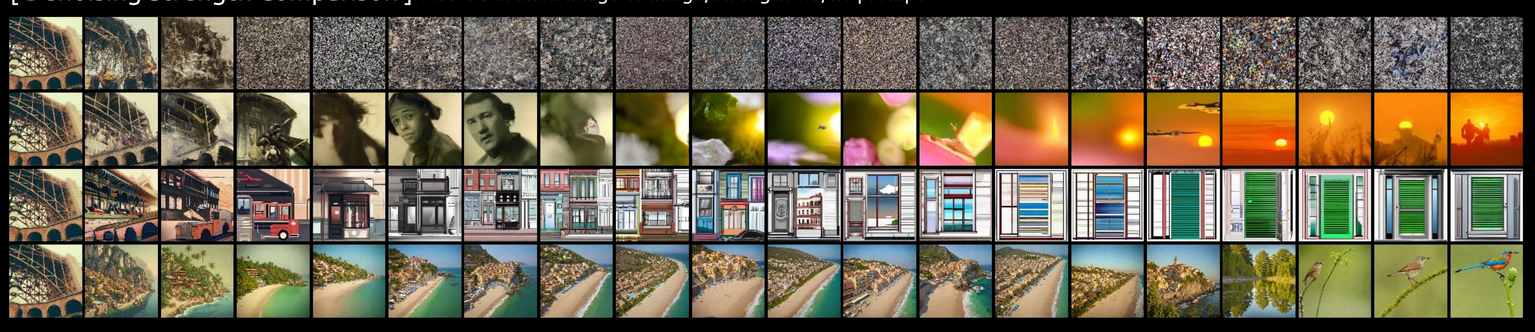
A closer inspection of latent diffusion models uncovers a complex pipeline of trained algorithms. Since the patterns is exclusively observed through recursive application, this prompted an investigation into the mechanisms driving these phenomena.

By dissecting the pipelines and applying consistent feedback loops, we isolate the components responsible for the observed effects.

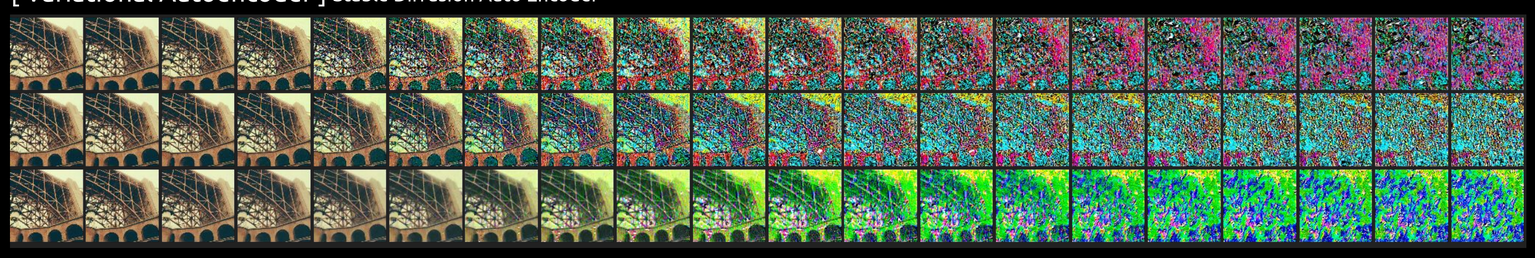
[Baseline] Stable Diffusion Image-to-image, strength 0.5, no prompt



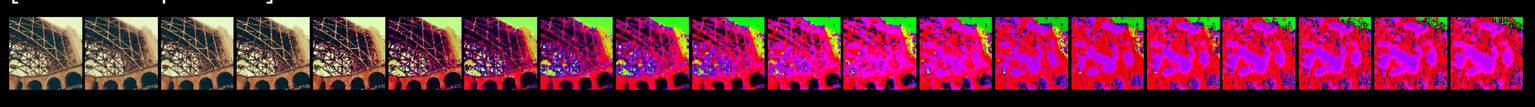
[Denoising Strength Comparison] Stable Diffusion Image-to-image, strength 0.8, no prompt



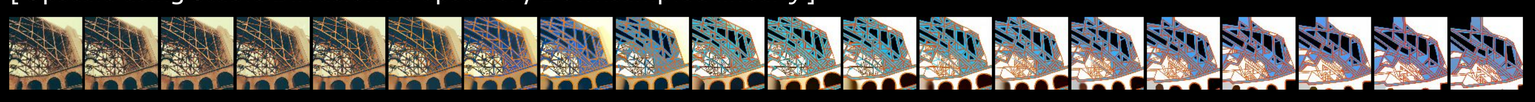
[Variational Autoencoder] Stable Diffusion Auto Encoder



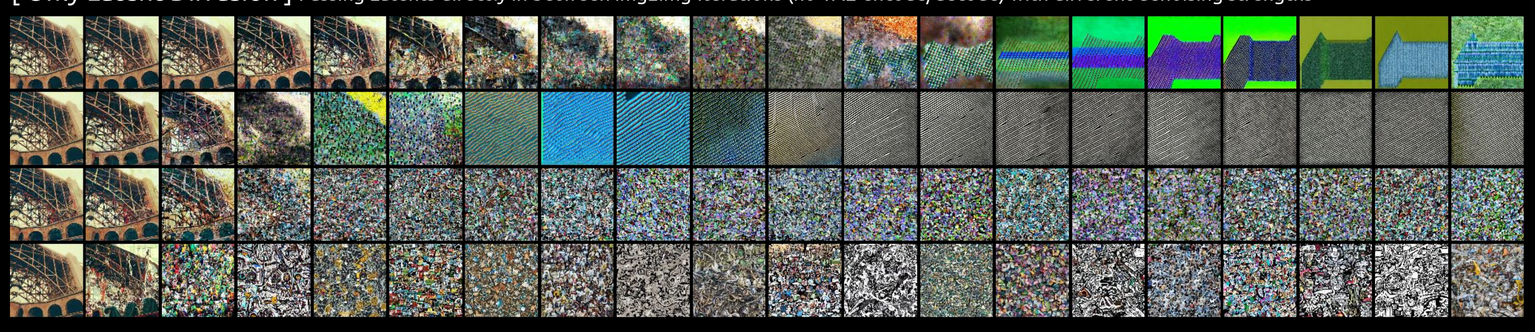
[Latent Compression]



[Upscale using Stable Diffusion x4 upscaler / Downsampled linearly]



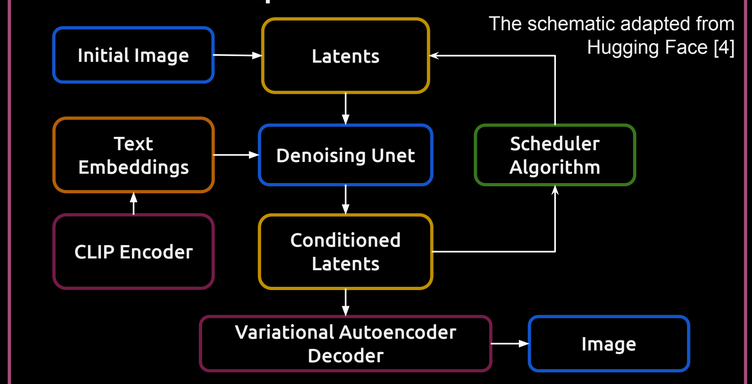
[Only Latent Diffusion] Passing Latents directly in between img2img-iterations (no VAE-encode/decode) with different denoising strengths



* The progression of images shown follows a power function rather than linear steps, starting from the original image on the left.

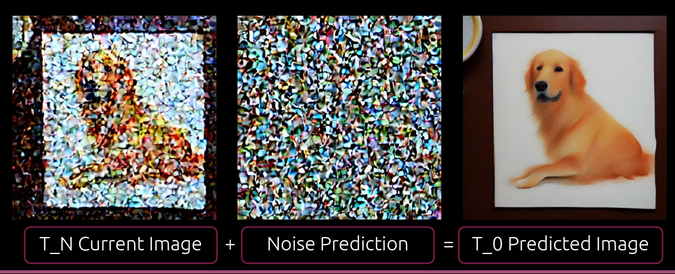
PIPELINES

Stable Diffusion Pipeline



Diffusion Model

A diffusion model gradually adds noise to an image and then learns to reverse this process step-by-step, removing the noise to generate a clear image.



Denoising Strength Comparison

Strength controls the amount of noise added to the input image before denoising, with high strength leading to cyclic, model-specific concepts and low strength converging to distinct patterns for each model.

Variational Autoencoder (VAE)

A VAE encodes images into a lower-dimensional latent space for model processing and then decodes them back into images. Repeating this encoding and decoding process highlights color shifts learned through training and reveals patterns indicative of the underlying convolutional architecture.

Latent Compression

Compressed latent representation of images using the diffusion process to progressively generate details, with patchy patterns indicating high-frequency detail loss.

Upscale / Downscale

Latent Diffusion Models (LDMs) refine low-dimensional image representations, gradually transforming a low-resolution image into a high-resolution version. Repeated use shows an examination of the close neighborhoods through low-frequency patterns.

Only Latent Diffusion (VAE skipped)

Directly passing latents to and from the denoising U-Net bypasses the VAE's influence, causing patterns to converge into pure noise, repeating motifs, or perfectly repeating patterns, depending on the noise strength.

CONCLUSION

Our experimentations discovered distinct differences in color and pattern distribution when specific algorithms and components are subjected to recursive processes. These variations reveal the accumulation of algorithmic biases and the models' tendencies to gravitate toward particular visual concepts, repetitive motifs, and colors, offering a glimpse into the models' working mechanisms and underlying structures. Our findings suggest that artificially introducing randomness might prevent significant degradation, which subsequently causes the patterns to occur. This raises concerns about future models deteriorating over time, as even a single use could introduce subtle degradation that compounds with further iterations. The currently unseen effects of latent diffusion models may, over time, contribute to a future model collapse.

This investigation initiates a philosophical and aesthetic discourse on mapping out the possibility space of diffusion models. Through our experiments, we've identified the components driving the observed effect. This research documents our artistic exploration of diffusion models' methodologies on a smaller scale. More broadly, it contributes to the discourse on AI agency by examining model collapse and anticipating scenarios where generative artificial intelligence iterates on its results.

REFERENCES

1. Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. "**The Curse of Recursion: Training on Generated Data Makes Models Forget.**" arXiv:2305.17493 (2024).
2. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "**High-Resolution Image Synthesis with Latent Diffusion Models.**" arXiv:2112.10752 (2022).
3. Lorem Picsum. "**Lorem Picsum: Placeholder Images for Testing.**" Available at: https://picsum.photos
4. Hugging Face, "Stable Diffusion," Hugging Face Blog, accessed December 24, 2023, https://huggingface.co/blog/stable_diffusion.

ACKNOWLEDGEMENT

We sincerely thank everyone who supported this research, particularly our mentor at the Academy of Media Arts Cologne, Prof. Dr. Georg Trogemann, and Christian Heck. We also acknowledge the broader research community for their influential contributions to generative artificial intelligence, cybernetics and aesthetics.

Disclaimer: No AI was harmed during the making of this poster and these experiments would not have been possible without the use of ChatGPT's code.