MAT 255 - Final Project

Investigation of CLIP and other image-to-text AI

The goal of this project was to learn something about the semantic space of generative AI. In this course we have mostly engaged with AI image generation through text prompts. This project looks at the reverse process: how an AI turns an image into text. In order to keep with the nature of this course, images were then created from the generated text.

CLIP interrogator

- [website]
- "CLIP Interrogator is a tool that uses the CLIP (Contrastive Language–Image Pre-training) model to analyze images and generate descriptive text or tags, effectively bridging the gap between visual content and language by interpreting the contents of images through natural language descriptions."
- [link to paper]
- Relavant quote:
    - "Although the technical specifications of CLIP Interrogator have not been published as a paper and there is no literature that can be referred to, it may be presumed from the code[6] and its operation that it first generates a base caption using BLIP and then selects and adds phrases that match the target image from a predefined set of phrases called Flavors. CLIP image/text encoders [13] are used to measure the degree of matching between a target image and the phrases in Flavors. Flavors contains approximately 100,000 words and phrases, including those referring to objects and entities (e.g., motorcycle, building, young woman), image styles (e.g., photo-realistic), and artist names (e.g., greg rutkowski). We used the code released by the developer (clip interrogator.ipynb, version 2.2)."
- Overview of CLIP interrogator process

1) Base Caption Generation:

Use the BLIP model to create an initial caption for the image. This gives a general description of what's in the image.

2) Enhancement with "Flavors":

Adds specific phrases, known as "Flavors," to the base caption. These phrases cover various categories like objects, styles, and artist names.

3) Matching with CLIP:

Uses the CLIP model to match the image with the most fitting phrases from the "Flavors". This ensures the final text is more detailed and closely aligned with the image's content.

- Details for each model

1) BLIP Model:

BLIP (Bootstrapped Language Image Pretraining) focuses on generating a basic, initial caption for an image.

It's designed to provide a general understanding of what the image depicts, creating a simple and straightforward description. This serves as the foundation for further analysis.

2) CLIP Model:

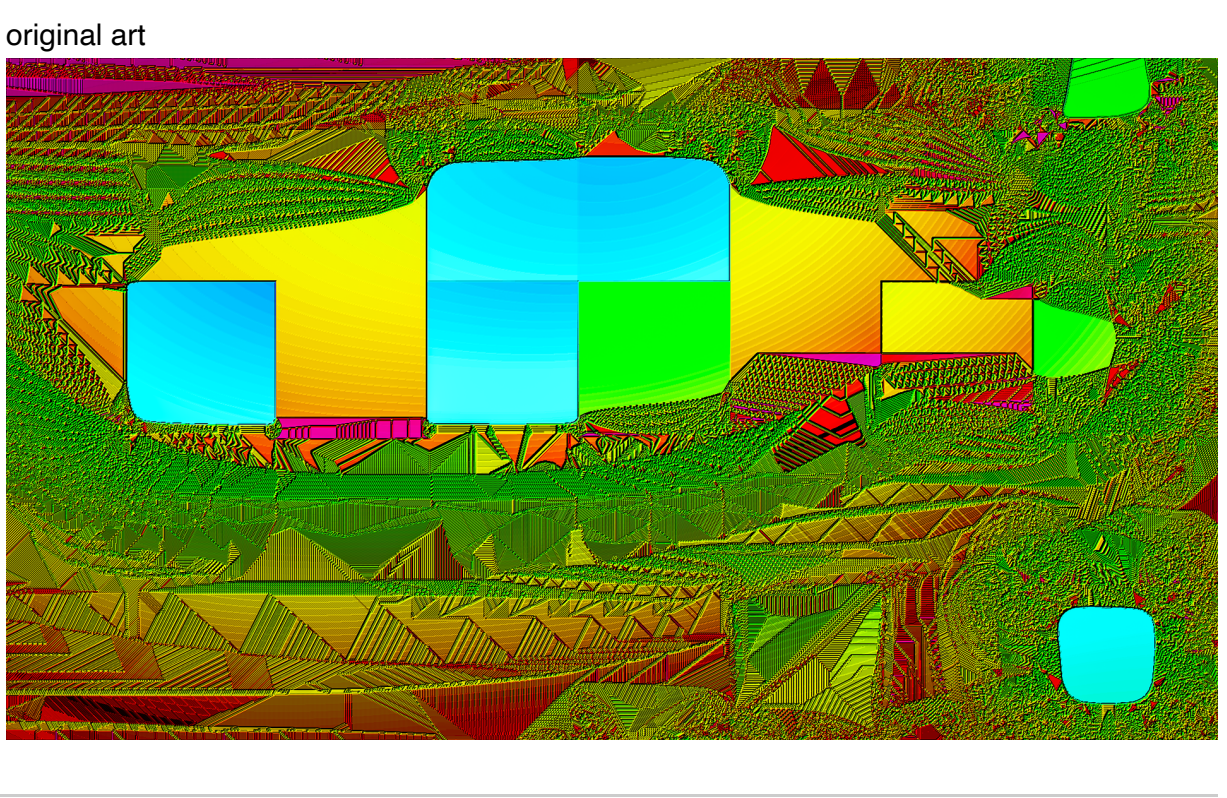
CLIP (Contrastive Language–Image Pre-training) takes the basic description from BLIP and enhances it. It compares the image with a variety of predefined phrases to add more details to the description.

This process ensures that the final text is much more detailed and closely aligned with the specific content and context of the image.

- Open AI CLIP info
    - Learns visual concepts from natural language supervision
    - [open AI website about CLIP]
    - "A critical insight was to leverage natural language as a flexible prediction space to enable generalization and transfer."
    - Source of supervision is text paired with images from the internet
    - "This data is used to create the following proxy training task for CLIP: given an image, predict which out of a set of 32,768 randomly sampled text snippets, was actually paired with it in our dataset.
    - "In order to solve this task, our intuition is that CLIP models will need to learn to recognize a wide variety of visual concepts in images and associate them with their names. As a result, CLIP models can then be applied to nearly arbitrary visual classification tasks."
    - "While CLIP usually performs well on recognizing common objects, it struggles on more abstract or systematic tasks such as counting the number of objects in an image and on more complex tasks such as predicting how close the nearest car is in a photo. On these two datasets, zero-shot CLIP is only slightly better than random guessing. Zero-shot CLIP also struggles compared to task specific models on very fine-grained classification, such as telling the difference between car models, variants of aircraft, or flower species."
    - "CLIP also still has poor generalization to images not covered in its pre-training dataset. For instance, although CLIP learns a capable OCR system, when evaluated on handwritten digits from the MNIST dataset, zero-shot CLIP only achieves 88% accuracy, well below the 99.75% of humans on the dataset. Finally, we've observed that CLIP's zero-shot classifiers can be sensitive to wording or phrasing and sometimes require trial and error "prompt engineering" to perform well."

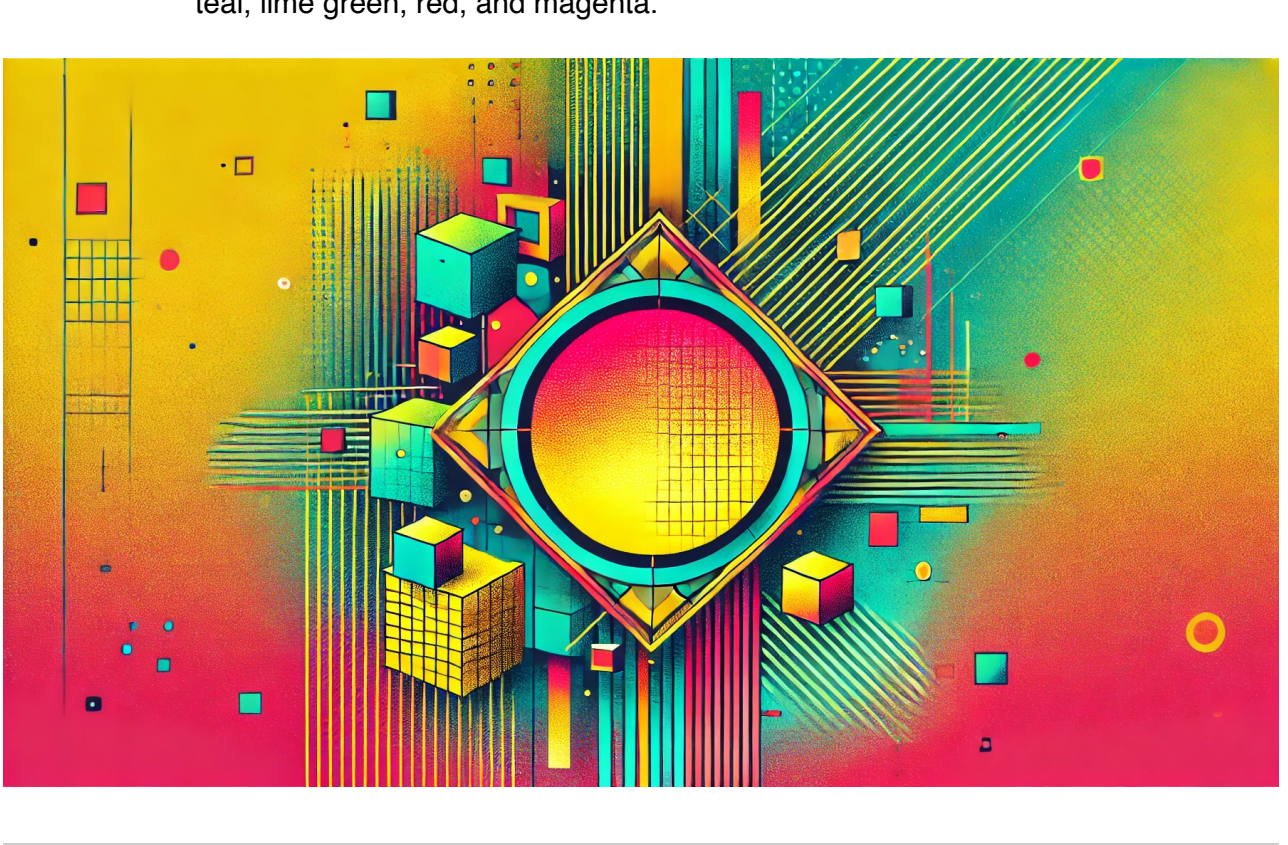
Results

- original art

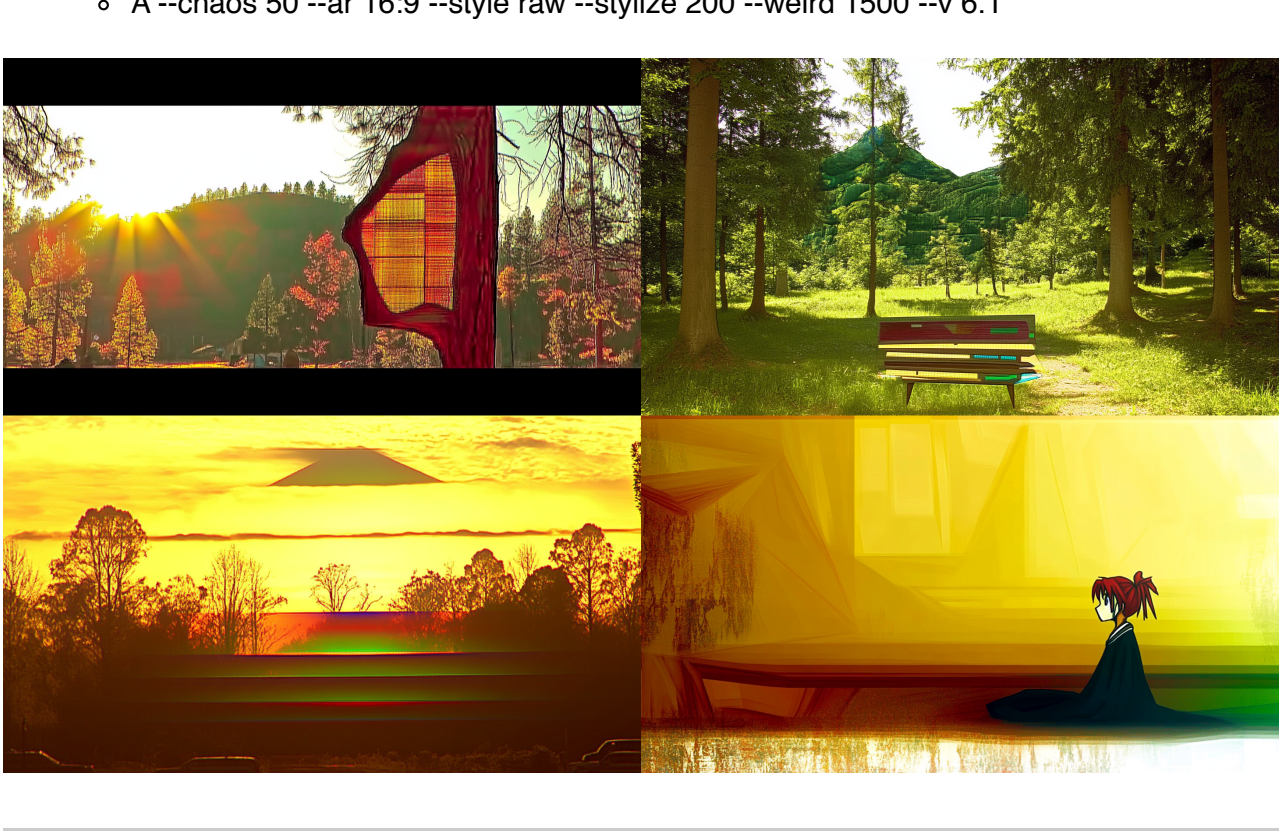


- genrerated CLIP: a digital painting of a colorful abstract background with a large square in the center of the image and a smaller square in the middle of the image, fractals, a raytraced image, objective abstraction, Benoit B. Mandelbrot Steps: 40, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 585850058, Size: 1280x720, Model hash: 31e35c80fc, Model: sdxlbase_1.0, Version: v1.6.0 -

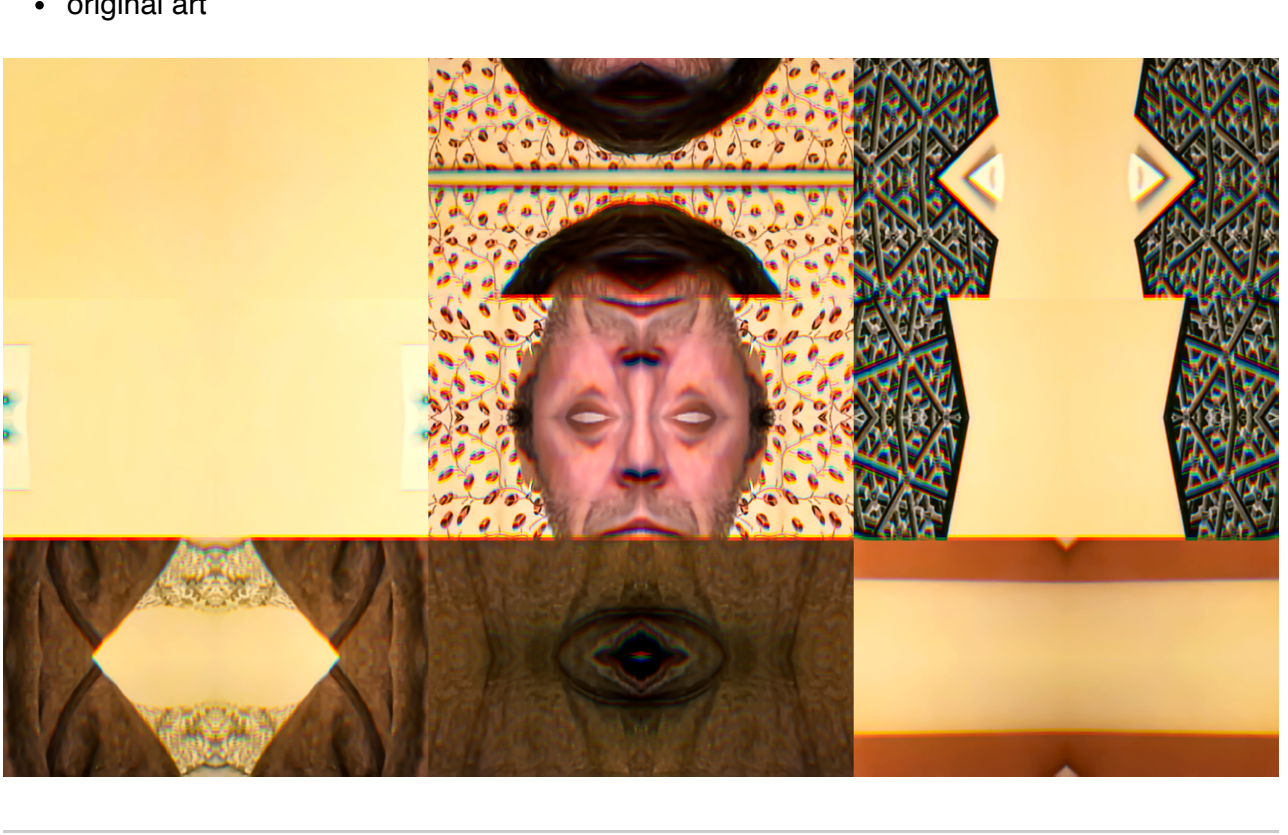
- from DALLE (chatGPT): - original: "An abstract digital artwork featuring a central large yellow gradient shape flanked by teal and lime-green squares, surrounded by intricate geometric patterns and tessellations. The background is a vibrant mix of red, magenta, and green with dense, textured linework. The style is futuristic, with sharp triangular patterns, flowing gradients, and a vibrant color palette of yellow, teal, lime green, red, and magenta. The composition has a kaleidoscopic and energetic feel, evoking a sense of digital complexity."
    - this failed
        - simplified: "An abstract digital artwork featuring a central yellow gradient shape surrounded by teal and lime-green squares. The background is vibrant with a mix of red and magenta, featuring textured linework and sharp geometric patterns. The composition is energetic and futuristic, using a limited color palette of yellow, teal, lime green, red, and magenta."



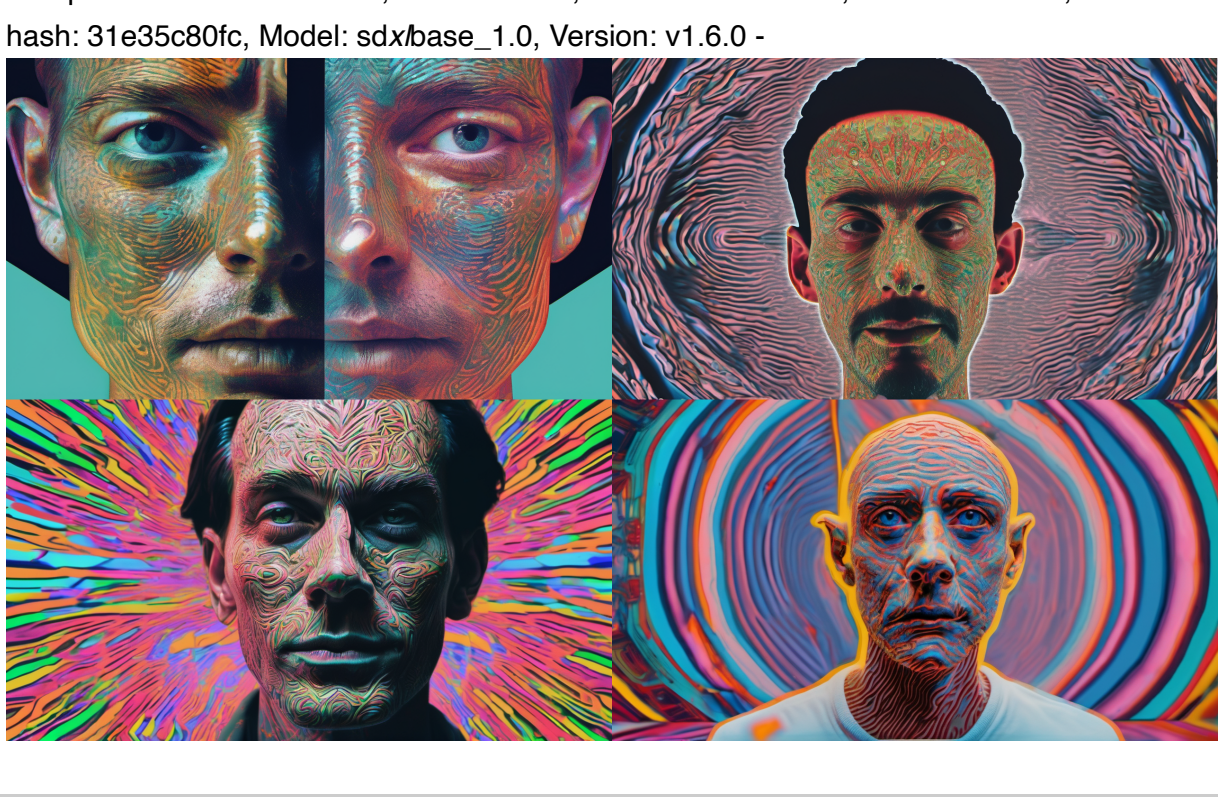
- Midjourney
    - midjourney doesn't have a image-text option so I uploaded the image and just used the letter A as the prompt. From the results it appears to do some subject analysis.
    - A --chaos 50 --ar 16:9 --style raw --stylize 200 --weird 1500 --v 6.1



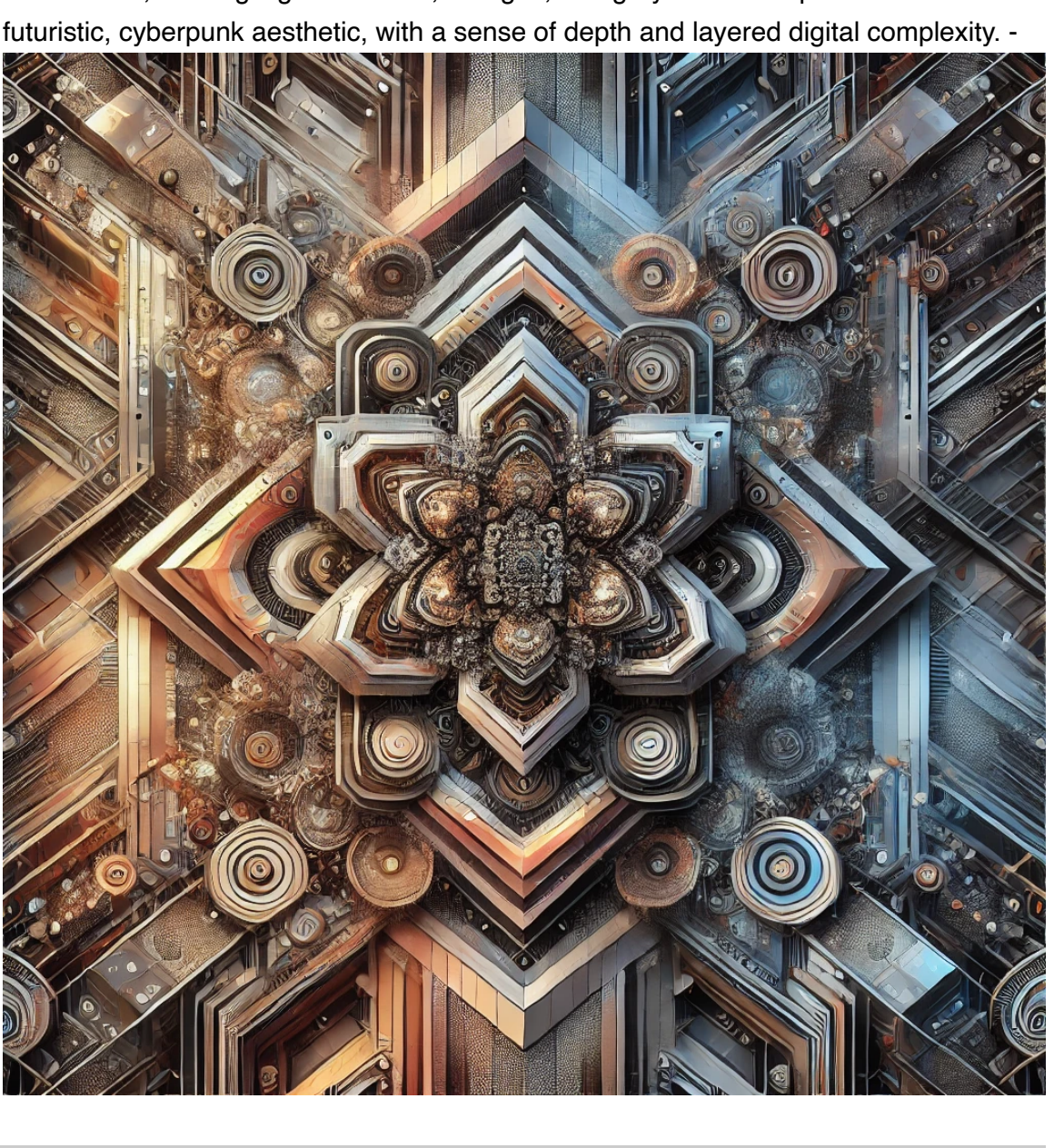
- original art



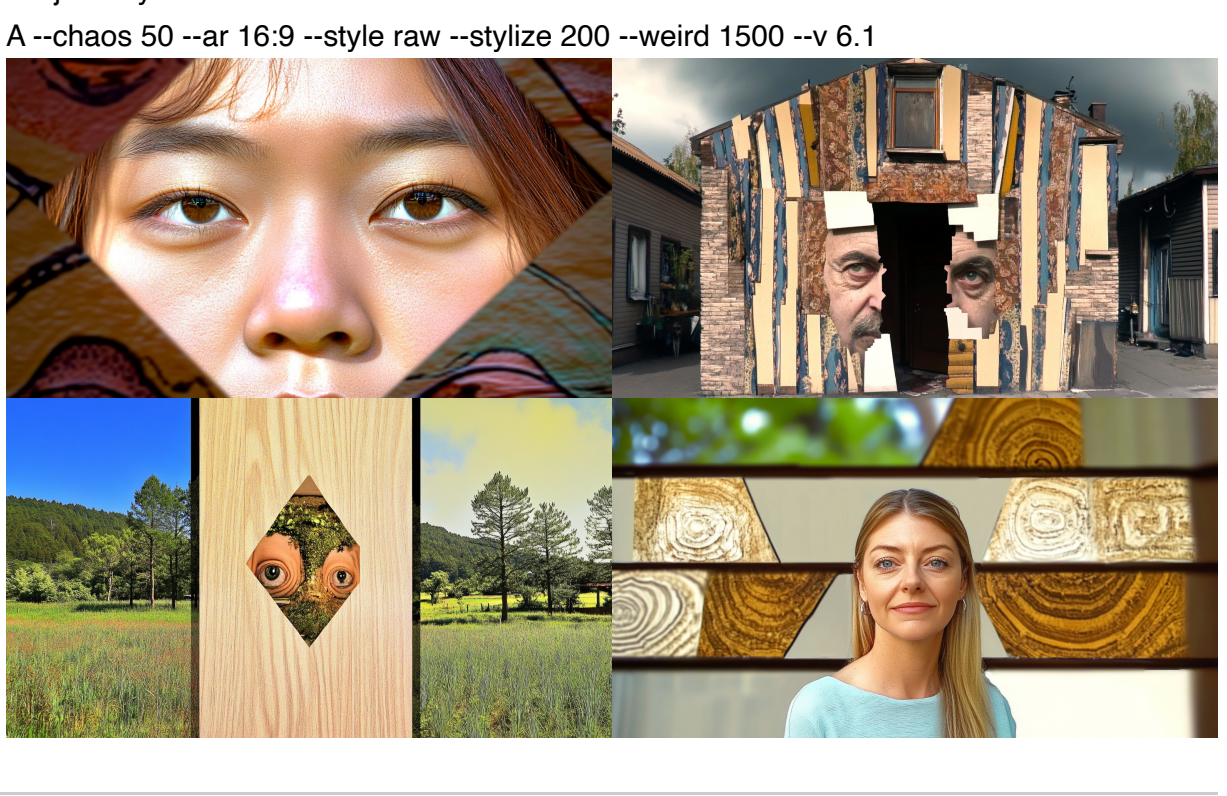
- generated CLIP: a man with a strange face and a strange background with a distorted image of a man's face, dmt, a character portrait, video art, Bedwyr Williams Steps: 40, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 2657358959, Size: 1280x720, Model hash: 31e35c80fc, Model: sdxlbase_1.0, Version: v1.6.0 -



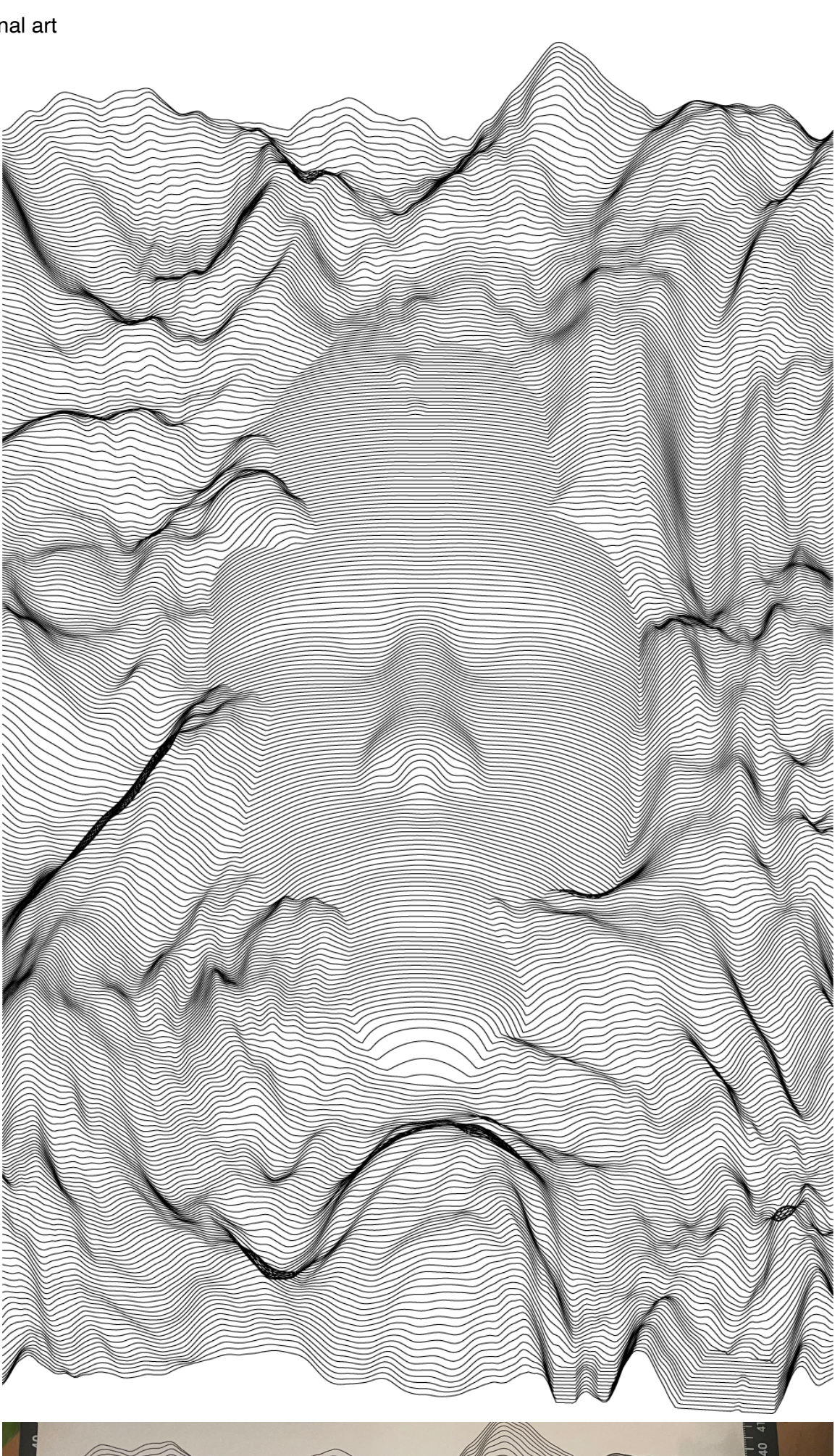
- DALLE
    - An abstract digital artwork with intricate geometric patterns and rich textures. The design features overlapping shapes with sharp edges and a blend of metallic and gradient tones. The background is filled with fractal-like details and repeating patterns, creating a complex and hypnotic visual field. The palette includes a mix of cool and warm tones, with highlights of blues, oranges, and grays. The composition has a futuristic, cyberpunk aesthetic, with a sense of depth and layered digital complexity. -

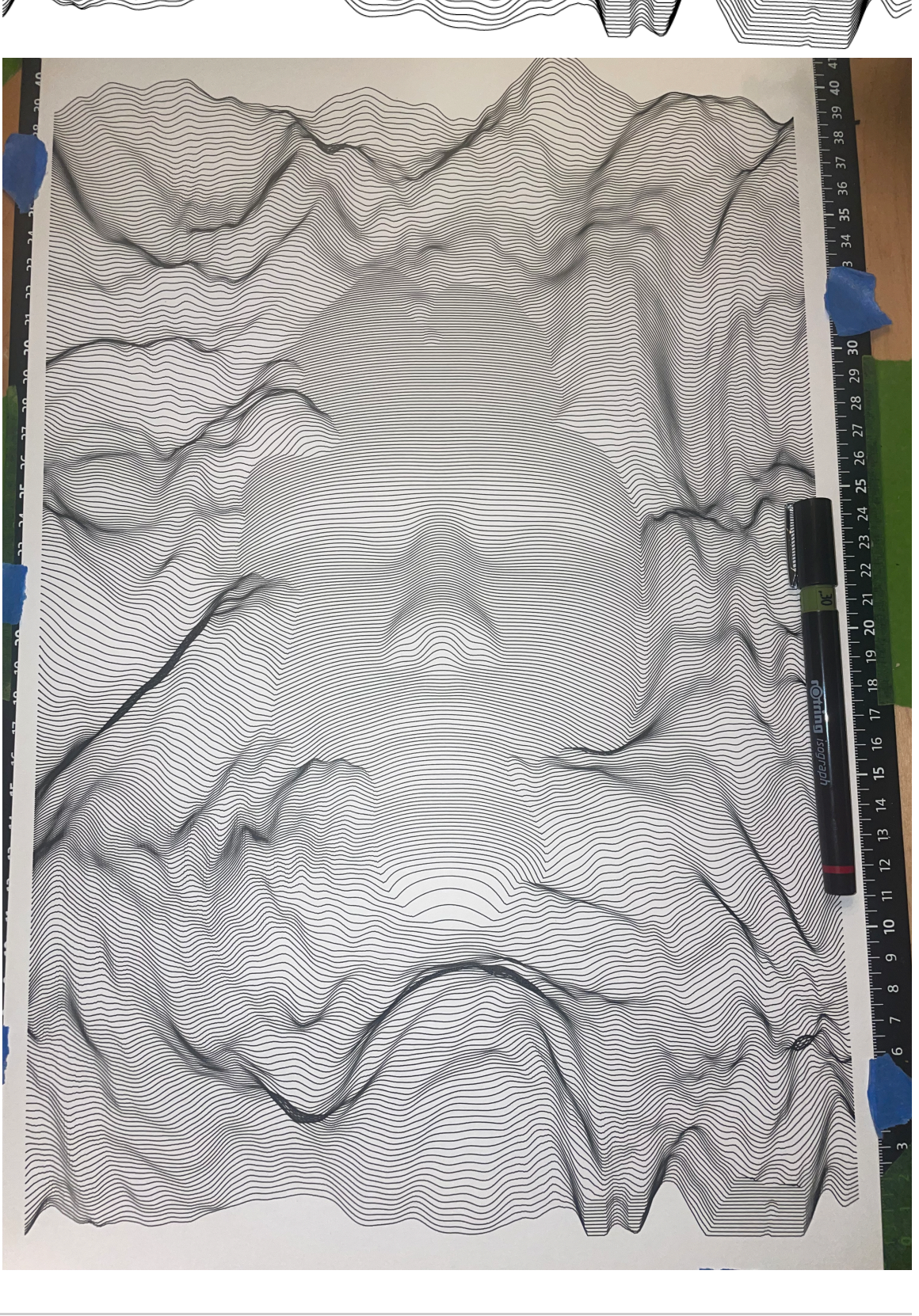


- Midjourney
- A --chaos 50 --ar 16:9 --style raw --stylize 200 --weird 1500 --v 6.1



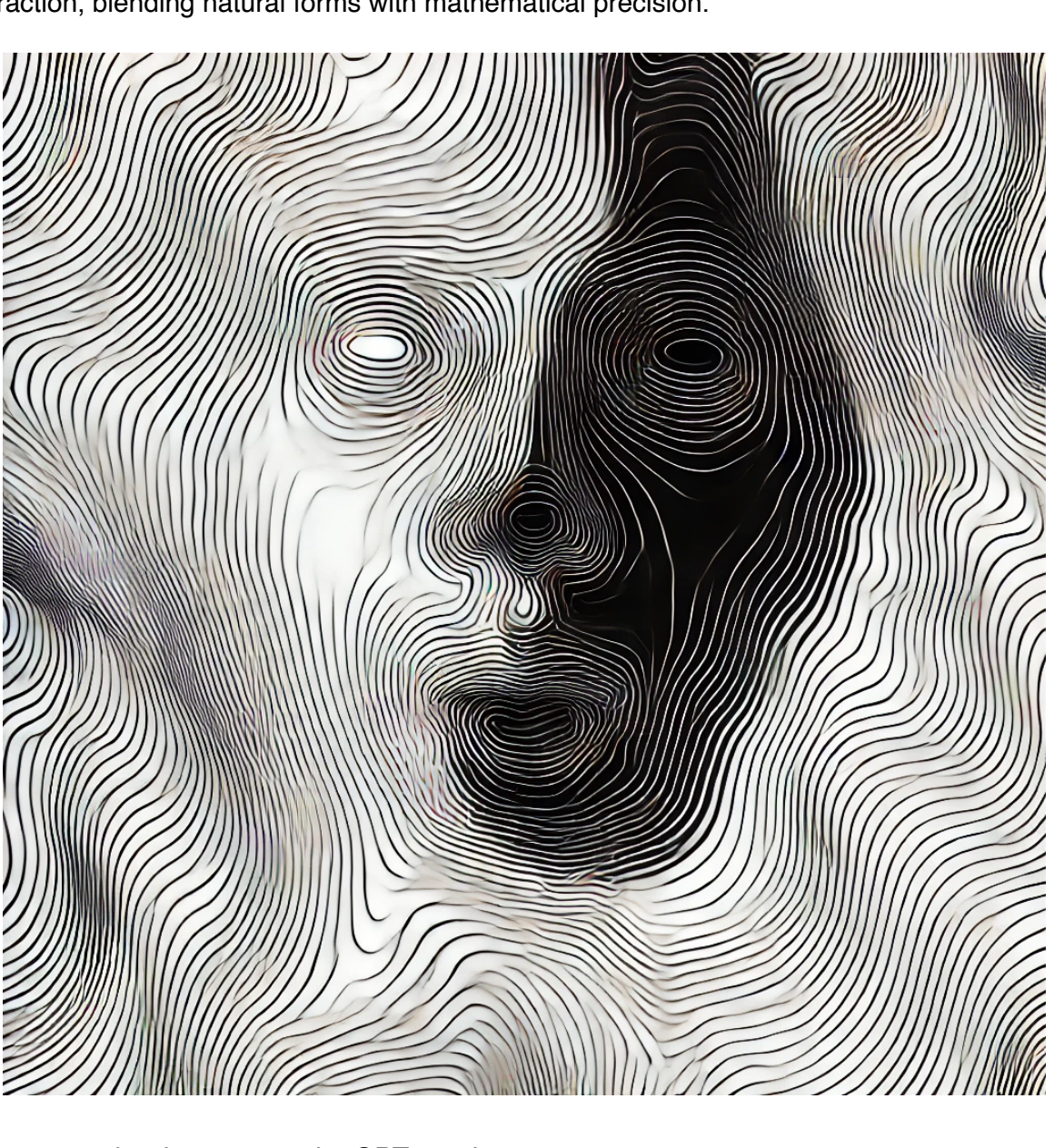
- original art





- CLIP: a white and black photo of a face in a wavy pattern of lines and curves on a white background, smooth face feature, an ambient occlusion render, generative art, Bedwyr Williams Steps: 40, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 585850058, Size: 1280x720, Model hash: 31e35c80fc, Model: sdxlbase_1.0, Version: v1.6.0

- DALLE: A surreal monochromatic artwork featuring a topographic, contour-line design that forms the abstract outline of a human face emerging from or submerged within undulating, wave-like terrain. The lines are evenly spaced, creating a flowing, hypnotic rhythm with a sense of depth and motion. The overall composition is minimalistic and geometric, with a high-contrast black-and-white palette. The artwork evokes themes of isolation, fluidity, and abstraction, blending natural forms with mathematical precision.



- Same prompt but from a new chatGPT session

- Midjourney
- A --chaos 50 --ar 16:9 --style raw --stylize 200 --weird 1500 --v 6. -