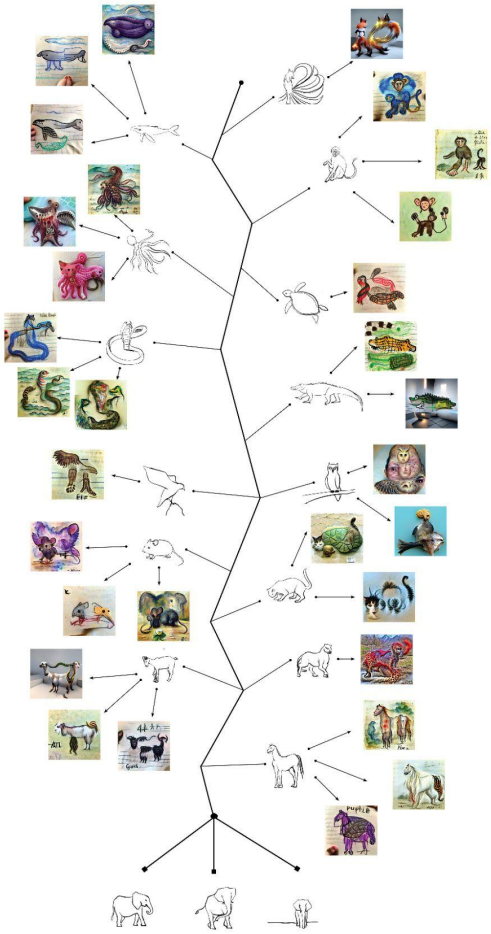
ART AND MACHINE LEARNING
CMU 2022 SPRING
PROJECT 2

A Dream of Fantasy Species



Jiaying Wei
Master of Science in Computational Design, School of Architecture
Zhenfang Chen
Master of Science in Computational Design, School of Architecture
Guanzhou Ji
PhD in Building Performance and Diagnostics, School of Architecture

Date: 2/28/2022

DESCRIPTION

The objective of this study is to study: (1) how text inputs affect the generated image; (2) the performance of the machine learning model in generating animals' image with different features; (3) explore the impact of the linework image on the visual results

Concept

In our childhood we always were amazed by the fairy stories that our parents described to us and we could not help thinking about what the creatures look like. They are so mysterious and different to the animals we can see in the world. Did they really exist in the long history? The only references we have are the texts and illustrations from the book. Since the illustrations were also drawn based on the text's description and the artist's imagination, sometimes we would wonder about our very own version of the species.

In early stages, we dreamed about hybridized creatures, starting from a cat with bird's feet. We refer to the book Classic of Mountains and Seas which is a compilation of mythic geography and beasts. How were these beasts imagined by our ancestors as a part of the historical and cultural context? Why were the details selected and arranged in such ways? What do they look like? Can we visualize them? We want to give it a shot.

The concept of this project is using text inputs and framework images to generate fantasy species. The existing animals shape people's visual stereotype when seeing the texts. Machine learning approach with image generation provides ambiguity during the process of generating visual results. In the decision making process of a human artist, image composition is one of the critical elements. Because text prompt contains limited ability to guide how drawings are composed, black contour lines in white background are employed to achieve such a goal.

Technique

This study is primarily based on the Machine Learning framework - CLIP, and its modified versions (VQGAN+CLIP, CLIP-Guided-Diffusion).

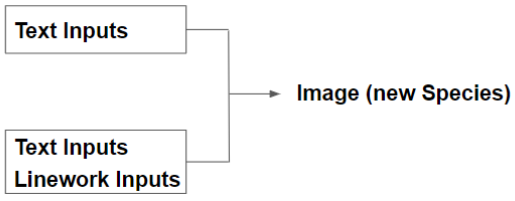


Figure 1. overview of workflow

We use the words in the same structure as the text input:

Exploration 1: ‘elephant + other animals‘ feature’ ’
(eg. fishtails, cat face, turtle shell, rabbit ears)

Exploration 2: ‘elephant + style’
(eg. unreal engine, Van Gogh, surrealism)

Exploration 3: using linework control image

Exploration 4: ‘elephant in the environment’
(e.g, environment: forest, classroom, playground)

Process and Reflection

We start playing with GLIDE (a diffusion-based text-conditional image synthesis model) to see what we can do with these kinds of image generation models that base on text prompt and familiarize ourselves with the workflow. The default epoch for running each image generation is 500.

The first step is to feed the model with the same text prompt but with different style keywords. We want to see how the keywords modify the final results.

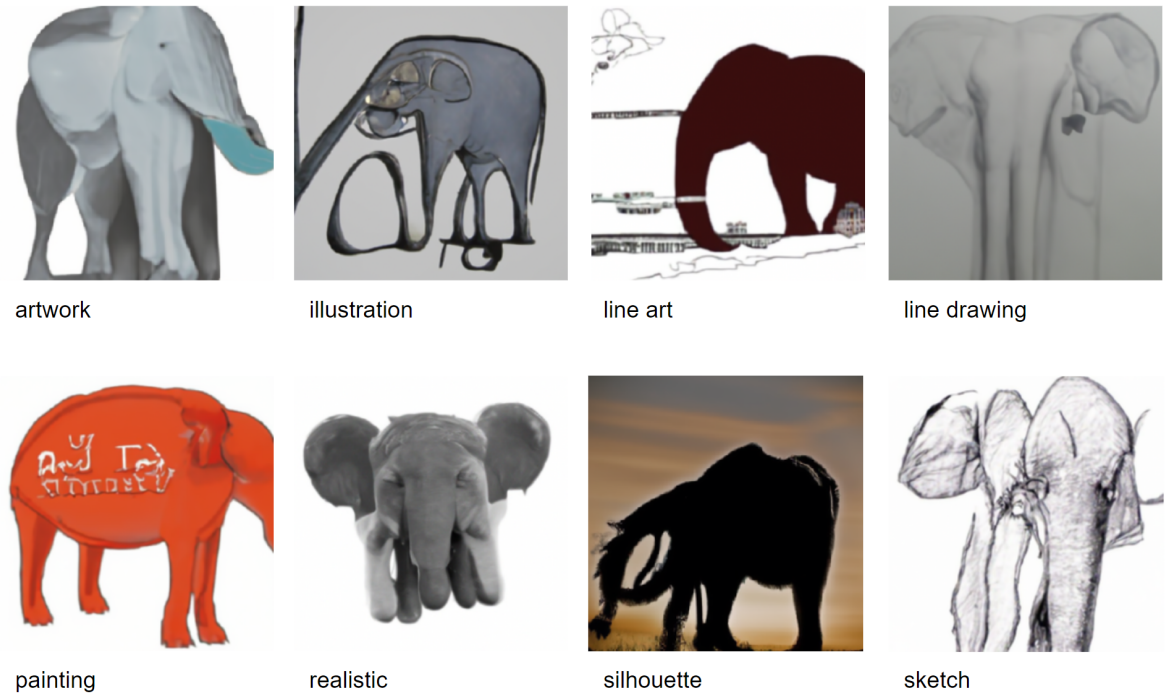


Figure 2. Testing Results: Elephant with Different Styles

We tested with many rounds and found that even with the same style keywords, sometimes the results were still not expected, but in general, the result is related to the style keyword.
Then we used the same text prompt to feed different clip variation models to see what the results can be and how differently they are.

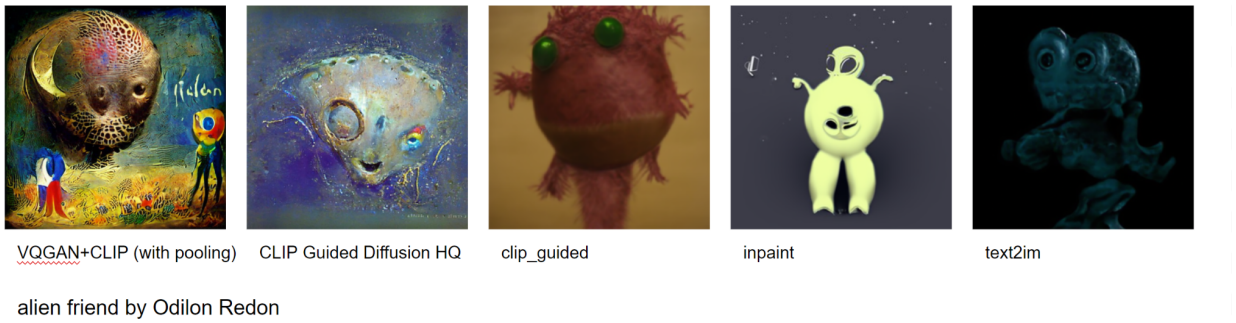


Figure 3. Testing Results: 'Alien Friend by Odilon Redon' under Different Clip Models

We also tested with several rounds and found that the results are not controllable. After careful consideration, we decided to go with VQGAN+CLIP and CLIP Guided Diffusion HQ, since we have the ability to change datasets and the resolution of the final result is better. Here, notably, the results from VQGAN+CLIP have more artistic textures based on modifiers (style keywords), but not so good at composing the subject matter; the results from CLIP Guided Diffusion HD are less artistic at times, but much better at composing the subject matter.

The outcome of the generated images with solely text input is highly variable in item localizations: animals may exhibit only partial body parts, or placed in corners/edges. So we decided to also go forth with the assistive support of the framework.

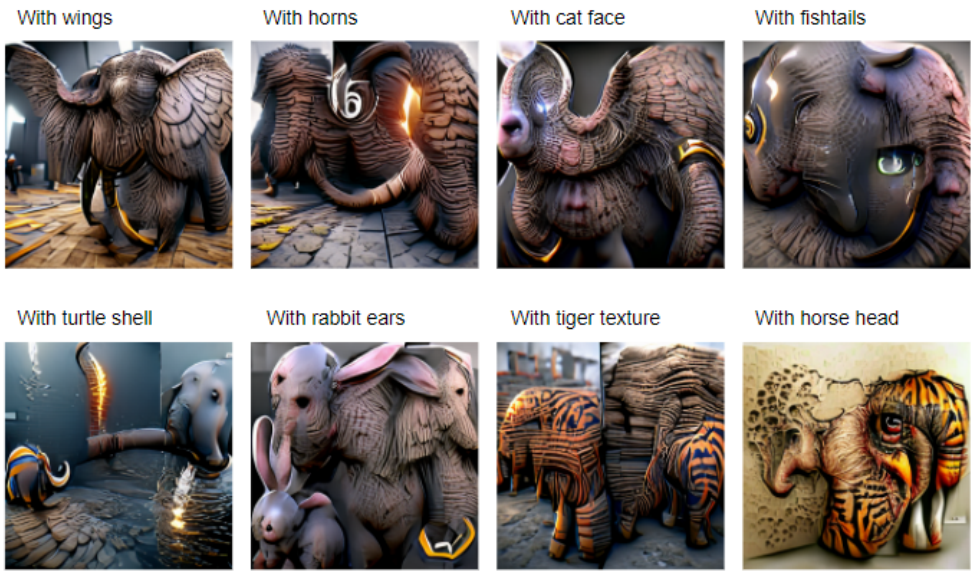


Figure 4. Elephant + other animals

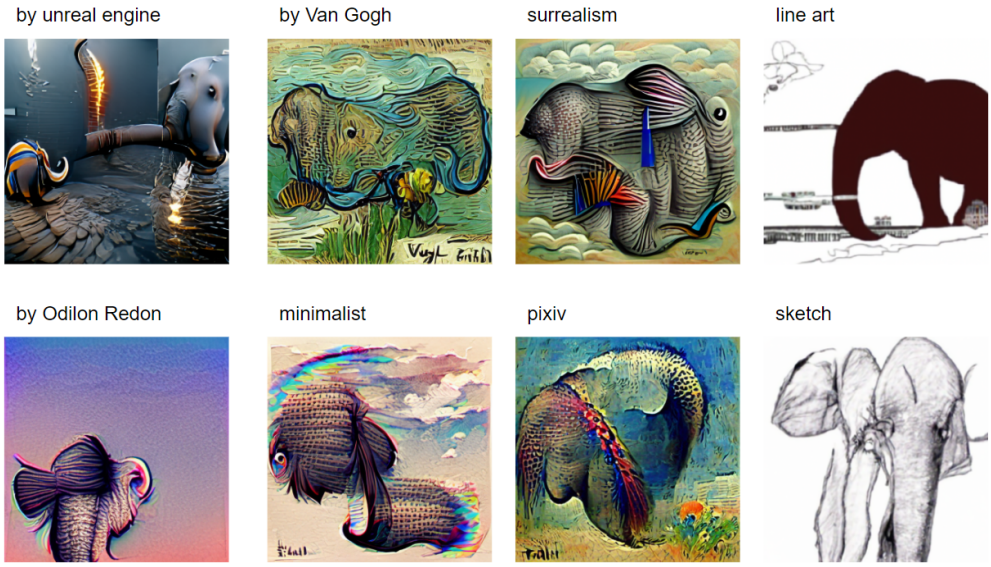


Figure 5. Elephant + style (from top left to downright:unreal engine, Van Gogh, surrealism, line art, Odilon Redon, minimalist, pixiv, sketch)

Initially We use photos as init_ images to guide the composition of the image, despite its accurate control of the focus of the image, photos have big limitations in color compositions and designer intentions. Therefore we seek another image guiding technique, which is the contourline framework that human artists frequently used for composition setups. We created 3 images for our elephant test in 1. front view 2. Perspective view 3. Side view.

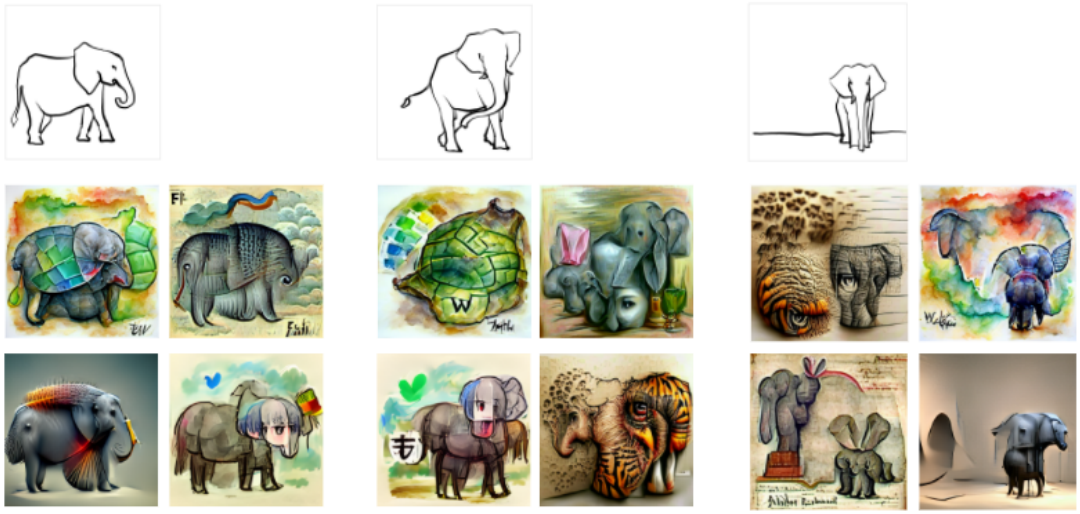


Figure 6. Linework Control for Elephant

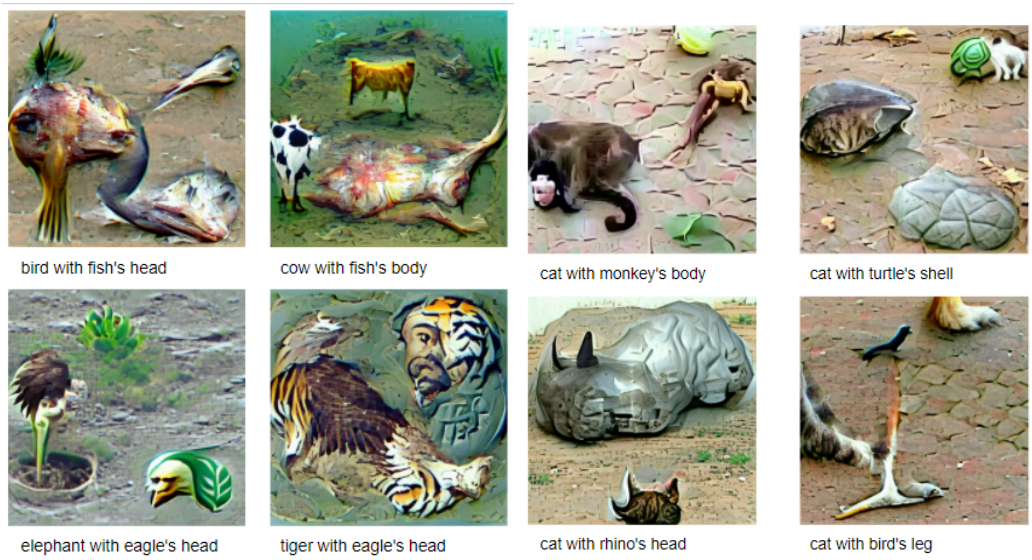


Figure 7. Left: Animal_1 with Animal_2's feature, Right: Cat with Animal_2's feature

Through multiple iterations of testing we found that results are better inside the view in which all body parts of animals are clearly displayed and separated. Also background lines create counter effects for clarity that we decided to remove them. Therefore we decided to proceed with such a technique to the framework series. We created 15 animal sketches in total as the foundation for our series of fantasy species, and simultaneously observe the following:
(1).How precisely can the contours control distinction between focused object and background?
(2).How well can the model identify the body parts, such as head, eyes, ears, limbs, tail, etc?
(3).Would the model replace/add/subtract body parts at designated areas? Eg. The fish head precisely replaced the snake head.

Figure 8. Linework Diagrams

We found that in failure cases, the thickness of the stroke, the number of iterations, and the overlapping of the body parts are the max influencer. For example in the case of the horse, with prompt ”A white horse with tiger's jaws and black tail”, images at 200 epoachs display a well balanced body shape but additional epochs start to blur the boundaries of the framework.

Figure 9. Left: epoch = 500, Right = epoch = 200, text inputs for two images : A white horse with tiger’s jaws and back tail

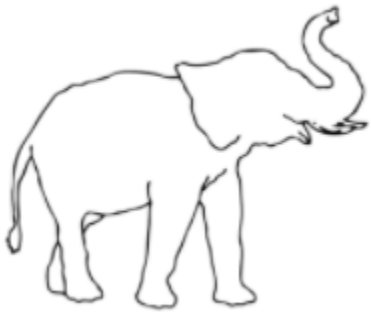


Figure 10. Linework Control Diagram for 'Elephant in the _____'

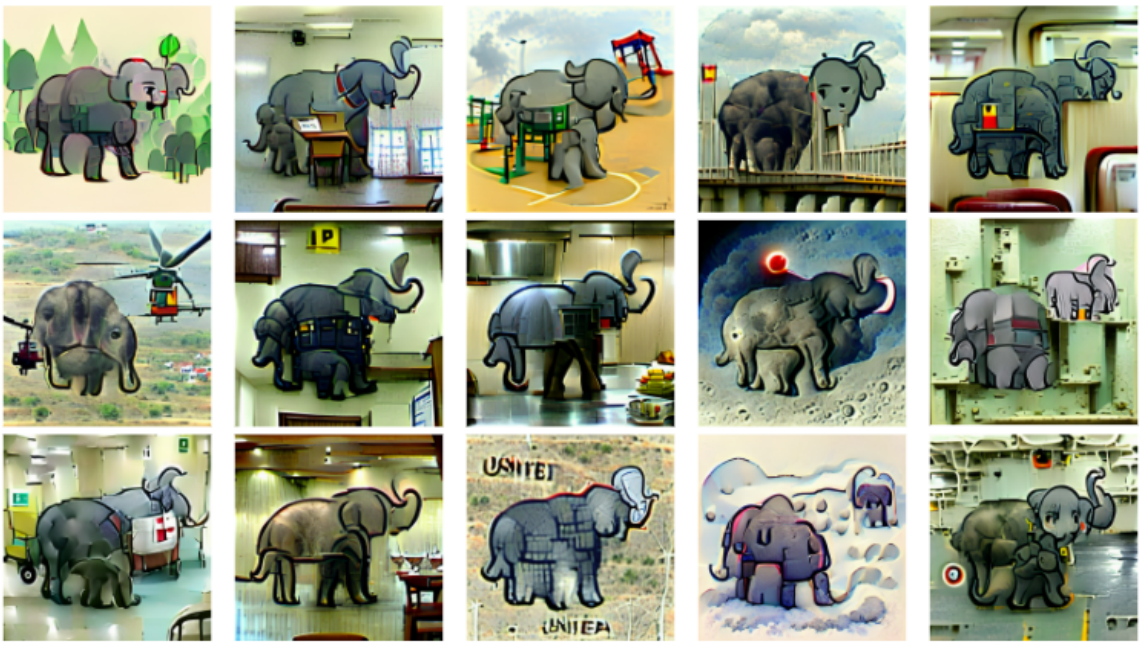


Figure 11. Elephant in the _____
(from top left to right corner: forest, classroom, playground, bridge, train, helicopter, police station, kitchen, moon, jail, hospital, restaurant, United States, snow, aircraft carrier)

The environment text input successfully creates the identifiable background texture outside the linework. However, the machine learning model cannot identify the integrated outline of the elephant which causes the unrecognizable elephant outline.

RESULT

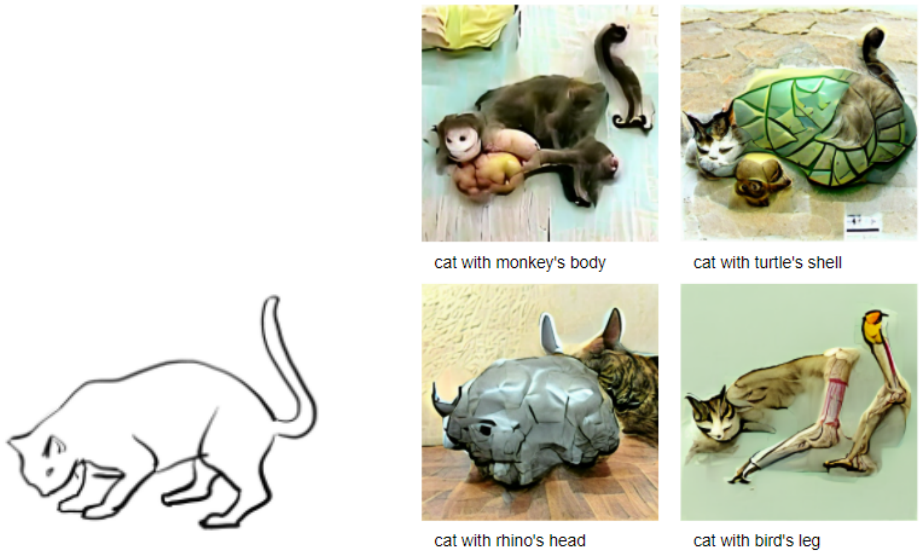


Figure 12. Left: Cat_1 Line Control diagram, Right: Cat with Animal_2's feature

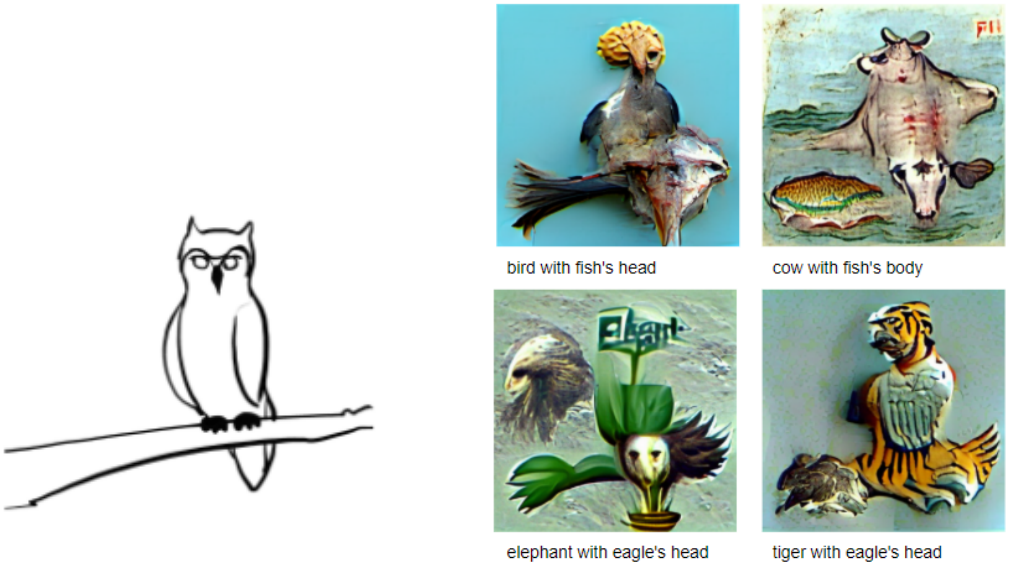
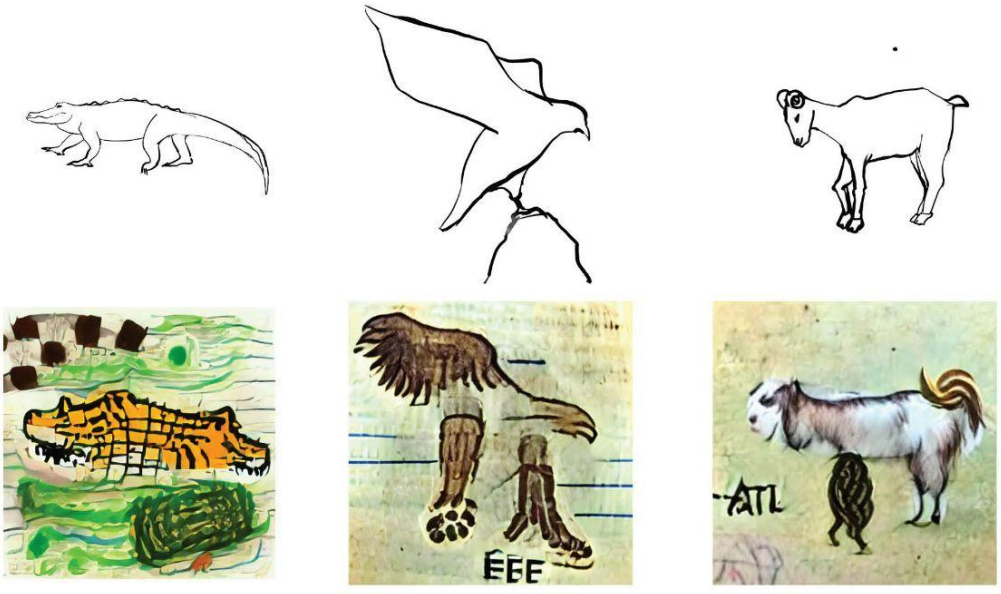


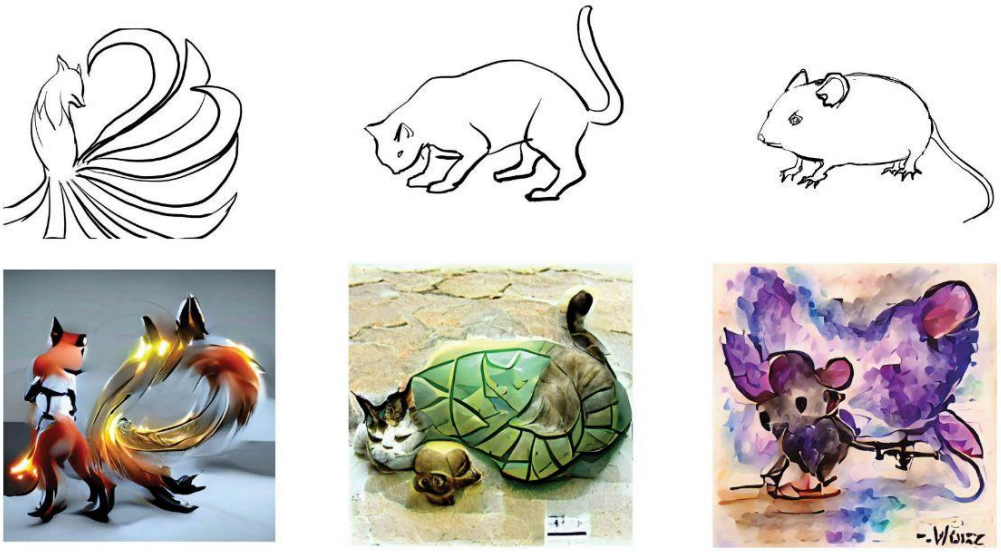
Figure 13. Left: Animal_3's Line Control, Right: Animal_1 with Animal_2's feature



A crocodile with tiger's texture

An eagle with human feet

A goat with horse tail



fox with 9 tails unreal engine

cat with turtle's shell

A mouse with purple wings watercolor

Figure 14. Animal Mixing

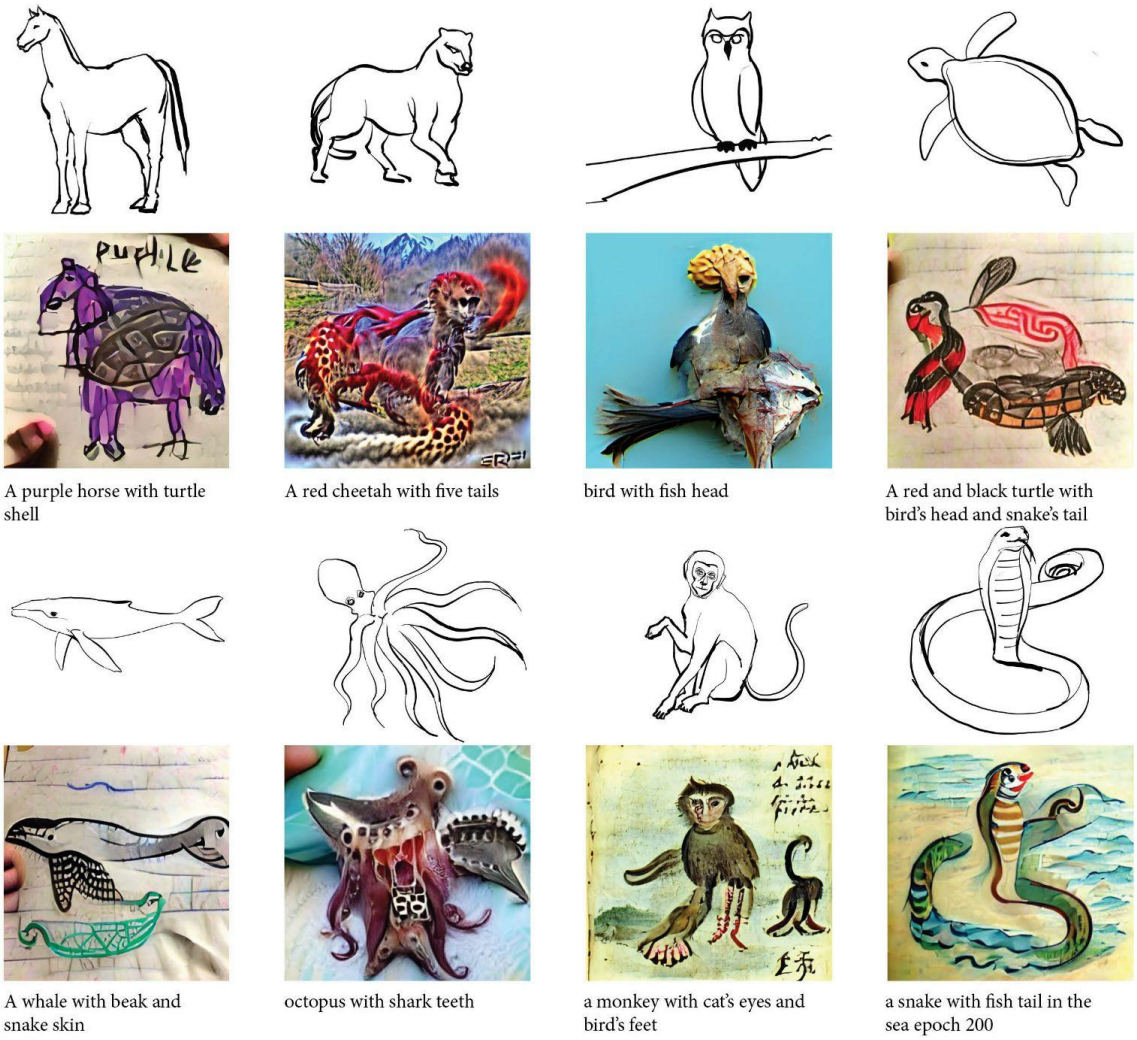
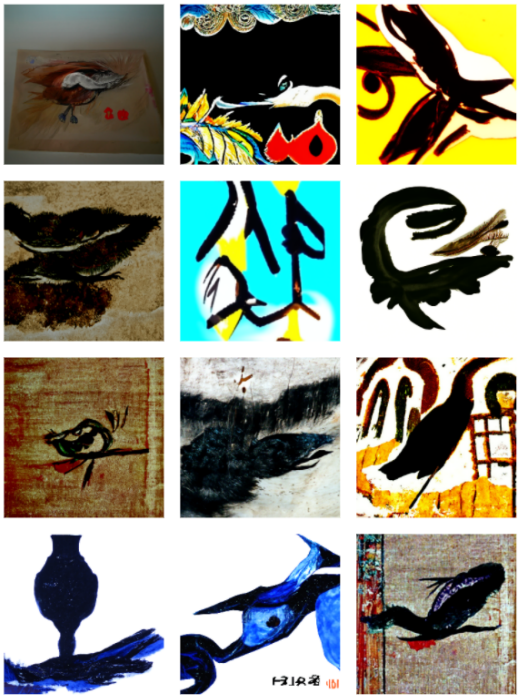


Figure 15. Animal with Animal_2's feature

Later, we started to generate images based on the text prompts from the book Classic of Mountains and Seas. We started with the text prompt 'bird with only one eye and one wing'. We did not use any initial image to guide the generation, since we think it will constrain the final results. We tested many different modifiers and many of the results are too abstract and we had to pick the ones we want. A side note here, we think the results from early training epochs are better than later epochs. Although it may not be as clear as the later images, but we think it has most of the spirit that we are looking for.



Chinese painting of two birds each with only one eye and one wing

Figure 16. Generated Image for Classic Mountains and Seas

Chinese painting of a fox like horse with horns on its back and yellow fur

Chinese painting of a pig shape creature with a head on the front and a head on the back and the whole body is black

Figure 17. Generated Image for Classic Mountains and Seas

The future development will focus on:
(1) Improve the generated image quality
(2) Improve the detailed control for generating animal's body features
(3) Test more fantasy animals and create a large image sets

CODE

(1). Generate images from text prompts with VQGAN and CLIP (z+quantize method) by ak92501 (we mainly used this one for the project).
(2). Generates images from text prompts with CLIP guided diffusion by RiversHaveWings (we also used this one for some of the images).
(3). We started playing with the CLIP model from this github repository
(4). We later tried some other variations that were mentioned in this reddit post. Some of the Colab notebooks work for us, some do not. We finally decided to use the aforementioned two notebooks since they are easy to use and very well documented.

REFERENCE

1. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2021). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. ArXiv, abs/2112.10741.