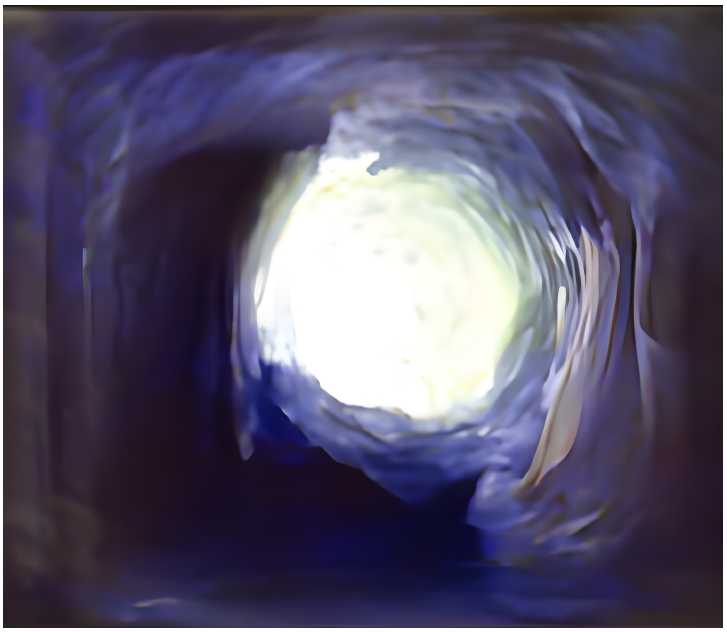
ART AND MACHINE LEARNING

CMU 2022 SPRING

PROJECT 2

The Journey to Self-Redemption



Kangning Chen, Ziyi Liu, Yinghao Ma, Ruibin Yuan

DESCRIPTION

Our project wants to raise the awareness of mental health issues and focuses on a character that initially suffers from depression and eating disorder, and later recovered by the practice of Buddhism. The entire project includes text generated by NLP text bigrams and the GPT-3 model, and pictures generated by the GLIDE model.

BACKGROUND

Our team members come from diverse backgrounds. Kangning and Ziyi are master's students of computer science major. Yinghao has an undergraduate background in mathematics and is currently pursuing a master's degree in music and technology. Ruibin is a master's student in music and technology.

CONCEPT

Many things in life can cause stress, and everyone has been under stress to some extent and at some point, in their lives. It could be a short but intense period, such as during midterm and finals or a day before interviews, or factors like uncertainty about future and being in a not so pleasant relationship can cause a mild but long-lasting feeling of anxiety. Different people have different ways of dealing with stress, some of which are healthy practices such as making plans ahead of time, while others might cause the stress to accumulate.

This project idea came into our minds when during one of our meetings, a team member started to stress eating, which we realized is a bad habit that we all do all the time. Stress eating not only doesn't solve the current problem at hand, but can potentially create other factors that lead to more stress like self-weight shaming, which is another topic that is being discussed more and more nowadays, especially among females. Although most people know it is not a right thing to do, images of a "perfect body shape" are everywhere on the internet, causing people wanting to lose/gain weight. Without a right mindset, some people might end up overeating or undereating, which in turn leads to more stress.

Therefore, we decided to depict a person with a strong sense of body management, who feels great guilt after stress eating, which leads to even more stress and overeating, and eventually suffers from depression. Despite the unfortunate beginning of the story, our best wish is that this person can overcome the hardship. We decided to use religion as "the hand" that pulls them out of the dark, specifically eastern religion, because religious practices and beliefs not only can

provide the support and tranquility they need, but are also associated with many visual artworks which we thought can improve the quality and variety of the images generated for the story. In addition, we chose eastern religion because it has some unique elements that might lead to interesting results, such as the “Yin and Yang”.

To make the story more creative, we made use of the idea from Project 1, where we gave GPT-3 a prompt and asked it to generate the rest. The generated images serve as illustrations of the scenes in the story. Our story is composed of four scenes, the first of which is an interview with the person when they have been suffering from depression for a while. The rest of the scenes are presented as a diary of the person, describing from being trapped in the overeating and self-weight shaming cycle, to the process of self-redemption through practicing religion, and finally reconciliation with food and themselves.

Through this project, we want to advocate for healthy stress management strategies and a right mindset and understanding of body management. We would like to encourage people going through similar difficulties to not be afraid to seek help using different ways. From the technical perspective, with the images generated, we want to show the model’s power of creating surrealist photography like artworks, style fusion, and object fusion.

TECHNIQUE

In terms of techniques, there are two parts to our project: text prompt generation and image generation. For prompt design, we used word clouds on uni-gram along with bi-grams that are inferred from related documents. We used GPT-3 to generate image captions and the main body of the story, which was inspired by project 1. Finally, for text-to-image generation, we used the GLIDE diffusion model.

We use the Natural Language Toolkit (NLTK) library for text processing, such as stopwords removal and creating n-grams, on datasets of four domains - religion, eastern, philosophy and depression. These n-grams, although cannot be directly input into the GLIDE model [3] to generate a good image, did inspire us with the prompt design.

Generative Pre-trained Transformer 3 (GPT-3) [2] is a high-quality language model that is trained on a large set of internet data for text generation. The full version of GPT-3 has 175 billion parameters, making it much bigger and more capable of handling a broader range of topics than the previous versions. We interacted with GPT-3 to get inspirations for better story prompts, creating captions for image generation, and most importantly, to generate the main narration of the story.

For text-to-image generation, we decided to use diffusion models for its good performance. There are many kinds of generative models that can synthesize images, including GAN, VAE, Flow-based models and diffusion models as shown in Figure 1. Among these models, GANs are potentially unstable to train and always lack diversity in generation, VAE relies on a surrogate loss and its generation is often obscure, Flow models have to use specialized architectures to construct reversible transforms, and diffusion models are proven to achieve the best results so far.

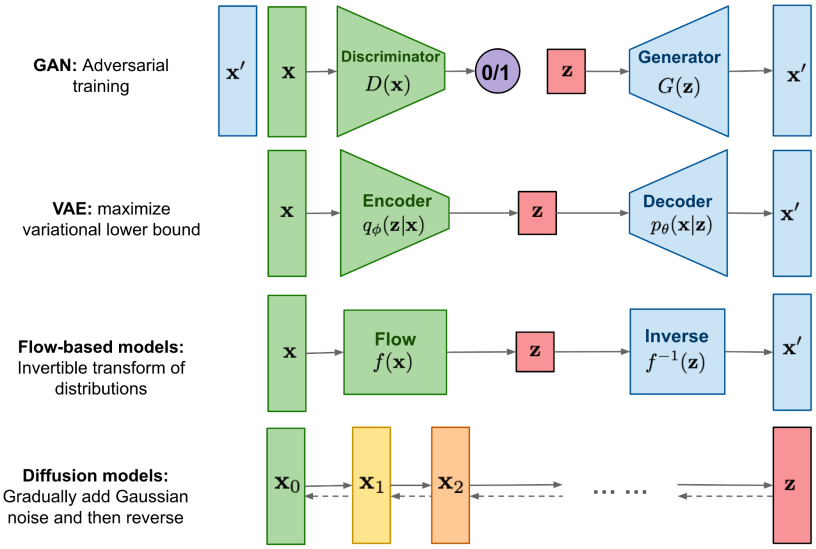


Fig. 1. Overview of different types of generative models.[5]

Inspired by nonequilibrium thermodynamics, diffusion models define a Markov chain of diffusion steps to slowly add random noise to the data x0 until it becomes complete Gaussian noise z. Like conditional GAN or CLIP guided GANs, we can use label text or text embedding which is related to the original image for training. In this way, we can obtain a generative model with text embedding or label text as input, which is a text2image model.

Specifically, the diffusion model we used is based on [6] due to the fact that it has recently been shown to generate high-quality synthetic images from text, outperforming even GANs, allowing a trade-off between diversity and fidelity, especially when trained under text model guidance e.g. guidance technique. From that, GLIDE [3] compared diffusion models for exploring image synthesis problems under text-based conditions, with training strategies including CLIP guidance. We use GLIDE for the cover generation and illustration generation which relate to the story we wrote interactively with GPT-3.

Since GLIDE can only generate 64*64 images, our cover image employs Real-ESRGAN [4] for superresolution to recover low-resolution images. This work extends the well-known ESRGAN

to a practical recovery application (i.e., Real-ESRGAN) by introducing a high-order degradation modeling process to better model the complex real world. We also consider the ringing and overshoot artifacts commonly found in the synthesis process. In addition, this model employs a U-Net discriminator with spectral normalization to improve the discriminator's capabilities and stabilize the training dynamics.

PROCESS

The Initial Attempts

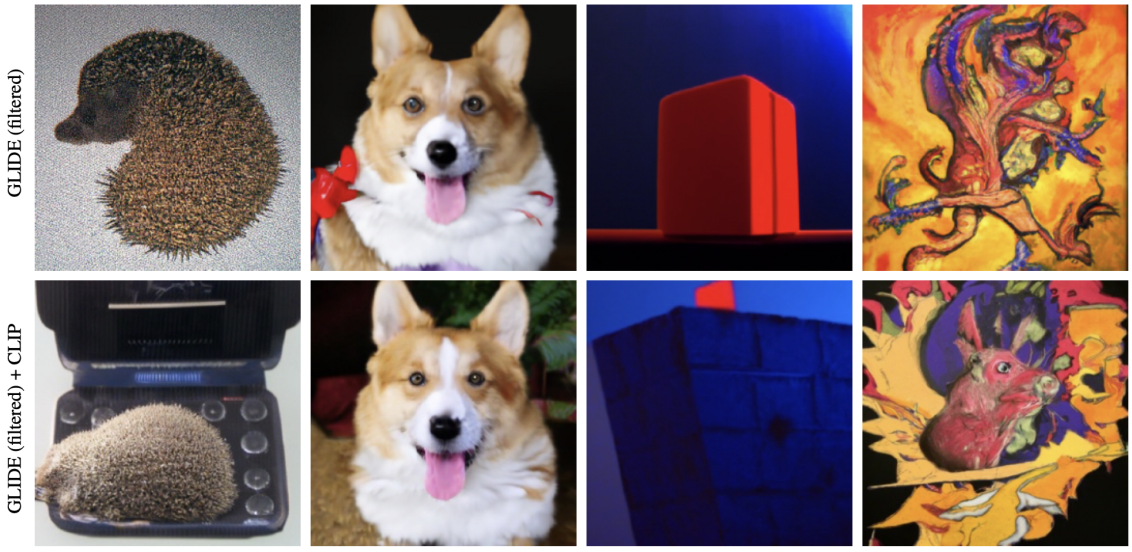
We first did not have a clear concept in mind, and we only wanted to do a text-to-image project. So to help us to form an idea, we started by testing out tools.

In recent studies, diffusion-based models have shown great success in image and audio generation tasks and even achieved better results than GANs. One of the significant breakthroughs in the text-to-image task is the guided diffusion model, GLIDE, which was also mentioned in Robert's guest lecture.

We started the project as soon as Project 1 ended. We set up an ubuntu GPU machine in a teammate’s apartment, serving a jupyterlab server to the public network so that our team can use it whenever they want. Then we quickly went through the GLIDE paper and set up the code in the server for our team usage. The initial model testing was conducted by Ruibin, and the exciting results can be found in https://helpful-newsboy-d5b.notion.site/Project-2-74dfe636a7034b3485b2a6fe4466103a

During the test, we found that even though the paper claims that the classifier-free version is better than the clip-guided version, the public available clip-guided version works better than the classifier-free version. Ruibin posted an issue in the GLIDE repo. They explained that the publicly available model is called "GLIDE (filtered)”, which is 10x smaller than the original model called “GLIDE”. Also, they applied a filter strategy to remove all humans, violent objects (weapons, etc.), and hate symbols (swastikas, etc.).

Comparison of clip-guided and classifier-free:



Some failing attempts on clip-guided (it still reflects some of the semantic, in a subtle, indirect way):

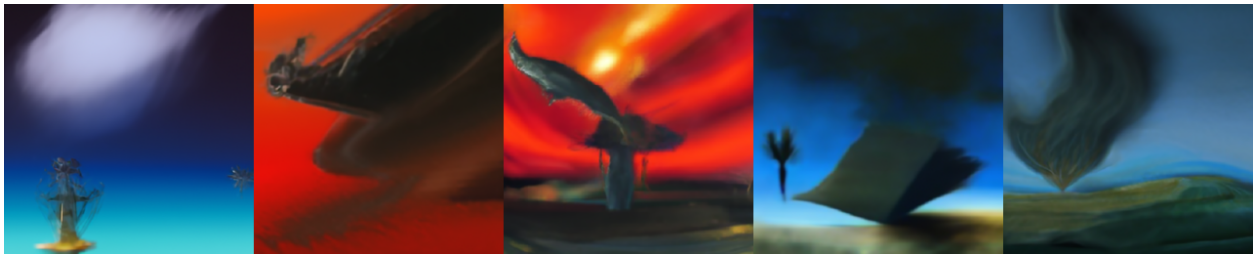
War

A flying man

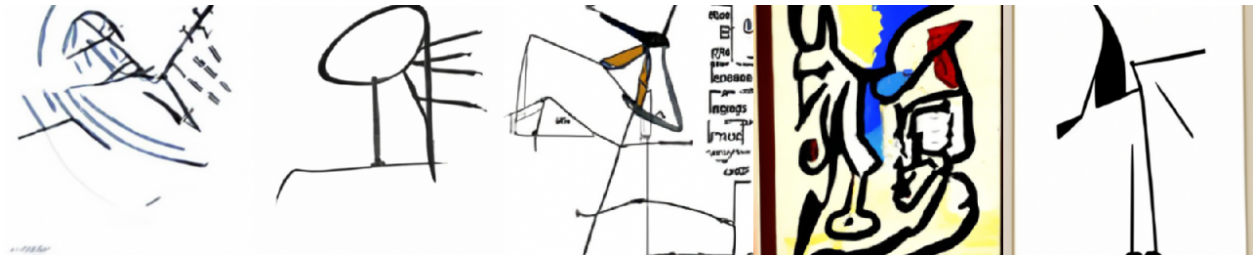
A man and a woman

More importantly, after the testing, we discovered lots of effective prompts and can even perform some text-guided style transfer.

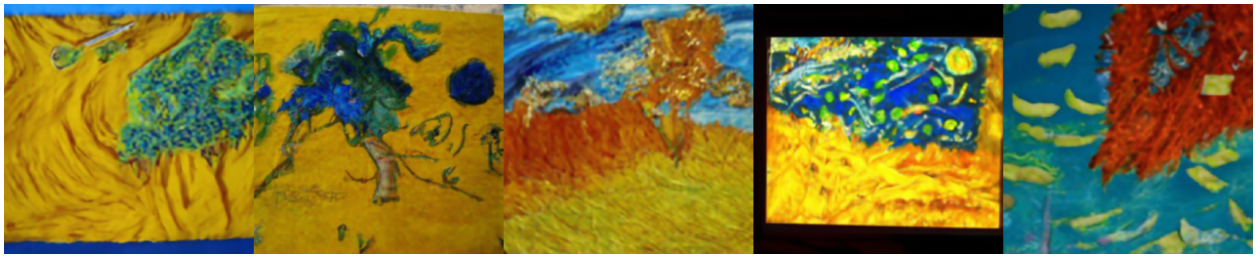
surrealist photography: int0thewind



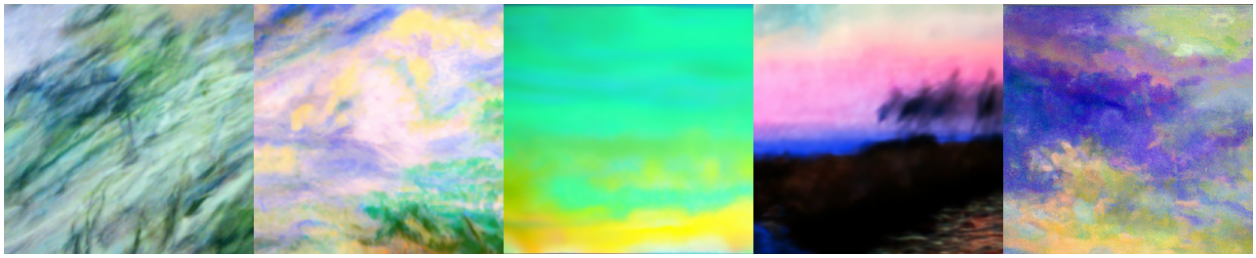
surrealist painting: int0thewind



Picasso: int0thewind



van Gogh: int0thewind



Monet: int0thewind

Qi Baishi: int0thewind

Constructing The Context

We decided to use the clip-guided version of GLIDE. After two days of brainstorming, our team decided to depict a person struggling with depression and eating disorder and later finds their peace with the help of religion in our story. More details can be found in the Concept part of our report.

Then the question is, how to tell the story? How to combine the story with images and make sure they have the same semantics?

We first started from the text because text can convey abstract concepts and complex stories efficiently, and humans can generate text easier than drawing pictures. However, we did not want to write the whole story by hand. We wanted more machine participation. So, we came up with a solution: using specific words as seeds and letting text generation models do most of the work.

We spent a night collecting text datasets from four domains: philosophy, religion, depression, and eastern. We collected works, articles, and publicly available datasets from these domains.

The combined dataset contains 15M words. Kangning helped with text preprocessing, Ziyi helped with bigram computing, Yinghao helped with word cloud visualization, and Ruibin helped with new word discovery.

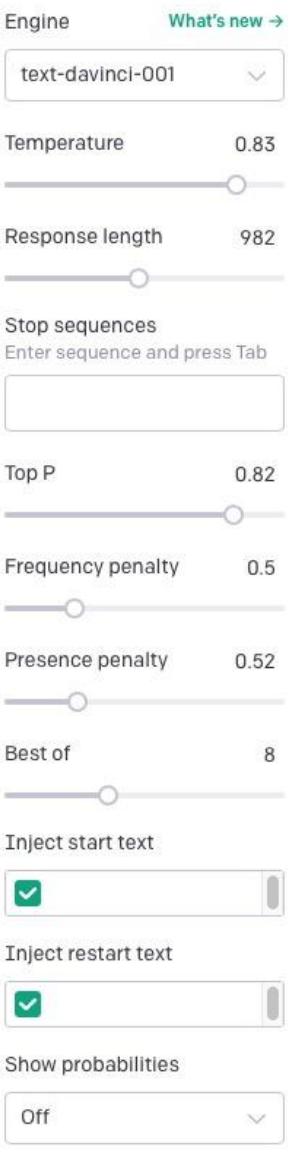
However, Ruibin and Yinghao did not reach meaningful results in a limited time, so our team decided to pick meaningful bigrams from Ziyi's results manually.

```
('want', 'die')     523 / ('suicidal', 'thought')       165 / ('attempted', 'suicide')      75
('depression', 'anxiety')       302 / ('deep', 'depression')        50 / ('struggling', 'depression')       34
('feel', 'alone')       130
('self', 'harm')        129
('feel', 'guilty')      91
('feel', 'empty')       76 / ('feel', 'worthless')      64
('stop', 'thinking')        64
('fall', 'asleep')      60
('talk', 'anyone')      60
('close', 'friend')     59
('falling', 'apart')        55
('cure', 'depression')      51
('like', 'burden')      51
('suffer', 'depression')        45
('like', 'failure')     44
('enjoy', 'life')       43
('eating', 'disorder')      39
('im', 'tired')     37
('year', 'relationship')        35
('cannabis', 'ease')        33
('enough', 'money')     33
('hypnotherapy', 'healing')     3
```

Some of the manually picked bigrams.

To generate paragraphs from seeds, we used GPT3. It was not easy to construct meaningful and appropriate stories with only a couple with seed words. We had lots of failures, and we only included our final solution. The key to a successful construction is to guide the generation throughout the whole paragraph. In other words, humans need to take control of the generation direction, by introducing prompts every several sentences, and incorporating the seed words in the prompts. One of the effective ways is “chatbot”. In the “chatbot” setting, humans chat with the machine by asking and answering the questions back and forth. The first section of our final results features an interview with our protagonist, which was constructed in this paradigm.

Also, we tuned the parameter settings of GPT3 to promote longer, detailed and novel results.

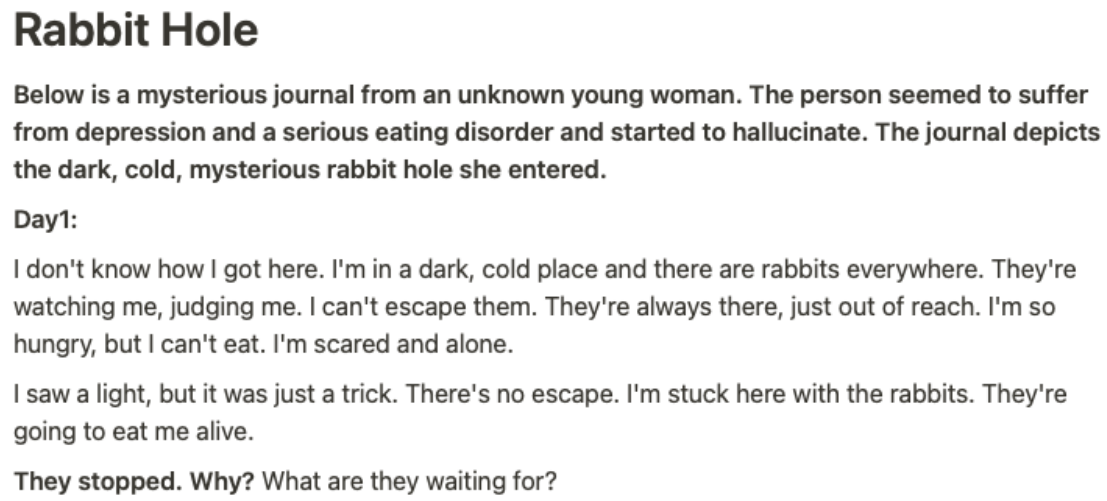


Generating Images with Context

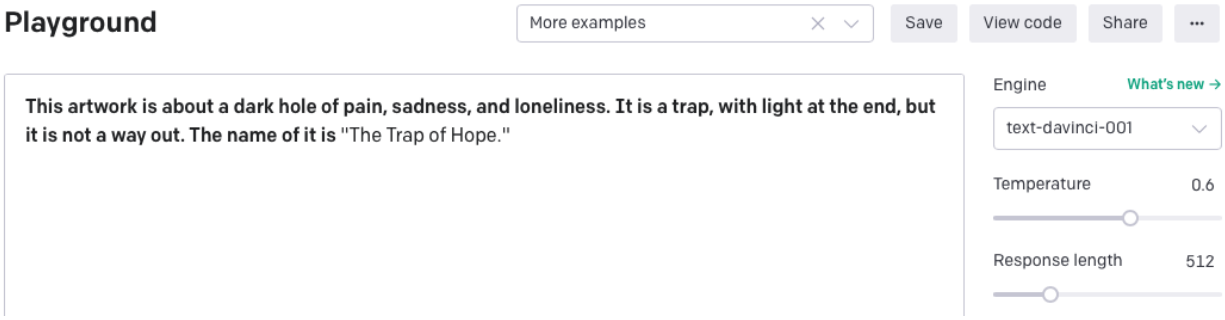
To make sure we can generate images from the context, we need to guide the GPT3 to describe objects or scenes that can be drawn. For example, prompts like “I picture myself ...” were used.

In the end, we came up with a paradigm for context-artwork pairs generation, our working solution:

1. Interact with GPT3, use prompts and seed words to construct a story interactively. The story will depict certain scene or objects. The result will serve as the context of the artwork. Below is an short example.



2. Summarize the scene or objects with a few sentences (summarization can also be done by GPT3), generate an artwork topic using GPT3. Example shown below.



3. Construct prompts for GLIDE and generate images. The format of the prompts will be {artwork name}: {description}. The description can be a sentence describing the scene or objects, along with a style transfer prefix like “a surrealistic painting of”.

4. Repeat 1~3 a few times, and pick the most interesting results. For example, a provocative context, an image with good composition, and the right tone.

5. (Optional) Do super resolution to the chosen output image with Real-ESRGAN.

With this method, the 4 members of our team constructed 4 paragraphs respectively with certain topics and corresponding artworks.

REFLECTION

Through many trials and failures, we have reached our final result with four scenes and each scene has its own visual style theme and topic. For text generation, we tried many different ways to guide GPT-3 and found that GPT-3 gives the best results when giving a pre-defined setting such as “chatbot” or indicating a theme like diary/journal in the text prompt. Therefore, we split the work into four different scenes so that GPT-3 can generate the best texts for each stage.

For the visual generation, we used GLIDE with text prompts in the format of “style: objects.” This is because GLIDE gives more striking, visually-attractive images when we give it a text prompt that defines both the visual styles and objects: for example, a surrealistic photograph of dark holes; a combination of snake and noodles; a van Gogh painting of a church. The images generated are fascinating: they do not exist in the world but mimic existing objects in the world: for example, a “noodle” snake, a van Gogh painting that van Gogh had never painted.

We are very satisfied with our final results because 1) the results are entirely original, visually appealing, and thought-provoking, 2) we were able to combine three different machine learning models (GPT-3, GLIDE, and bigrams). Unlike regular style transfers, our base images are generated from topics we define instead of existing images. Our prompts are mainly automatically generated from text data bigrams and GPT-3, for which we tried many different combinations of model parameters and prompt styles. This process is very educational and deepened our understanding of those models.

Through this project, we also noticed several issues in these models. For the GLIDE model, we found its visual data are biased towards English, western elements. For example, there is not much source data for Buddhism, so the results are monotonous and of poor quality, if the prompt is Buddhism-related. Also, the GLIDE model can only generate non-human/non-sensitive images, so if any human-related prompts are given (human, angel, war, etc.), the results are unpredictable and unrelated to the prompt. GPT-3 also censors sensitive words and adds difficulty to our creative process.

Those trials and obstacles also lead us to new ideas, that in future projects, we could train a complete set of parameters on our own if we want to generate more customized and uncensored results.

RESULT

Our results have four scenes:

1) Dark hole: the protagonist suffers from depression and pictures herself in a dark hole that they are afraid to leave..
2) A vicious cycle: the character is trapped in the stress eating cycle and associates food with negative objects and feelings.
3) Fusion rejuvenation: the character is recovering from depression and the food disorder is relieving through the practice of Buddhism.
4) Rapprochement foods: after recovery, seeing food or arts on food no longer make the character feel depressive or anxious.

The initial attempts can be found at:

https://helpful-newsboy-d5b.notion.site/Project-2-74dfe636a7034b3485b2a6fe4466103a

The final result can be found at:
https://helpful-newsboy-d5b.notion.site/Project-2-Final-Result-7d54011055aa479e977a174d46360340

CODE

https://github.com/a43992899/10615-p2

REFERENCE

[1] Eric j. Heller, Transport 2, 2000
https://www.thedailybeast.com/larry-nassar-judge-aquilina-sentencing-victim-death-warrant?ref=scroll)

[2] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[3] Nichol A, Dhariwal P, Ramesh A, et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models[J]. arXiv preprint arXiv:2112.10741, 2021.

[4] Wang X, Xie L, Dong C, et al. Real-esrgan: Training real-world blind super-resolution with pure synthetic data[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1905-1914.

[5] What are Diffusion Models? https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

[6] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[J]. Advances in Neural Information Processing Systems, 2021, 34.