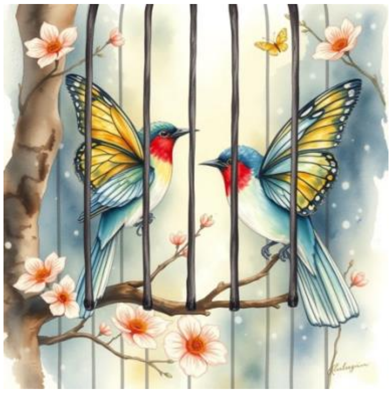
Josh Yang’s MP3 submission.

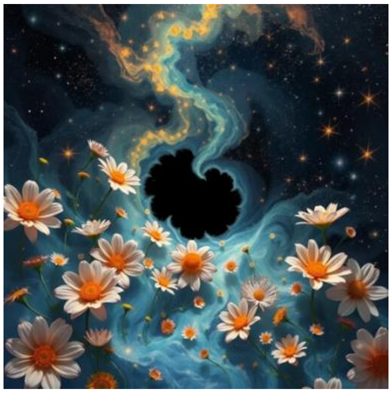
Part 1.

Task 0.1:

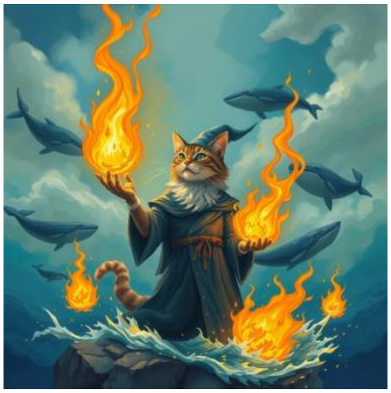
Generated with FLUX.1 Shnell:



(1) “We two alone will sing like birds i’ the cage: And pray, and sing, and tell old tales, and laugh at gilded butterflies, watercolor illustration”



(2) “flowers melting into a cosmic black hole swirling around the cosmos like stars, fantasy”

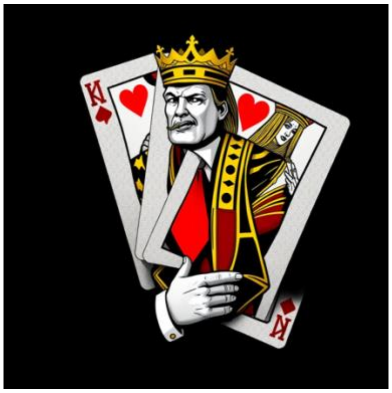


(3) “A gouache cat wizard casting fireball hexes surrounded by whales swimming in the sky, painterly”

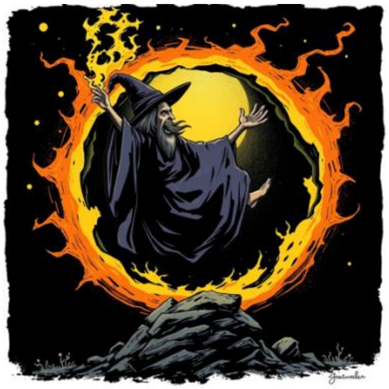
I believe my images demonstrate generalization because (1) is a Shakespeare quote that doesn’t have any sort of ground truth image but we got an image that matches the romance in his words and gives birds butterfly wings which is probably not in the dataset. (2) and (3) are a hodge-podge of concepts that don’t exist together, like cosmic flowers and aeronautical whales, and we still get reasonable depictions of what these may look like.

I believe diffusion models are capable of this behavior because, intuitively, guiding the score with language conditioning (i.e. via CLIP) means we want the image embedding to match the text embedding. So somehow the key words such as “cat”, “fireball”, “whale”, and “sky” need to visually appear somewhere in the diffused image. Even though there is no data with all of them combined, our model’s vast dataset makes sure it has seen them all individually (which is why data size is so important to hit rarer data points like “gouache”) and then it figures out a way to synthesize them all together. The diffusion sampling process makes sure we get a nice image from the distribution while matching the score objectives which makes it a super powerful technique.

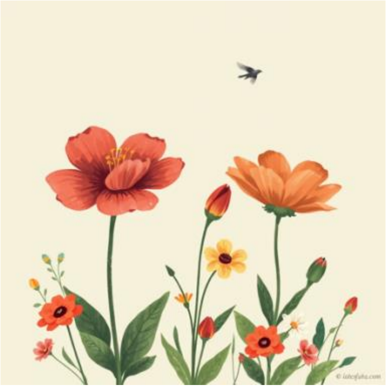
Task 0.2:



(1) “a suit of Trump playing cards”



(2) “a witch holding up a black hole, illustration”



(3) “пастораль 节日, illustration” (translated “pastoral holiday”)

Here for (1) I tried using double language where it could be read as either A) a set of cards from the same suit or B) Trump’s suit playing with cards. Words have multiple meanings and introduce ambiguity which can confuse CLIP in pathological constructions. Similarly with (2) a black hole can be either A) a celestial object or B) a hole that is black and the diffusion model gets confused. In these two cases I think current diffusion models have difficulty understanding ambiguous prompts due to a fundamental problem of language not being exact. (3) fails with a foreign language prompt which is interesting to see. I think this is a problem with the language conditioning moreso than the actual diffusion modeling approach. Having more data of ambiguous labels in both realized cases would definitely help as it can see how context influences our human ability to disambiguate. And more accurate foreign language data would be good too (a lot of auto-translated stuff is wrong).

Task 0.3:

The biggest problem with directly using these image generation technologies is that they have biases, i.e. towards normative views of the world. These models display bias towards men, Caucasian features, Western aesthetics, and more—simply because the data they are trained on contain this bias. For instance take (2) from Task 0.2: I asked for a “witch” (female) and the model gave me back a bearded wizard (male). Collecting diverse datasets and validating representation is key to mitigating this danger because manually crafting rules to “compensate” can end disastrously (i.e. Gemini’s fiasco circa 2024). Another more broader social issue is that these image generation technologies are trained on artist data without compensation. While one can argue fair use, it is morally wrong to use millions of artists’ work, learning their styles, and then selling it as an SaaS without compensation as a means to replace them. I think a Spotify-esque model where artists can get paid royalties through diffusion generations would be a fair way to make this happen and Adobe is currently researching this process.

Positive impacts include democratizing creation of visual imagery and bringing this skill to broader masses. However, I want to emphasize that using generative models as a tool is OK but it should not and cannot replace human art. If the so-called “art” that generative models produces is “democratized” then human art as a skill dies, fails to move further, and synthetic images collapse on themselves with only the natural world to guide what images, visuals, really are. Google Images is poisoned already; when I look up historical artists generated images pop up masquerading as authentic ones. There is a distinction between using generative models as a tool for art versus creating the art entirely.

Part 2.

Task 1: DenoisingScoreNetwork loss: 0.0016

Task 2: NCSN loss: -0.0053

Part 3.

I learned to model a distribution from its samples by indirectly optimizing a neural network that models the corresponding score function. We then perturbated with varying levels of noise to help learn the less-densely-sampled areas of the distribution.

Part 4.

I referenced the slides, the website linked in the handout, and this website.

Part 5.

I liked the GIF animations code was given as part of the stencil so I can see how the denoising works over the varying levels of noise. I also appreciated getting multiple perspectives on how diffusion works from both denoising and Langevin dynamics (and coding them shows they are the same thing).