Conditional Generation for Inverse Problems and Class/Text-Based Conditioning

Adam Imdieke (imdie022@umn.edu)

I. Text to image Generation

A. (A)No Prompt

For the no-prompt generation, I used the same model as in part B, with the parameters $\omega CFG = 0.0$, num steps $= 50$, and $eta = 0.0$, with a fixed seed at 42. The resulting image is shown below.

Fig. 1. No Prompt Image Generation

B. (B-C) 5X3 prompts- Manual evaluation

I implemented a pipeline that will read from a set of prompts, and then use them to generate the images from the diffusers StableDiffusionPipeline module, with the $\omega CFG = 10$, num steps $= 50$, $eta = 0.0$, and a fixed seed at 42. The prompts for each image are included in the caption of the figures.

- Space
  - Simple: Human: 5/10, CLIP: 28.49
  - Medium: Human: 8/10, CLIP: 32.65
  - Detailed: Human: 9/10, CLIP: 33.75
- Ocean
  - Simple: Human: 9/10, CLIP: 31.39
  - Medium: Human: 6/10, CLIP: 32.93
  - Detailed: Human: 5/10, CLIP: 30.34
- Castle
  - Simple: Human: 8/10, CLIP: 29.36
  - Medium: Human: 7/10, CLIP: 31.98
  - Detailed: Human: 9/10, CLIP: 30.38
- Cyberpunk
  - Simple: Human: 8/10, CLIP: 33.70
  - Medium: Human: 8/10, CLIP: 34.95
  - Detailed: Human: 7/10, CLIP: 29.07
- Cat-Bird
  - Simple: Human: 2/10, CLIP: 25.97
  - Medium: Human: 2/10, CLIP: 27.57
  - Detailed: Human: 6/10, CLIP: 37.96

Fig. 2. Space. Left: Simple - "An astronaut in space" (Human: 5/10, CLIP: 28.49). Center: Medium - "An astronaut in a spacesuit floating above Earth" (Human: 8/10, CLIP: 32.65). Right: Detailed - "A single astronaut in a shiny white spacesuit drifting serenely against the stars in the sky. There is a silent planet below with swirling clouds and blue oceans, with their ship orbiting in the distance" (Human: 9/10, CLIP: 33.75).

In almost all cases, the most detailed prompts are the best looking images. While the short prompts also did well, the longer prompts alligned better with what I was expecting from the model. I think that in most cases, if you don't know the prompt, the short and long prompts produce images of similar fidelity, which is expected as the model is trained to approach the image manifold similarly, even without a prompt as seen in part A.

The CLIP and the human scores are pretty well alligned in terms of relitive changes. While the scale of the measurements was not alligned well, when the human score

Fig. 3. Ocean. Left: Simple - "A coral reef" (Human: 9/10, CLIP: 31.39). Center: Medium - "A colorful coral reef with tropical fish of and sunlight filtering through the water" (Human: 6/10, CLIP: 32.93). Right: Detailed - "An underwater coral reef that has all sorts of life, with many fish and sharks swimming around. It has bright corals of all colors and shapes, with sunlight filtering through the clear blue water from above." (Human: 5/10, CLIP: 30.34).

Fig. 4. Castle. Left: Simple - "A medieval castle" (Human: 8/10, CLIP: 29.36). Center: Medium - "A lively medieval castle surrounded by a moat and lush greenery" (Human: 7/10, CLIP: 31.98). Right: Detailed - "A beautyful german day, with a large castle made of stone, with a few vines climbing up the spires. The vilage around the castle is full of life, with people walking around the market." (Human: 9/10, CLIP: 30.38).

changes, the clip score will usually also have a similar change in score, at least in terms of magnitude.

C. (D) Negative Prompts

I chose the prompt: "An astronaut in a spacesuit floating above Earth" with the negative prompt: "blob, blurry, low-definition, water, clouds, fingers" because the original image had a odd blob on the plannet, and I wanted to see if it could remove the fingers. The results are shown in figure 7.

CFG Scale vs CLIP Score results:

- CFG 0.0: CLIP Score 32.8%
- CFG 2.0: CLIP Score 33.9%
- CFG 5.0: CLIP Score 33.6%
- CFG 8.0: CLIP Score 35.6%
- CFG 12.0: CLIP Score 33.7%
- CFG 15.0: CLIP Score 33.0%

At CFG 0, the image is very noisy and low quality, and it does not look like it is on the image manifold. As the CFG increases, the astronaut looks more realistic, where at 5 it looks the best. At 5, there is no longer any water on the plannet, and the fingers are gone as was in the

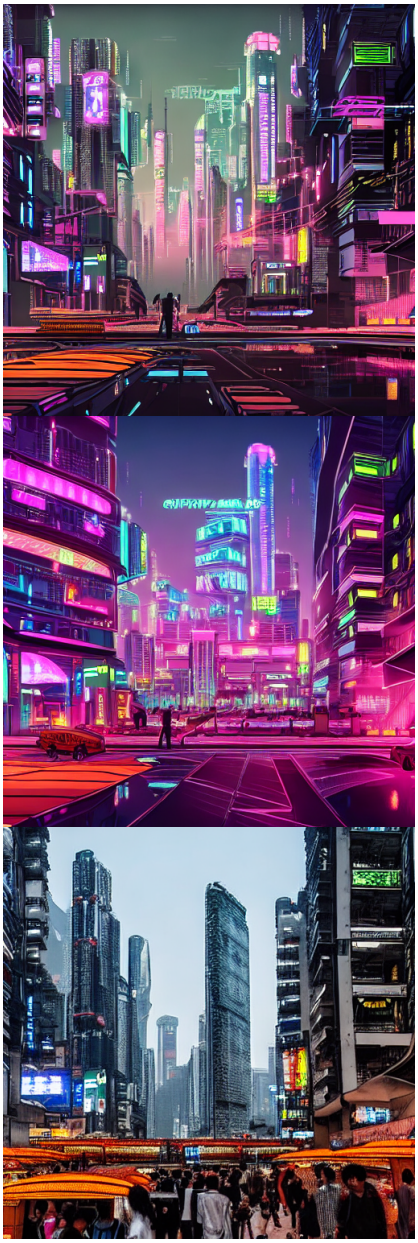

Fig. 5. Cyberpunk. Left: Simple - "A cyberpunk city" (Human: 8/10, CLIP: 33.70). Center: Medium - "A distopian cyberpunk city, with neon lights and flying cars." (Human: 8/10, CLIP: 34.95). Right: Detailed - "A breathtaking cyberpunk megacity that has bustling streets filled with people and vendors. The skyline has many towering skyscrapers, and there are futuristic flying cars." (Human: 7/10, CLIP: 29.07).

negative prompt. At higher CFG, the image becomes a bit degenerate, where the saturation increases, where at 15 the planet is no longer round. This makes sense because at high CFG, there is too much emphasis on the prompt, potentially driving the image off the manifold.

Fig. 6. Cat-Bird. Left: Simple - "A cat with a bird body" (Human: 2/10, CLIP: 25.97). Center: Medium - "A chimera with the body of a cat, wings of a bird." (Human: 2/10, CLIP: 27.57). Right: Detailed - "A beautiful chimera creature that has the body of a maine coon cat, with large majestic wings of an eagle." (Human: 6/10, CLIP: 37.96).

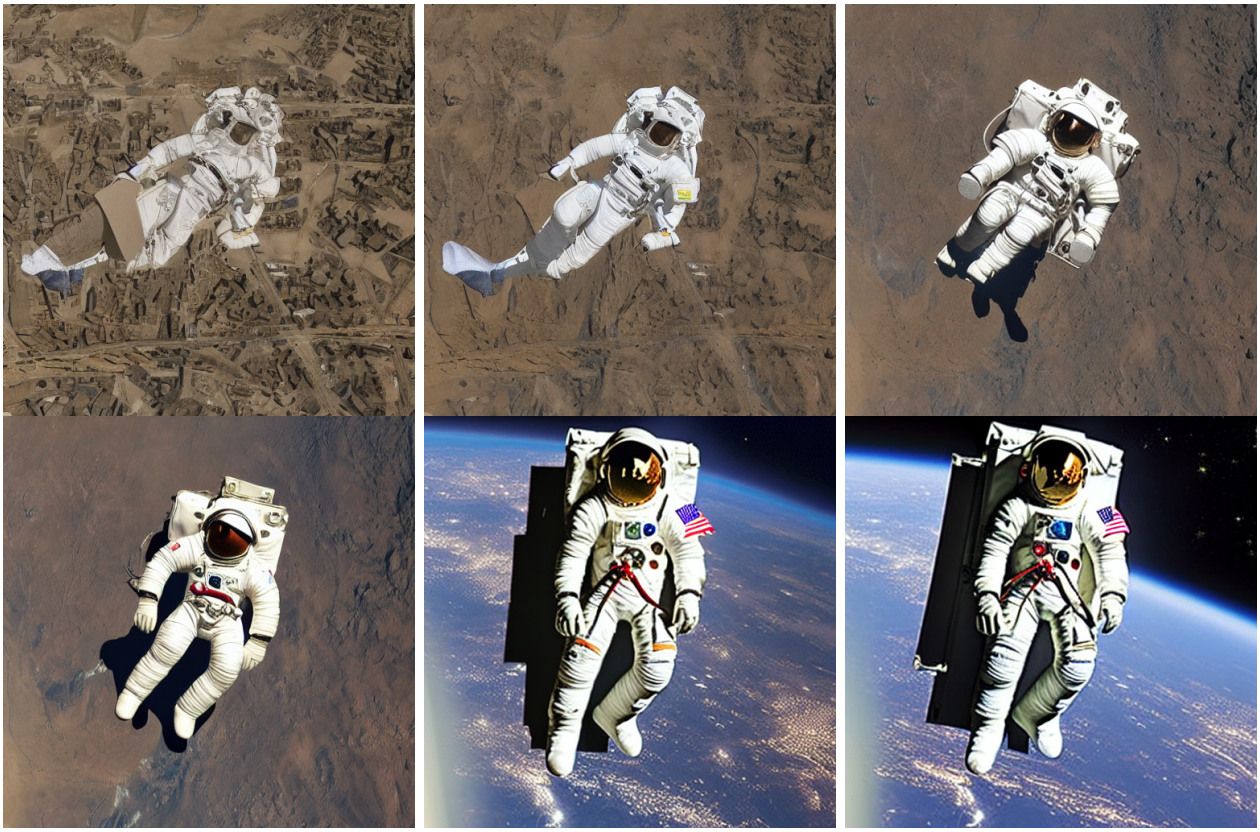

Fig. 7. Negative Prompt Results for "An astronaut in a spacesuit floating above Earth" with various CFG scales. From top left to bottom right: CFG 0.0 (CLIP: 32.8%), CFG 2.0 (CLIP: 33.9%), CFG 5.0 (CLIP: 33.6%), CFG 8.0 (CLIP: 35.6%), CFG 12.0 (CLIP: 33.7%), CFG 15.0 (CLIP: 33.0%).