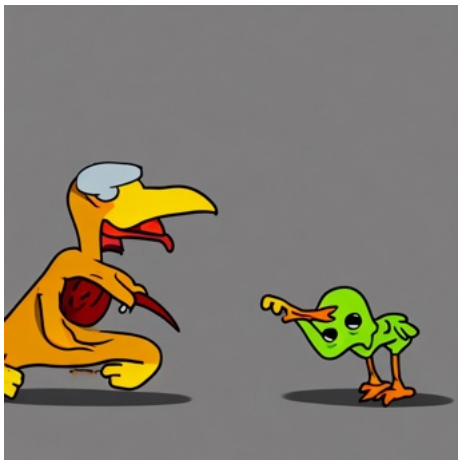
Assignment on Stable Diffusion

1. Basic Stable Diffusion

StabilityAI: Image of a duck fighting with a zombie

Here, I am using the stabilityAI's stable-diffusion, stable-diffusion-2 model from Hugging Face. Although stabilityAI has the latest stable-diffusion-3 models, it takes a significant amount of time and memory to load all the model, pipeline components, checkpoints, tokenizers, and other prerequisites. This ultimately results in a session crash, so I had to use the older version. The image generated with stable-diffusion-2 fails to include all the prompt details with accuracy.

		
Prompt: Image of a duck fighting with a zombie in Simpson style cartoon	Prompt: Realistic image of a duck fighting with a zombie in 4K with vibrant colors	Prompt: Realistic image of a duck fighting with a zombie duck, both ducks are holding lightsaber, in 4K with vibrant colors

The first image seems to have included almost all the information in the prompt. However, by the word "zombie", it assumes the contender of the duck as another duck. In the case of the second image, "4K with vibrant colors" is generated but "zombie" is no longer a keyword for this image. In the case of the third image, the model struggles to incorporate all the information in the prompt.

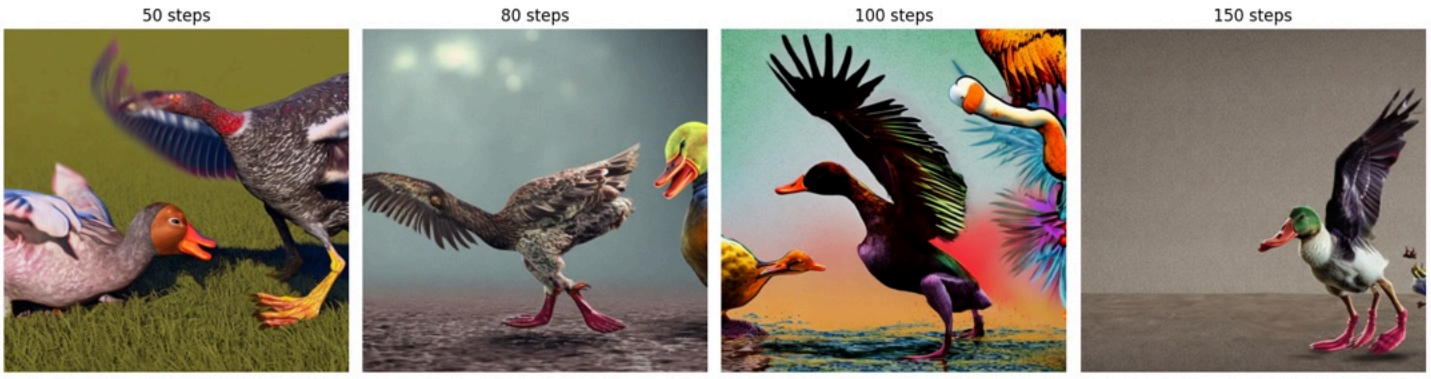
DALLE-3: Image of a duck fighting with a zombie

For comparison, I also used the DALLE-3 model from ehristoforu/dalle-3-xl-v2 which obviously generates highly detailed images, almost fully incorporating the prompt. As it is obvious from these results, the DALLE-3 has a way better quality of generated images than the stabilityAI model. However, this model being heavy sometimes led to session crashes.

 Prompt: Image of a duck fighting with a zombie in Simpson style cartoon	 Prompt: Realistic image of a duck fighting with a zombie in 4K with vibrant colors	 Prompt: Realistic image of a duck fighting with a zombie duck, both ducks are holding lightsaber, in 4K with vibrant colors

2. Messing with the Inference Process

Here, I used `stabilityai/stable-diffusion-2` as the base model and tried changing the `num_inference_steps` with a random manual seed. As shown in the figure below, the quality of image gets better with the inference steps. This is because the denoising part of the diffusion process gets more and more intricate with the increasing number of steps and consequently taking a longer time.



Prompt: Realistic image of ducks fighting in 4K with vibrant colors

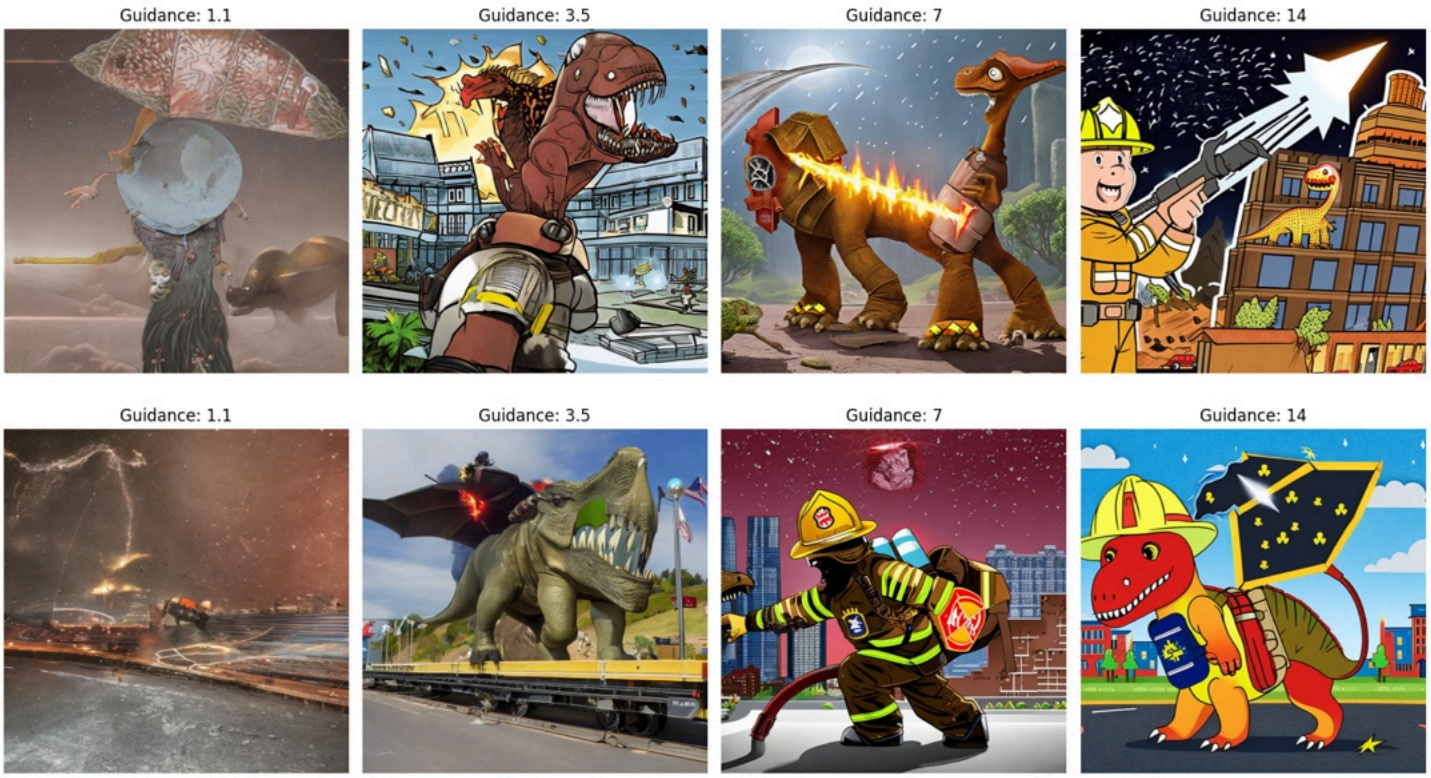
3. Classifier-free Guidance

Here, I experiment with different values of classifier-free guidance, which affects how closely the generated image aligns with the given prompt. As shown in the figure below, the low guidance values lead to creative, however, incoherent images. Moderate values are balanced, providing both creativity and structure. On the other hand, the higher values generate clear, detailed, and accurate images that align closely with the prompt. However, very high values may produce overly literal outputs, sacrificing some degree of creativity for accuracy.

Prompt: A robot baking cookies with a bunch of curious frogs watching

4. Negative Prompts

Here, I repeat the classifier-free guidance experiment but this time with negative prompts as well. The negative prompt acts as a filter that can reduce the undesirable features in the generated images. Its effect increases with the increasing guidance values..



Strength 0.2

Prompt: A dinosaur dressed as a firefighter saving a city from a meteor shower, Negative prompt: blurry, distorted, low quality, abstract

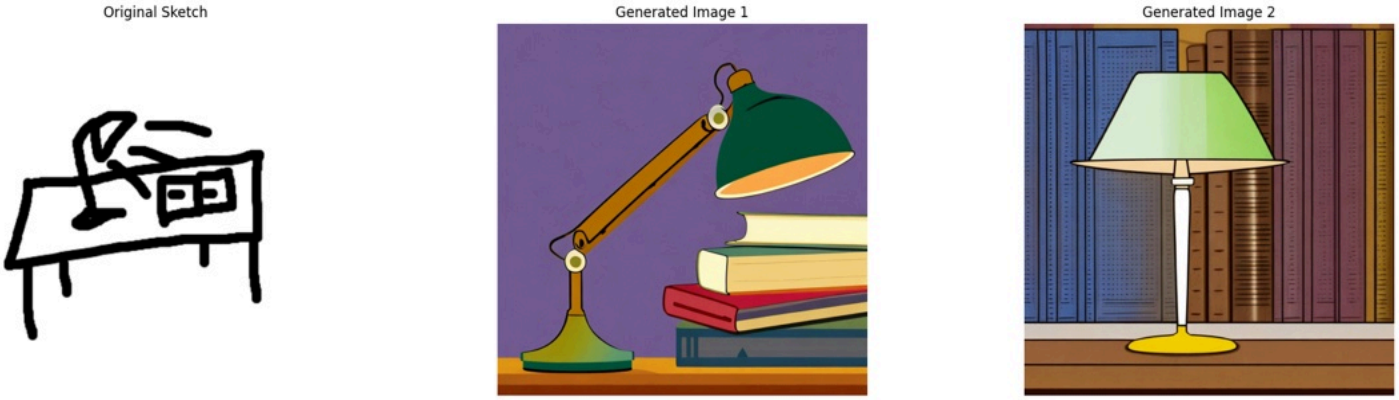
5. Image-to-image

Starting from a sketch

Prompt: A person reading newspaper, sitting on a chair, photorealistic 4K

The generated images generally follow the composition of the original sketch, which contains a person sitting on a chair and reading a newspaper. The model preserved the main elements of the sketch: a person, seated posture, and a newspaper in hand.

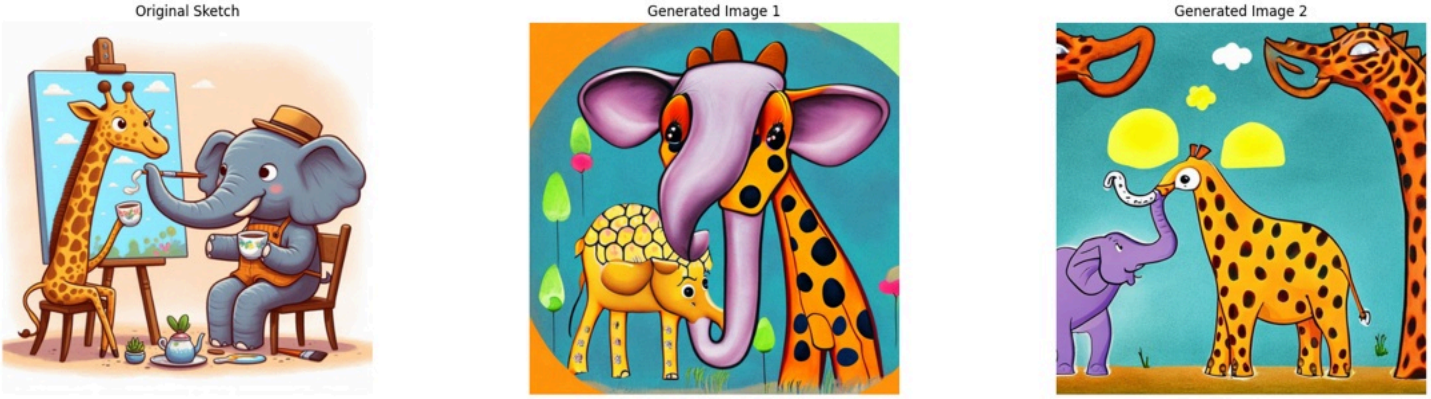
In the case of the first generated image, the background is filled with a large newspaper page as if the person is a part of the newspaper. The image has muted color, emphasizing the silhouette. Similarly, in the second generated image, the main elements of the sketch are preserved. However, the model has now added details to the person, such as a defined face, hairstyle, double column newspaper, and shadow of chair and person. Unlike the first image, the background is now plain gray, which makes the main components more prominent.



Prompt: A table lamp on a table, focused on a book, 90s style cartoon

Like the first sketch the main components of this sketch are also preserved in the generated images. The first generated image has the study lamp on a table with light focused on a stack of books while the second one has a table lamp on a table without any books. Unlike the first generated image, which has a blank background, in the second image, the model added the books in the bookshelf as the background.

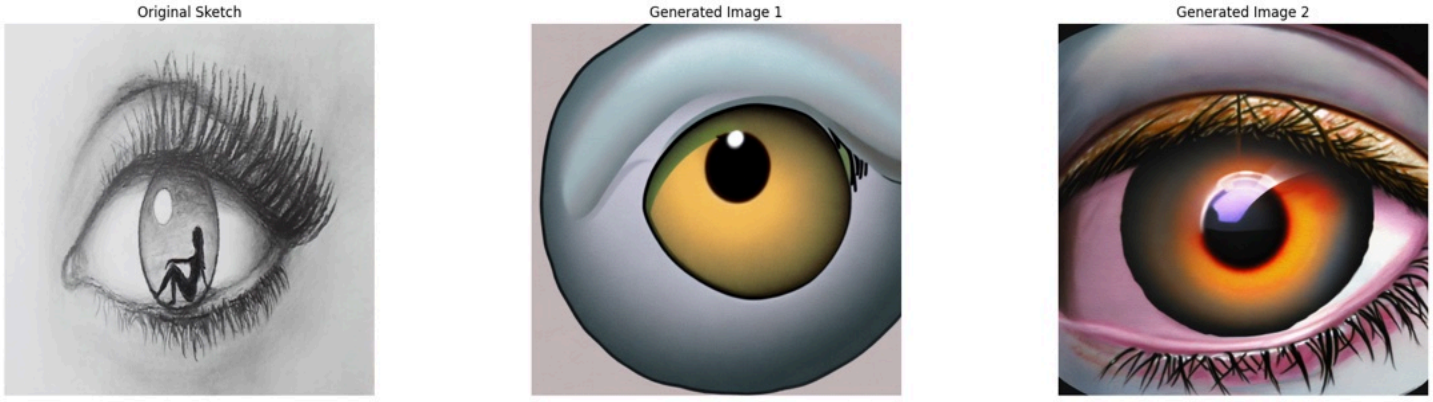
Starting from an image



Prompt: An elephant painting a picture of a giraffe, in a cute cartoon style, vibrant colors

Here the main components of the image are elephant, giraffe, and painting. The model has preserved these main components on both the images. Although the elephant isn't painting a picture of the giraffe, both generated images have a canvas painting style. The cartoon style and vibrant colors are also retained.

In the first generated image, the giraffe has the nose, ears, and trunk of an elephant while the elephant has the skin pattern of a giraffe. In the second generated image, the model has added two more giraffes with the trunk and nose of elephants with the main giraffe at the center having the nose of an elephant and a weird cartoonish bulging eye. In both images, the environment is whimsical, cartoonish and vibrant with interacting characters.



Prompt: An eye with a shadow of female sitting inside the eyeball, realistic, detailed, vibrant color

In this sketch input with a shadow of a female in sitting posture inside the eyeball, the model failed to generate the female shadow. However, the photorealism, details and vibrant colors are well retained especially in the second generated image.

Prompt: Bob Marley face, smiling, trippy eye, caricature cartoon style, black and white

Similarly in this image, the image of Bob Marley is preserved with cartoon style, smiling, trippy eye, and grayscale format. In the first image, the model has added details to the face, hair and smile. In the second generated image, in addition to the facial details, the model has added a trippy feel to the whole face instead of just the eyes.