Assignment

Problem Statement: *Dissecting image generation*

With the recent advent of image generation using Stable Diffusion [1], generating descriptive and photo-realistic images has become easy. This image generation can be guided and controlled by the user's text prompt.
Further, recent works like ControlNet [2-3] also support guidance from a wide variety of options like, depth maps, normal maps, etc. An example of such an image generation for some text prompt conditioned on an input depth is here.
For this assignment you are given the metadata. The metadata contains:

1. prompts: text description for the image to be generated, and
2. images: Inside this folder you will find the input depth maps for each of the corresponding prompts above. The folder contains 7 depth maps (5 in png format and 2 in npy format).

The objective of this assignment is to do **thorough analysis** on the image generation pipeline involving a critique on the various guidance and conditioning possible as per user's input. To achieve this, some tasks are listed below.
You can go through the references for additional guidance and feel free to write at harshil.bhatia@avataar.ai for any clarifications.

Tasks:

The assignment has three "main" deliverables:

1. For the given metadata, i.e., text description and depths, generate the "best" possible output images
    a. When needed feel free to combine depth with other forms of conditioning in ControlNet like normals, canny etc. (you can extract them using the given depths).
2. Can we generate images of different aspect ratios (use "Metadata/No crop/2_nocrop.png" to test this out) using SD? Comment on the generation quality with respect to the aspect ratio of 1:1 for the same image.
3. What is the generation latency? Do you see some quick fixes to reduce it? Comment on how much latency you can reduce. What happens to the generation quality with reduced latency?

Guidelines to be followed:

1. Use the following checkpoint: Stable Diffusion Checkpoint.
2. Keep the seed fixed to **12345** (for the SD model inference only).
3. For the generated outputs, verify that the generated depths are the same as the input depths.
4. The pipeline/heuristic should work consistently across all the images. Do not "manually" change any metadata.

Deliverables:

As part of the submission, provide a link of the github repository with the code and a readme file describing the execution of the code, the thought process and the visual results. Also analyse where your approach works well, where it fails, ideas on what would be done to fix it etc. and include this in your report.

References:

[1] Stable Diffusion
[2] ControlNet
[3] Huggingface space