# Project #2 for CEG5304: Generating Images through Prompting and Diffusion-based Models.

Spring (Semester 2), AY 2023-2024

In this exploratory project, you are to explore how to generate (realistic) images via diffusion-based models (such as DALLE and Stable Diffusion) through prompting, in particular hard prompting. To recall and recap the concepts of prompting, prompt engineering, LLVM (Large Language Vision Models), and LMM (Large Multi-modal Models), please refer to the slides on Week 5 ("Lect5-DL_prompt.pdf").

Before beginning this project, please read the following instructions carefully, failure to comply with the instructions may be penalized:

1. This project does not involve compulsory coding, complete your project with this given Word document file by filling in the "TO FILL" spaces. **Save the completed file as a PDF file** for submission. Please do NOT modify anything (including this instruction) in your submission file.
2. The marking of this project is based on how detailed the description and discussion are over the given questions. To score, please make sure your descriptions and discussions are readable, and adequate visualizations are provided.
3. The marking of this project is **NOT** based on any evaluation criteria (e.g., PSNR) over the generated image. Generating a good image does NOT guarantee a high score.
4. You may use ChatGPT/Claude or any online LLM services for polishing. However, purely using these services for question answering is prohibited (and is actually very obvious). If it is suspected that you generate your answers holistically with these online services, your assignment may be considered as committing plagiarism.
5. Submit your completed PDF on Canvas before the deadline: 1759 SGT on 20 April 2024 (updated from the slides). Please note that the deadlines are strict and late submission will be deducted 10 points (out of 100) for every 24 hours.
6. The report must be done **individually**. You may discuss with your peers, but NO plagiarism is allowed. The University, College, Department, and the teaching team take plagiarism very seriously. An originality report may be generated from iThenticate when necessary. A zero mark will be given to anyone found plagiarizing and a formal report will be handed to the Department/College for further investigation.

Task 1: generating an image with Stable Diffusion (via Huggingface Spaces) and compare it with the objective real image. (60%)

## Answer

In this task, you are to generate an image with the Stable Diffusion model in Huggingface Spaces. The link is provided here: CLICK ME. You can play with the different prompts and negative prompts (prompts that instructs the model NOT to generate something). Your objective is to generate an image that looks like the following image:

1a) First, select a rather coarse text prompt. A coarse text prompt may not include a lot of details but should be a good starting prompt to generate images towards our objective. An example could be “A Singaporean university campus with a courtyard.”. Display your generated image and its corresponding text prompt (as well as the negative prompt, if applicable) below: (10%)

Prompt :

Positive: A courtyard in a Singaporean university campus

Negative: Low quality

1b) Describe, in detail, how the generated image is compared to the objective image. You may include the discussion such as the components in the objective image that is missing from the generated image, or anything generated that does not make sense in the real world. (20%)

Compared the generated image to the objective image, we can find their similarities and differences:

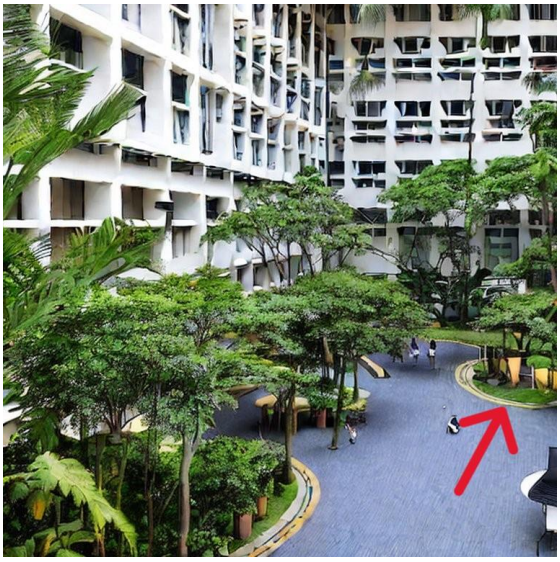
## Similarities:

1. **Green Environment**: Both images feature lush green plants, creating a pleasant atmosphere.
2. **Multi-Story Buildings**: The courtyards are surrounded by multi-story buildings, which could be academic buildings, dormitories, or other functional structures.
3. **Pathways or Walkways**: Both images show pathways leading to the buildings, providing access for walking, leisure, or transit.

## Differences:

1. **Architectural Style**: The first image showcases modern architecture, while the architectural style in the second image is unknown.
2. **Plant Density:** The plant in the first picture is more sparsely branched, while the plant in the second picture is more dense.

To accelerate the procedure of image generating, I use the negative prompt “Low quality”, which made the generated image coarse. As it can be seen, the window in the picture is distorted. Also the silhouette of the person in the picture is somewhat distorted. The image in the area pointed to with the red arrow in the image below also gives an indication of splicing.



## End

Next, you are to improve the generated image with prompt engineering. Note that it is highly likely that you may still be unable to obtain the objective image. A good reference material for prompt engineering can be found here: PROMPT ENGINEERING.

1c) Describe in detail how you improve your generated image. The description should include display of the generated images and their corresponding prompts, and detailed reasoning over the change in prompts. If the final improved image is generated with several iterations of prompt improvement, you should show each step in detail. I.e., you should display the result of each iteration of prompt change and discuss the result of each prompt change. You should also compare your improved image with both the first image you generated above, as well as the objective image. (30%)

## Answer

Prompt :

Positive: A courtyard in a Singaporean university campus, overlook

Negative: Low quality

First of all, we analyze our picture from the structure, based on the word “a courtyard in a Singaporean university campu” generated model already has the characteristics of the Courtyard, but the perspective and the target picture is much worse, so we add the word “overlook” to generate the picture as above. You can see that the perspective has changed, but the layout of the roads on the ground is not as good as in the first picture, and the layout of the buildings has changed from rectangular to circular.

Prompt :

Positive: A courtyard in a Singaporean university campus, hallway with crossroad, overlook

Negative: Low quality

Because we didn't give the structural details of the ground surface before, we added the word “hallway with crossroad” here, and generated the image shown above. We can see that the building structure of the courtyard is similar to that of the target, both of them are rectangular, and the generated image is closer to the target, but the figure in the image is still a bit distorted, which will have an effect on the diffusion.

Prompt :

Positive: A courtyard in a Singaporean university campus, hallway with crossroad, overlook

Negative: Low quality

In order to reduce the impact of people, so in the negative prompt to add the “people” entry, the final image generated as shown above, closer to the real NUS EA courtyard.

# End

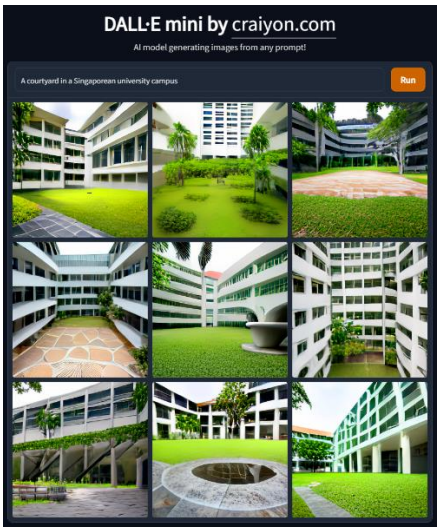
## Task 2: generating images with another diffusion-based model, DALL-E (mini-DALL-E, via Huggingface Spaces). (40%)

Stable Diffusion is not the only diffusion-based model that has the capability to generate good quality images. DALL-E is an alternative to Stable Diffusion. However, we are not to discuss the differences over these two models technically, but the differences over the generated images qualitatively (in a subjective manner). The link to generating with mini-DALL-E is provided here: MINI-DALL-E.

2a) You should first use the same prompt as you used in Task 1a and generate the image with mini-DALL-E. Display the generated image and compare, in detail, the new generated image with that generated by Stable Diffusion. (10%)

# Answer

Prompt : A courtyard in a Singaporean university campus, hallway with crossroad, overlook




I used the same prompt “a courtyard in a Singaporean university campus” as in 1 to generate the image as shown above, selecting the image closest to the target image, we get the following image:

Compared the new generated image with that generated by Stable Diffusion, these two images share some similarities and differences:

Similarities:

1. **Multi-Story Buildings**: The courtyards are surrounded by multi-story buildings, which could be academic buildings, dormitories, or other functional structures.
2. **Element Type**: The two images show courtyards with plants and roads, both surrounded by buildings.

Differences:

1. **Plant Density:** The plants in this image are not as dense as in the other one, and the plants in this image are only trees and grass, not bushes as in the other one.
2. **Element Layout**: Element layout: in this image the road is surrounded by grass, while in the other image the road and the greenbelt are integrated with each other.

# End

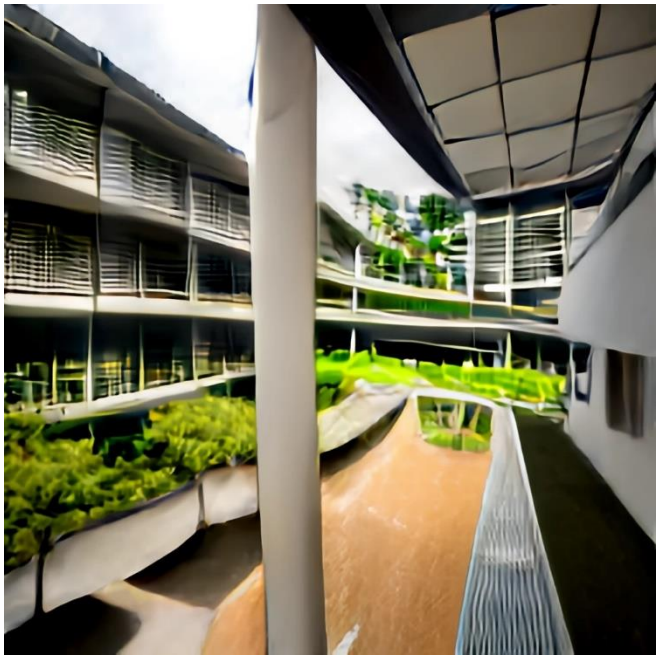
2b) Similar to what we performed for Stable Diffusion; you are to again improve the generated image with prompt engineering. Describe in detail how you improve your generated image. Similarly, if the final improved image is generated with several iterations of prompt improvement, you should show each step in detail. The description should include display of the generated images and their corresponding prompts, and detailed reasoning over the change in prompts. You should compare your improved image with both the first image you generated above, as well as the objective image.

In addition, you should also describe how the improvement is similar to or different from the previous improvement process with Stable Diffusion. (10%)

# Answer

Similar to the steps in Task 1, we first add a change in perspective.

Prompt : A courtyard in a Singaporean university campus, hallway with crossroad, overlook



As you can see from the image above, unlike Stable Diffusion, once I added the “overlook” prompt, the building construction and layout of the resulting image is very similar to the target image. It can be seen that the courtyard building construction learned from training in this model is similar to the target image. However, the same as using Stable Diffusion in this step is that the layout of plants and roads in the courtyard is not similar to the target.

Prompt : A courtyard in a Singaporean university campus, crossroad, overlook

Thus, I added a new prompt “crossroad” to generate the image shown above, and finally got the result as shown in Figure A. You can see that after adding “crossroad”, on the basis of the previous generated image, the road and the plants have changed from surrounding to blending with each other, which is closer to the target. Although the quality of the image was sacrificed

for the speed of generation, the layout and shape of the elements in the image are similar to the target.

## End

2c) From the generation process in Task 1 and Task 2, discuss the capabilities and limitations over image generation with off-the-shelf diffusion-based models and prompt engineering. You could further elaborate on possible alternatives or improvements that could generate images that are more realistic or similar to the objective image. (20%)

## Answer

Diffusion-based models and prompt engineering have certain capabilities in image generation, but also have some limitations:

### Capabilities:

1. **Diversity**: Diffusion-based models can generate a variety of images, from simple geometric shapes to complex natural landscapes.
2. **Control**: By adjusting model parameters and prompts, the content and style of the generated images can be controlled to some extent.

### Limitations:

1. **Realism**: Although diffusion-based models can generate a variety of images, the generated images may not achieve a level of realism similar to the real world. This is because the training data and algorithms of the model limit its understanding and modeling capabilities of the real world.
2. **Consistency**: In a continuous sequence of images, diffusion-based models may not maintain consistency, resulting in visually disjointed generated images.

### Possible alternative or improvement methods:

1. **Try other models**: For example, use models based on Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs), which have made significant progress in image generation.
2. **Improve training data**: Using larger and more diverse training data can help the model better understand and simulate the real world.
3. **Introduce additional constraints**: For example, some physical or geometric constraints can be introduced during the training process of the model to improve the realism and consistency of the generated images.

## End