Ayna Assignment Report

Name-Atharva Vaidya
Roll No-2022201046
Branch-M.Tech CSE

Experimentation:

Details->

1. Models used:-

-> Stability AI's Stable Diffusion xl 1 model.
-> Stable Diffusion model pretrained using concept of Lora.

2. Technology used:-

-> Pytorch
-> OpenCv

3. Libraries used:-

-> Hugging Face
-> Various other supporting libraries

Pipeline Process->

Phase-1. Input Prompt:-

-> The input prompt is to given to the string variable "prompt". In order to check for steerability, I was thinking of adding NLP procedures like basic preprocessing and removing stepwords, but when I researched online, I found that these diffusion models already take care of these in their input text processing and with low time decided not to go fot it.(Although would love to do it and see the result)
-> I tried and generated many vague inputs for seeing steerabilty, but model was able to understand the meaning and semantics of the sentence.

Phase-2. Calling Model:-

->A function which takes the given prompt as input and calls the model to generate the image according to the prompt given.

->This model can be first or second model according to the given model id.
->The model generates the image and saves it first in a created directory “Images”.
->After this the model displays the output on notebook for viewing the result.
->The images generated by base model were 1024*1024 images shown as 512*512(using cv window function).

Phase-3. Upscaling the image:-

->As the task mentioned to upscale the image, I upscaled the output base image to 2048*2048 image using a concept of bilinear interpolation as directly rescaling it distorts the pixel and reduces the image quality.
->P.S I tried to add ESRGAN model for upscaling the image, but was getting some error which I was not able to correct, and with given time constraints decided to go with bilinear interpolation which is also very good for the task of upscaling.

Metrics Focused->

->Photorealism:- Photorealism refers to the artistic movement or style in which paintings, drawings, or other artworks closely resemble high-resolution photographs.

->Steerability:- Steerability often relates to the ability to control or influence the output of a model in a specific way.

->Optimization(Computing and resources):- Optimization refers to the most efficient use of the compute power with judicious use of the resources to perform the task saving lot of energy and power.

Analysis of the models->

Model-1:

Description:

-> In the first model, I have loaded the weights of the Stability AI's stable diffusion xl 1 model. It is one of the State of the Art diffusion model for generating image from text prompt.
-> Loaded these weights from the hugging face library.
-> The pipline would be the same as described above with model Id of this model.

Experiments:

->Tried many other open-source image generation models, but were not at par with this model so decided not to go with those. Also, the session would fully exhaust GPU, so could not run all these models one after another.

Outputs:-
1->

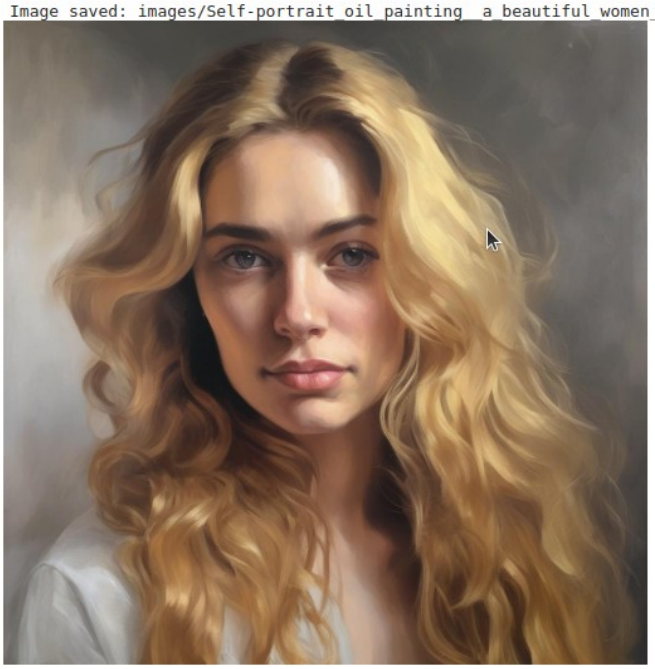
```
Running Stable diffusion to generate an image from text:
A young person, quite muscular wandering in the forest, wearing jacket in 4K
Please wait...
Loading pipeline components...: 100% 7/7 [00:01<00:00, 2.99it/s]
100% 50/50 [00:40<00:00, 1.15it/s]
Done in 48 seconds
Image saved: images/A young person  quite muscular wandering in the forest__wearing_jacket_in_4K_25Feb2024_082657_with_prompt.jpg
```

2->

```
Running Stable diffusion to generate an image from text:
A old person, quite muscular wandering in the forest, wearing jacket in 4K
Please wait...
Loading pipeline components...: 100% 7/7 [00:01<00:00, 3.04it/s]
100% 50/50 [00:43<00:00, 1.20it/s]
Done in 50 seconds
Image saved: images/A_old_person__quite_muscular_wandering_in_the_forest__wearing_jacket_in_4K_25Feb2024_092447_with_prompt.jpg
```

3->

```
Running Stable diffusion to generate an image from text:
Self-portrait oil painting, a beautiful women with golden hair, 8k
Please wait...
Loading pipeline components...: 100% 7/7 [00:01<00:00, 2.85it/s]
100% 50/50 [00:42<00:00, 1.14it/s]
Done in 49 seconds
Image saved: images/Self-portrait_oil_painting__a_beautiful_women_with_golden_hair__8k_25Feb2024_084749_with_prompt.jpg
```



4->

```
Running Stable diffusion to generate an image from text:
 Background of a war ongoing, a beautiful prince on horse with black hair, 8k
Please wait...
Loading pipeline components...: 100% 7/7 [00:01<00:00, 2.84it/s]
100% 50/50 [00:43<00:00, 1.20it/s]
Done in 50 seconds
Image saved: images/_Background_of_a_war_ongoing__a_beautiful_prince_on_horse_with_black_hair__8k_25Feb2024_090758_with_prompt.jpg
```

Model-2:

Description-

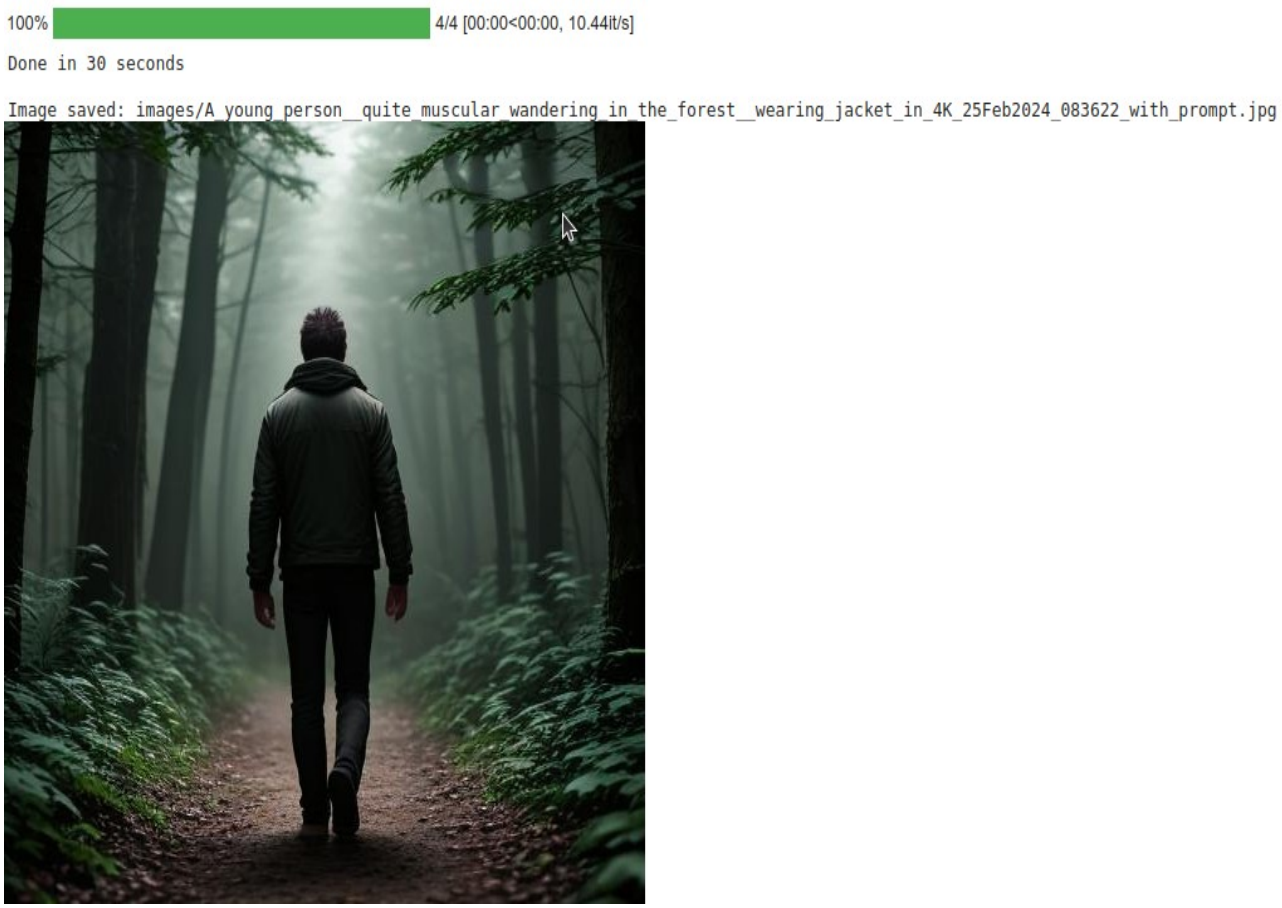
-> In the second model, I have loaded the lora pretrained model weights on the Stability AI's stable diffusion xl 1 model. It is one of the State of the Art diffusion model for generating image from text prompt.
-> Loaded these weights from the hugging face library.
-> The pipline would be the same as described above with model Id of this model.

Experiments:

->Tried many other open-source checkpoints, but were not at par with this model so decided not to go with those. Also, the session would fully exhaust GPU, so could not run all these models one after another.

Outputs:-

1->

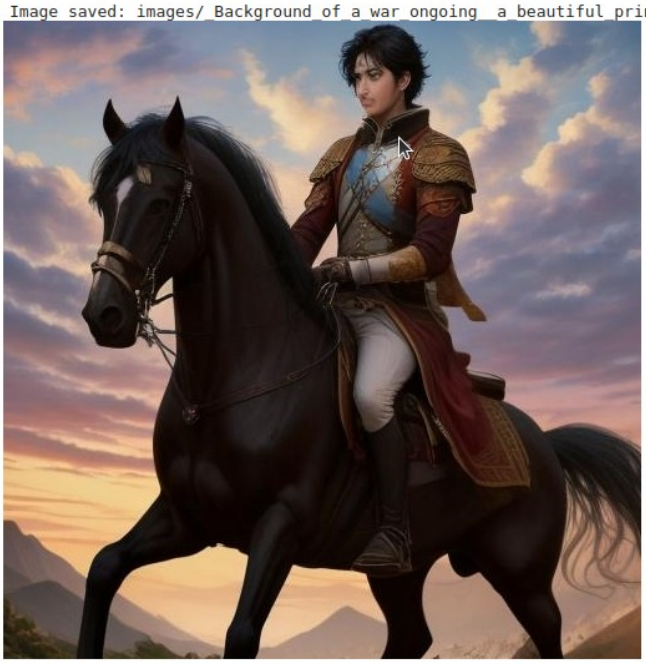


2->

```
Self-portrait oil painting, a beautiful women with golden hair, 8k

Please wait...
Loading pipeline components...: 100% ████████████████████ 7/7 [00:01<00:00, 4.26it/s]
The config attributes {'skip_prk_steps': True} were passed to LCMScheduler, but are not expected and will be ignored. Pl
on configuration file.
100% ████████████████████ 4/4 [00:00<00:00, 10.46it/s]
Done in 4 seconds

Image saved: images/Self-portrait_oil_painting__a_beautiful_women_with_golden_hair__8k_25Feb2024_084458_with_prompt.jpg
```

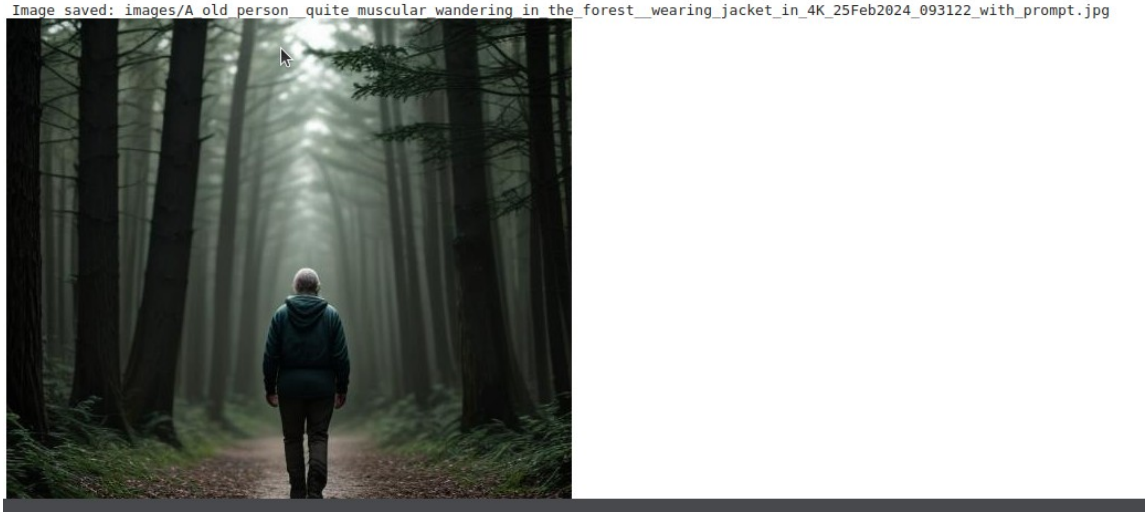
3->

```
Running Stable diffusion to generate an image from text:

 Background of a war ongoing, a beautiful prince on horse with black hair, 8k

Please wait...
Loading pipeline components...: 100% 7/7 [00:01<00:00, 3.91it/s]
The config attributes {'skip_prk_steps': True} were passed to LCMScheduler, but are not expected and will be ignored. Please verify
on configuration file.
100% 4/4 [00:00<00:00, 11.19it/s]
Done in 4 seconds

Image saved: images/ Background of a war ongoing  a beautiful prince on horse with black hair__8k_25Feb2024_090858_with_prompt.jpg
```



4->

```
Running Stable diffusion to generate an image from text:

A old person, quite muscular wandering in the forest, wearing jacket in 4K

Please wait...

Loading pipeline components...: 100% 7/7 [00:01<00:00, 3.74it/s]

/opt/conda/lib/python3.10/site-packages/transformers/models/clip/feature_extraction_clip.py:28: FutureWarning: The class CLIPFeatur
and will be removed in version 5 of Transformers. Please use CLIPImageProcessor instead.
  warnings.warn(
The config attributes {'skip_prk_steps': True} were passed to LCMScheduler, but are not expected and will be ignored. Please verify
on configuration file.
/opt/conda/lib/python3.10/site-packages/diffusers/loaders/lora.py:1078: FutureWarning: `fuse_text_encoder_lora` is deprecated and w
0.27. You are using an old version of LoRA backend. This will be deprecated in the next releases in favor of PEFT make sure to inst
ransformers packages in the future.
  deprecate("fuse_text_encoder_lora", "0.27", LORA_DEPRECATION_MESSAGE)

100% 4/4 [00:00<00:00, 10.81it/s]

Done in 4 seconds

Image saved: images/A old person__quite muscular wandering in the forest__wearing jacket in 4K_25Feb2024_093122_with_prompt.jpg
```



Comparisions:-

->**Photorealism:-** Model 1 and Model-2 compete in this aspect because it can be clearly seen that the images produced by both the images are realistic, although model-2 images feel more like dream and are crystal clear.

->**Steerability:-** Both the models are quick to adapt to the smallest of the changes in the prompt as shown below:-

Model-1: Here we observe that tweaking small details like young to old though gramatically incorrect, gave expected image.

Running Stable diffusion to generate an image from text:

A young person, quite muscular wandering in the forest, wearing jacket in 4K

```
Please wait...
Loading pipeline components...: 100% 7/7 [00:01<00:00, 2.99it/s]
100% 50/50 [00:40<00:00, 1.15it/s]
Done in 48 seconds
Image saved: images/A young person  quite muscular wandering in the forest__wearing jacket in 4K_25Feb2024_082657_with_prompt.jpg
```

Running Stable diffusion to generate an image from text:

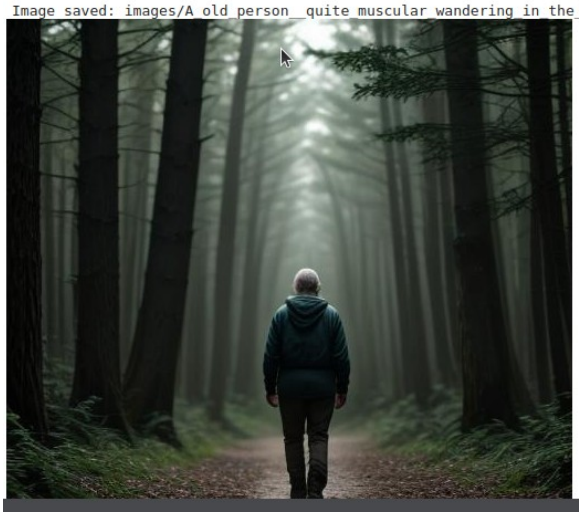
A old person, quite muscular wandering in the forest, wearing jacket in 4K

```
Please wait...
Loading pipeline components...: 100% 7/7 [00:01<00:00, 3.04it/s]
100% 50/50 [00:43<00:00, 1.20it/s]
Done in 50 seconds
Image saved: images/A old person  quite muscular wandering in the forest__wearing jacket in 4K_25Feb2024_092447_with_prompt.jpg
```

Model 2: We can clearly see that talthough the person’s face is not revealed, the hair is white and body is slightly fragile which can make us point out that the person is old.

```
100% 4/4 [00:00<00:00, 10.44it/s]
Done in 30 seconds
Image saved: images/A young person__quite muscular wandering in the forest__wearing jacket in 4K_25Feb2024_083622_with_prompt.jpg
```

```
Running Stable diffusion to generate an image from text:
A old person, quite muscular wandering in the forest, wearing jacket in 4K
Please wait...
Loading pipeline components...: 100% 7/7 [00:01<00:00, 3.74it/s]
/opt/conda/lib/python3.10/site-packages/transformers/models/clip/feature_extraction_clip.py:28: FutureWarning: The class CLIPFeatur
and will be removed in version 5 of Transformers. Please use CLIPImageProcessor instead.
  warnings.warn(
The config attributes {'skip_prk_steps': True} were passed to LCMScheduler, but are not expected and will be ignored. Please verify
on configuration file.
/opt/conda/lib/python3.10/site-packages/diffusers/loaders/lora.py:1078: FutureWarning: `fuse_text_encoder_lora` is deprecated and w
0.27. You are using an old version of LoRA backend. This will be deprecated in the next releases in favor of PEFT make sure to inst
ransformers packages in the future.
  deprecate("fuse_text_encoder_lora", "0.27", LORA_DEPRECATION_MESSAGE)
100% 4/4 [00:00<00:00, 10.81it/s]
Done in 4 seconds
Image saved: images/A old person__quite muscular wandering in the forest__wearing jacket in 4K_25Feb2024_093122_with_prompt.jpg
```



->**Optimization:-** Model 2 clearly outperforms here because of the Lora or low rank adaption we used and it is evident in the image that the result using model 1 was generated in 50-60 s whereas using Lora in model 2 was generated in 4-5 s, a speed gain of nearly 10 times.
->Also the VRAM requirement for Lora while processing was in MB's whereas for the first model was reaching nearly limit of 15GB.

->**Inferences:-** According to me, Model-2 clearly outperforms model-1 when judged considering all the metrics listed above and is evident from the quality, resources and time taken to generate these images by both the models.