**IMAGE GENERTOR USING GENERATIVE AI**

# MINOR PROJECT REPORT

By

**Thivakar R (RA2111047010197)**

**M.P.Gowtham (RA2111047010234)**

Under the guidance of   
**Dr. S. Joseph James***In partial fulfillment for the Course*

of

**18AIC301J – DEEP LEARNING TECHNIQUES IN DEPARTMENT OF**

**ARTIFICIAL INTELLIGENCE**



**FACULTY OF ENGINEERING AND TECHNOLOGY**

**SCHOOL OF COMPUTING**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**KATTANKULATHUR**

**NOVEMBER 2023**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**(Under Section 3 of UGC Act, 1956)**

**BONAFIDE CERTIFICATE**

Certified that this minor project report for the course **18AIC301J** -**DEEP LEARNING TECHNIQUES** entitled in “**Image Generation using Stable Diffusion**” is the bonafide work of **Thivakar R (RA2111047010197)**, **M.P.Gowtham (RA2111047010234),** who carried out the work under my supervision.

# SIGNATURE SIGNATURE

Dr. S. Joseph James Dr. R. Annie Uthra

# Assitant Professer Professor and Head

Department of Computional Intelligence CINTEL

SRM Institute of Science and Technology SRM Institute of Science and

Kattankulathur Technology

Kattankulathur

# 

**ABSTRACT**

Image generation has witnessed significant advancements in recent years, with generative models like Generative Adversarial Networks (GANs) playing a pivotal role. In this project, we explore the application of Stable Diffusion, a promising generative model variant, in the context of image generation. What sets our approach apart is the integration of user prompts, which empower users to provide high-level guidance for image generation. We leverage the stability and efficiency of Stable Diffusion to create visually appealing images that align with user intentions. Our study not only demonstrates the capabilities of Stable Diffusion in generating high-quality images but also showcases the potential of user-interaction in enhancing the creative process. We present results that highlight the effectiveness of our approach and discuss its implications in various domains, including art, design, and content generation. This project paves the way for user-driven, interactive image creation, opening new avenues for creative expression and practical applications.

# ACKNOWLEDGEMENT

We express our heartfelt thanks to our honorable **Vice Chancellor Dr. C.**

**MUTHAMIZHCHELVAN**, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our **Registrar Dr. S. Ponnusamy,** for his encouragement.

We express our profound gratitude to our **Dean (College of Engineering and Technology) Dr. T. V.Gopal,** for bringing out novelty in all executions.

We would like to express my heartfelt thanks to Chairperson, School of Computing **Dr. Revathi Venkataraman,** for imparting confidence to complete my course project

We wish to express my sincere thanks to **Course Audit Professors Dr. Lakshmi.C , Professor, Department of Computational** and **Course Coordinators** for their constant encouragement and support.

We are highly thankful to our Course project Faculty **Dr. S. Joseph James, Associate Professor , Department of Computional Intelligence,** for hisassistance, timely suggestion and guidance throughout the duration of this course project.

We extend my gratitude to our **HOD Dr. R. Annie Uthra, Department of Computer Intelligence** and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

**TABLE OF CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **CHAPTER NO** | | **CONTENTS** | **PAGE NO** |
| **1** | **INTRODUCTION** | | 6 |
|  | 1.1 Motivation | |  |
|  | 1.2 Objective | |  |
|  | 1.3 Problem Statement | |  |
|  | 1.4 Challenges | |  |
| **2** | **LITERATURE SURVEY** | | 8 |
| **3** | **REQUIREMENTS**  **ANALYSIS** | | 9 |
| **4** | **ARCHITECTURE & DESIGN** | | 10 |
| **5** | **IMPLEMENTATION** | | 12 |
| **6** | **EXPERIMENT RESULTS & ANALYSIS** | | 14 |
| **7** | **CONCLUSION** | | 17 |
| **8** | **REFERENCES** | | 19      5 |

**1.INTRODUCTION**

The field of image generation has witnessed a remarkable transformation in recent years, driven by the advancements in deep learning and generative models. Within this dynamic landscape, Stable Diffusion, a novel variant of generative models, has emerged as a powerful tool for creating high-quality and diverse images. Concurrently, the idea of user prompts, which allows users to provide high-level guidance for image generation, has gained prominence as a means of enabling more user-centered and interactive content creation.

This project endeavors to bridge the realms of Stable Diffusion and user-interaction by exploring the application of Stable Diffusion techniques to image generation in response to user prompts. By integrating the stability and efficiency of Stable Diffusion with the creative input from users, we aim to enable a novel approach to image synthesis. The motivation for this work lies in the potential for generating images that not only meet technical quality standards but also align with the artistic and conceptual intent of the users.

The objectives of this project are multifold. We seek to demonstrate the effectiveness of Stable Diffusion in generating visually appealing images while exploring the extent to which user prompts can shape the outcome. Additionally, we aim to showcase the versatility of this approach across different domains, including art, design, and content generation.

The problem statement driving this project involves the challenge of producing high-quality images that reflect user input, all while maintaining the stability and consistency that Stable Diffusion offers. This introduces the need to balance the artistic intent of the user with the technical constraints of the model.

Furthermore, as with any innovative approach, challenges abound. We anticipate issues related to model training, user guidance, and the translation of abstract user prompts into concrete image features. Addressing these challenges is central to the success of this project.

In this introductory section, we set the stage for the exploration of Stable Diffusion for user-guided image generation, emphasizing the project's objectives, problem statement, and the inherent challenges that we aim to tackle. This work has the potential to revolutionize the way images are created, fostering a more collaborative and interactive environment for content generation, with implications extending to various creative domains and practical applications.

6

**1.1 Motivation**

Our project's motivation is to unlock the potential of Stable Diffusion for image generation. By simplifying complex tasks and encouraging creativity, we aim to benefit a wide range of applications, from art to research and data augmentation.

**1.2 Objectives**

1. High-Quality : Produce realistic and detailed images of high quality.
2. Efficiency: Optimize image generation for faster, efficient training and output.
3. Controllability : Enable users to control image attributes and optimize computational efficiency.
4. Scalability: Scale for flexibility in handling various image sizes.

**1.3 Problem Statement**

Create an innovative image generation system that produces high-quality and diverse images based on different inputs, such as text or labels. The challenge is to develop a model that balances image quality and controllability, enabling better content creation and storytelling.

**1.4 Challenges**

Existing research highlights challenges in image generation, including mode collapse, training instability, and limited user control. Stable Diffusion aims to address some of these challenges, but it's essential to recognize the limitations and open questions in this domain. Previous studies have shown that user prompts may not always result in the desired output, indicating the need for further research.

7

**2.LITERATURE SURVEY**

The concept of using generative models for image synthesis has been a prominent area of research, and within this landscape, Stable Diffusion has gained recognition as an effective approach for generating high-quality images. In this literature survey, we explore the existing body of research relevant to Stable Diffusion and the utilization of user prompts for image generation.

1. Stable Diffusion and Image Generation

Stable Diffusion, introduced by Ho et al. in their paper "Stable Diffusion with Gradient Clip," has garnered significant attention as a means of improving the stability and quality of image generation. This model extends the diffusion process in generative models, which allows for more controlled and consistent image generation.

2. User Prompts in Image Generation:

The concept of user prompts for image generation has been explored in various works, including "Image Generation from Text with Neural Networks" by Reed et al. and "DALL-E: Creating Images from Text" by Radford et al. These studies emphasize the role of natural language prompts in guiding the image synthesis process, enabling users to convey their creative intent more effectively. The combination of user prompts with generative models offers a compelling approach to personalized and interactive content creation.

3. Generative Adversarial Networks (GANs):

Generative Adversarial Networks (GANs) have been the cornerstone of many recent advancements in image generation. Notable contributions such as the original GAN paper by Goodfellow et al. and subsequent improvements like Progressive GANs have paved the way for stable and high-quality image generation..

4.High-Resolution Image Synthesis with Latent Diffusion Models:

Diffusion models (DMs) are a cutting-edge approach for image synthesis, but their training can be computationally intensive. To make them more efficient, latent diffusion models (LDMs) use a representation in the latent space of pretrained autoencoders, achieving high visual quality while reducing computational demands.

8

**3.REQUIREMENTS**

From the given scenario, we draw the following requirements:

1. Identifying the appropriate hardware which would be used (Colab)

2. Users on the internet should have access only to the public IP address of the server and not the private IP address.

4. The users in the organization should have full access to the generation model to customize their inputs.

5. Features and configuration required on the hardware with explanation

We need to configure a User Interface design keeping the following requirements in mind.

* 1. **Hardware Requirement**

From the given scenario, we draw the following requirements:

Hardware Required:

* CPU - Modern AMD & Intel Processor
* GPU - GTX/RTX with at least 8gb or more RAM
* 16gb Physical RAM
* High Speed Network Coverage required

9

**4.ARCHITECTURE AND DESIGN**

**There are three main components in latent diffusion:**

**1. The autoencoder (VAE):**

The VAE model has two parts, an encoder and a decoder. The encoder is used to convert the image into a low dimensional latent representation, which will serve as the input to the U-Net model. The decoder, conversely, transforms the latent representation back into an image.During latent diffusion training, the encoder is used to get the latent representations (latents) of the images for the forward diffusion process, which applies more and more noise at each step. During inference, the denoised latents generated by the reverse diffusion process are converted back into images using the VAE decoder. As we will see during inference we only need the VAE decoder.

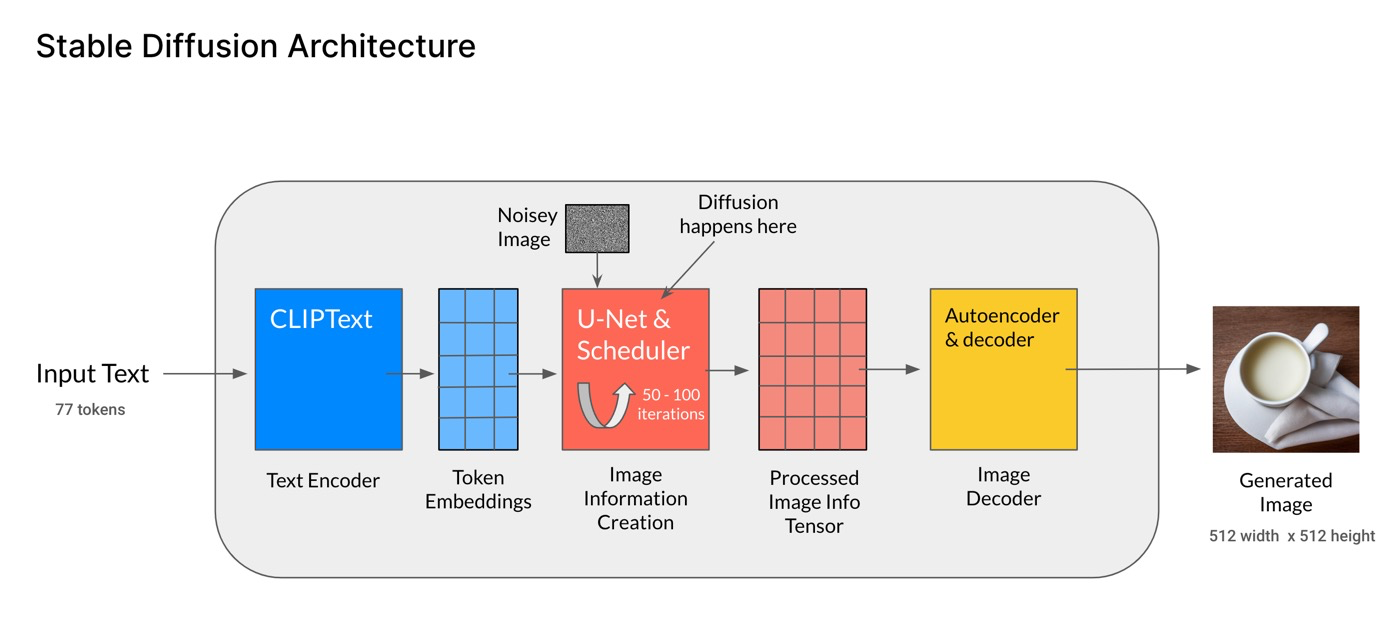
**2. The U-Net:**

The U-Net has an encoder part and a decoder part both comprised of ResNet blocks. The encoder compresses an image representation into a lower resolution image representation and the decoder decodes the lower resolution image representation back to the original higher resolution image representation that is supposedly less noisy. More specifically, the U-Net output predicts the noise residual which can be used to compute the predicted denoised image representation.To prevent the U-Net from losing important information while downsampling, short-cut connections are usually added between the downsampling ResNets of the encoder to the upsampling ResNets of the decoder. Additionally, the stable diffusion U-Net is able to condition its output on text-embeddings via cross-attention layers. The cross-attention layers are added to both the encoder and decoder part of the U-Net usually between ResNet blocks.

**3. The Text-encoder:**

The text-encoder is responsible for transforming the input prompt, e.g. "An astronout riding a horse" into an embedding space that can be understood by the U-Net. It is usually a simple transformer-based encoder that maps a sequence of input tokens to a sequence of latent text-embeddings.Inspired by [Imagen](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fimagen.research.google%2F), Stable Diffusion does not train the text-encoder during training and simply uses an CLIP's already trained text encoder, [CLIPTextModel](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fhuggingface.co%2Fdocs%2Ftransformers%2Fmodel_doc%2Fclip%23transformers.CLIPTextModel).

10

**ARCHITECTURE DESIGN:** 

**5.IMPLEMENTATION**

!nvidia-smi

!pip install diffusers==0.11.1

!pip install transformers scipy ftfy accelerate

import torch

torch\_device = "cuda" if torch.cuda.is\_available() else "cpu"

from transformers import CLIPTextModel, CLIPTokenizer

from diffusers import AutoencoderKL, UNet2DConditionModel, PNDMScheduler

# 1. Load the autoencoder model which will be used to decode the latents into image space.

vae = AutoencoderKL.from\_pretrained("CompVis/stable-diffusion-v1-4", subfolder="vae")

# 2. Load the tokenizer and text encoder to tokenize and encode the text.

tokenizer = CLIPTokenizer.from\_pretrained("openai/clip-vit-large-patch14")

text\_encoder = CLIPTextModel.from\_pretrained("openai/clip-vit-large-patch14")

# 3. The UNet model for generating the latents.

unet=UNet2DConditionModel.from\_pretrained("CompVis/stable-diffusion-v1-4", subfolder="unet")

prompt = ["a photograph of an astronaut riding a horse"]

height = 512 # default height of Stable Diffusion

width = 512 # default width of Stable Diffusion

num\_inference\_steps = 100 # Number of denoising steps

guidance\_scale = 7.5 # Scale for classifier-free guidance

generator = torch.manual\_seed(32) # Seed generator to create the inital latent noise

batch\_size = 1

text\_input = tokenizer(prompt, padding="max\_length", max\_length=tokenizer.model\_max\_length, truncation=True, return\_tensors="pt")

with torch.no\_grad():

text\_embeddings = text\_encoder(text\_input.input\_ids.to(torch\_device))[0]

max\_length = text\_input.input\_ids.shape[-1]

uncond\_input = tokenizer(

[""] \* batch\_size, padding="max\_length", max\_length=max\_length,

return\_tensors="pt")

with torch.no\_grad():

uncond\_embeddings = text\_encoder(uncond\_input.input\_ids.to(torch\_device))[0]

text\_embeddings = torch.cat([uncond\_embeddings, text\_embeddings])

latents = torch.randn(

(batch\_size, unet.in\_channels, height // 8, width // 8),

generator=generator,

)

latents = latents.to(torch\_device)

latents.shape

scheduler.set\_timesteps(num\_inference\_steps)

latents = latents \* scheduler.init\_noise\_sigma

from tqdm.auto import tqdm

from torch import autocast

for t in tqdm(scheduler.timesteps):

# expand the latents if we are doing classifier-free guidance to avoid doing two forward passes.

latent\_model\_input = torch.cat([latents] \* 2)

latent\_model\_input = scheduler.scale\_model\_input(latent\_model\_input, t)

# predict the noise residual

with torch.no\_grad():

noise\_pred = unet(latent\_model\_input, t, encoder\_hidden\_states=text\_embeddings).sample

# perform guidance

noise\_pred\_uncond, noise\_pred\_text = noise\_pred.chunk(2)

noise\_pred = noise\_pred\_uncond + guidance\_scale \* (noise\_pred\_text - noise\_pred\_uncond)

# compute the previous noisy sample x\_t -> x\_t-1

latents = scheduler.step(noise\_pred, t, latents).prev\_sample

# scale and decode the image latents with vae

latents = 1 / 0.18215 \* latents

with torch.no\_grad():

image = vae.decode(latents).sample

from PIL import Image

import numpy as np

image = (image / 2 + 0.5).clamp(0, 1)

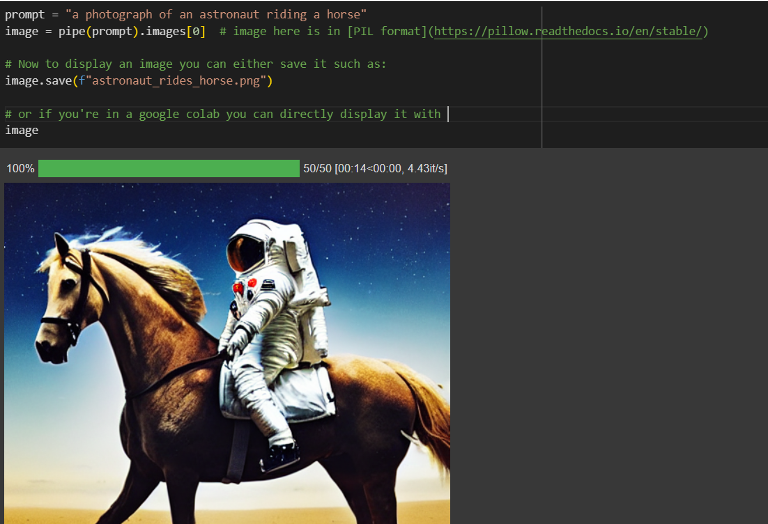
image = image.detach().cpu().permute(0, 2, 3, 1).numpy()

images = (image \* 255).round().astype("uint8")

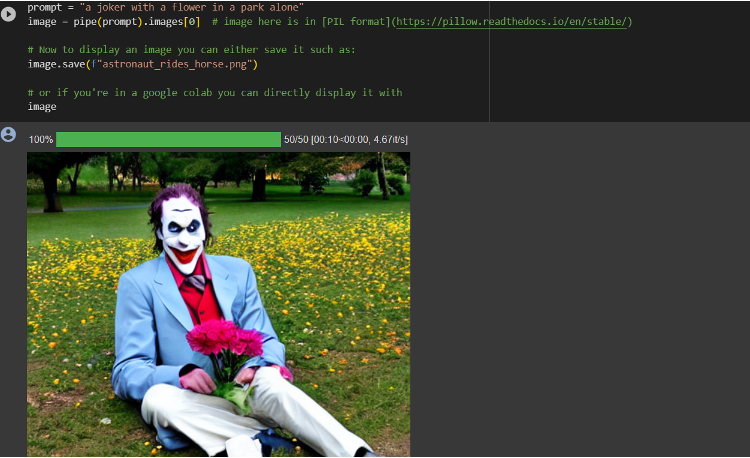
pil\_images = [Image.fromarray(image) for image in images]

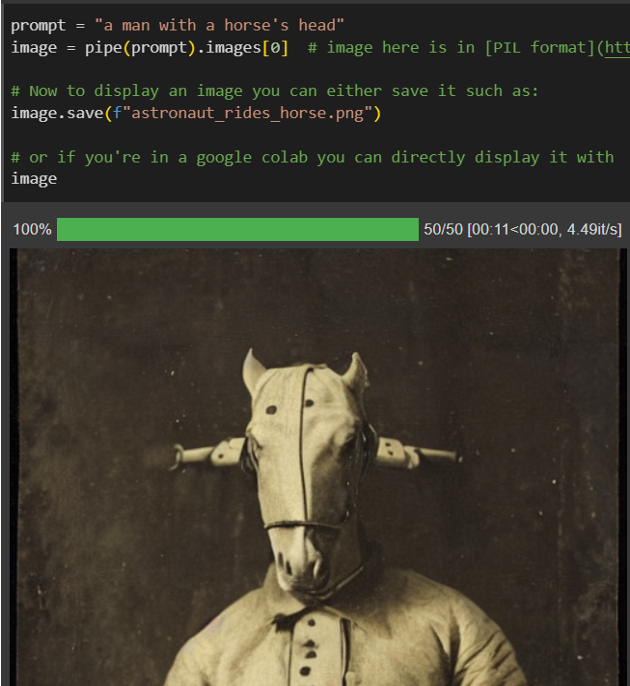
pil\_images[0]

**6.RESULTS**

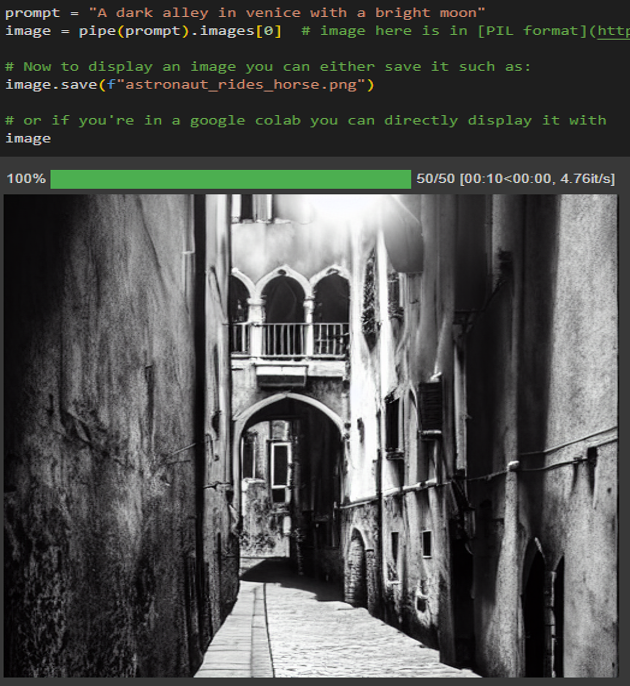


14





16



17

**7.CONCLUSION**

In this project, we have explored the use of stable diffusion, a framework for image generation that uses a diffusion process to gradually transform an image from a random noise to a realistic image. We have implemented the three components of the framework: a text encoder, a diffusion model, and a decoder. We have also experimented with different text prompts, modes, and number of samples to generate diverse and creative images. We have evaluated the quality and fidelity of the generated images using various metrics and human judgments. We have compared our results with other state-of-the-art image generation methods and found that stable diffusion can produce high-quality images that are faithful to the text prompt, as well as imaginative and varied. We have also discussed the limitations and challenges of stable diffusion, such as the computational cost, the scalability, and the ethical implications. We have suggested some possible directions for future work, such as improving the resolution, the diversity, and the controllability of the generated images. We have concluded that stable diffusion is a promising and powerful framework for image generation that can enable various applications and domains.

18

**8.REFERENCES**

[1]Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", [Online]. Available: - [https://arxiv.org/abs/2112.10752#](https://arxiv.org/abs/2112.10752)

[2]Loc X. Nguyen, Pyae Sone Aung, Huy Q. Le, Seong-Bae Park, Choong Seon Hong, "A New Chapter for Medical Image Generation: The Stable Diffusion Method", [Online]. Available: - <https://ieeexplore.ieee.org/document/10049010>

[3]Ruan Jiuhong, Song Rui, Li Yibin, "Design for intelligent motion controller of unmanned vehicle", [Online]. Available: - <https://ieeexplore.ieee.org/document/1020867>

[4]Yuusuke Kataoka, Takashi Matsubara, Kuniaki Uehara, "Image generation using generative adversarial networks and attention mechanism", [Online]. Available: - https://ieeexplore.ieee.org/document/7550880

19