**Image Generation using stable diffusion & Comfy UI**

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning

with

TechSaksham – A joint CSR initiative of Microsoft & SAP

by

**Manav Gyaneshwar Patil,**

**manavpatil3010@gmail.com**

Under the Guidance of

**Jay Rathore**

**ACKNOWLEDGEMENT**

I would like to take a moment to express my sincere and heartfelt gratitude to everyone who has supported me throughout this internship project. Without their invaluable help and encouragement, this journey wouldn’t have been as fulfilling or successful.

First and foremost, I want to extend my deepest thanks to my supervisors, **Jay Rathod**, **Pavan Sumohana**, and **Adharsh P**. Their unwavering guidance, patience, and dedication have truly been a driving force behind this project. From the very beginning, they have provided me with not only expert advice but also constant motivation to push my limits. Their belief in my potential has been a constant source of confidence, and I cannot thank them enough for their continuous support.

I feel incredibly fortunate to have had the opportunity to work with such brilliant minds. Their feedback was always constructive, encouraging, and insightful, which helped me grow both professionally and personally. The lessons they shared extend beyond this project and will remain with me throughout my career. It has been an absolute privilege to learn from them, and I am truly grateful for their mentorship.

This project wouldn’t have been possible without their dedication to my success, and I will forever be thankful for everything they have taught me.

#### **ABSTRACT**

This project focuses on the development and implementation of an AI-based image generation system using Stable Diffusion, a powerful deep learning model for text-to-image synthesis. The objective of this project is to explore the capabilities of Stable Diffusion in creating high-quality, realistic images from textual descriptions, and to leverage ComfyUI as a user-friendly interface for model interaction. The problem at hand is the need for efficient and accessible tools that allow users to generate customized images for various applications such as digital art, marketing, and content creation.

The methodology employed in this project involves training and fine-tuning a pre-trained Stable Diffusion model to generate images based on user-provided text prompts. The ComfyUI interface was integrated to facilitate seamless interaction with the model, providing users with an intuitive platform to input prompts and view generated images. Various techniques, including image refinement and prompt engineering, were employed to optimize the output quality and relevance.

Key results of the project include the successful generation of high-quality images that align with the provided prompts. The integration of ComfyUI enabled an interactive and customizable experience for users, making it easy to experiment with different input parameters. The model demonstrated flexibility in creating diverse image styles, from abstract concepts to photorealistic visuals.

In conclusion, this project successfully demonstrated the potential of Stable Diffusion for text-to-image generation, highlighting its value in creative and professional applications. The development of ComfyUI provided an accessible way for users to interact with the model, showcasing the practical benefits of AI-driven image generation in various industries.

**TABLE OF CONTENT**

**Abstract I**

**Chapter 1.**  **Introduction 1**

1.1 Problem Statement 1

1.2 Motivation 1

1.3 Objectives 2

1.4. Scope of the Project 2

**Chapter 2.**  **Literature Survey 3-4**

**Chapter 3.**  **Proposed Methodology 5-6**

3.1 System Design: 5-6

3.2 Requirements 6

**Chapter 4.**  **Implementation and Results 7-15**

4.1 Snap Shots of Result: 7-15

4.2 GitHub Link for Code 15

**Chapter 5. Discussion and Conclusion 16**

5.1 Conclusion 16

5.2 Future Work 16

**References**  **17**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Caption** | **Page No.** |
|  | Mystical Forest | **7** |
|  | Cyberpunk City | **8** |
|  | Majestic Dragon | **9** |
|  | Steampunk Airship | **10** |
|  | New City | **11** |
|  | Japanese Garden | **12** |
|  | Horror Asylum: | **13** |
|  | Chinese Buddhist Goddess | **14** |
|  | Post-Apocalyptic World | **15** |

**CHAPTER 1**

**Introduction**

* 1. **Problem Statement:**

In this project, I address the challenge of generating high-quality images from text descriptions. As demand for visual content grows, many individuals and businesses face the difficulty of creating customized images without specialized skills or expensive software. Traditional image creation methods are often time-consuming and require significant expertise, limiting accessibility for many.

This problem is significant because visual content is essential in today’s digital world, from marketing to social media. An AI-driven solution like Stable Diffusion, combined with a user-friendly interface like ComfyUI, can simplify the process and make high-quality image generation accessible to everyone. My goal is to provide a tool that allows users to easily create images from text, removing barriers to creativity and content production.

* 1. **Motivation:**

I chose this project to explore how AI-driven image generation can simplify the creative process and make it accessible to a wider audience. With the rise of AI technologies, I see a significant opportunity to empower individuals and businesses to generate high-quality images from text prompts without requiring advanced technical skills.

The motivation behind this project is to address the growing demand for digital content across industries like art, marketing, and gaming. By using Stable Diffusion, I aim to create a tool that reduces the time and cost involved in image production, while also enabling more people to bring their creative ideas to life.

Furthermore, by integrating ComfyUI, I want to ensure that the technology is easy to use, even for those without a technical background. The potential applications are vast, from personal artwork creation to generating customized marketing visuals, and I believe this project has the power to make a real impact in various fields.

* 1. **Objective:**

The primary objective of this project is to develop an AI-powered image generation system using Stable Diffusion, with the goal of creating high-quality, contextually accurate images from textual descriptions. By leveraging the power of Stable Diffusion, I aim to explore the capabilities of text-to-image synthesis and provide users with a tool that can generate images based on simple text prompts.

Additionally, I plan to integrate ComfyUI, a user-friendly interface, to ensure that the technology is accessible to both technical and non-technical users. This platform will allow individuals to easily experiment with different input prompts and generate customized images without needing advanced knowledge of AI or coding.

The broader goal is to demonstrate how AI can be harnessed for creative applications, enabling anyone—from digital artists to marketers—to create high-quality images quickly and affordably. By achieving this, I aim to make AI-driven image generation more accessible, fostering creativity and streamlining content production.

* 1. **Scope of the Project:**

In this project, I focus on utilizing Stable Diffusion to generate high-quality images based on text prompts. My goal is to develop an AI-powered platform where users can input text descriptions and receive corresponding images. To make the system user-friendly, I will integrate ComfyUI, ensuring that it is accessible to both technical and non-technical users.

The project scope includes the creation of a proof-of-concept that demonstrates the potential of AI in text-to-image generation. I will explore various techniques like prompt engineering and image refinement to enhance the quality and relevance of the generated images.

**Limitations**

* This project will rely on the pre-trained Stable Diffusion model, and I will not train a new model from scratch.
* The focus will be on generating static images, so I won’t be working on dynamic image generation, such as videos or animations.
* The performance of the system may vary based on the complexity of the input prompts and the available computational resources, which might impact the speed and quality of the output for highly detailed requests.
* Advanced image manipulation features, beyond basic text-to-image generation, will not be part of this project.

**CHAPTER 2**

**Literature Survey**

* 1. **Review relevant literature or previous work in this domain.**

The field of AI-driven image generation has evolved significantly over the past decade, with the introduction of various deep learning models aimed at synthesizing high-quality images from text. One of the most notable advancements is the development of Generative Adversarial Networks (GANs), introduced by Ian Goodfellow and his colleagues in 2014. GANs demonstrated the potential to create realistic images, but they were often limited by their training instability and difficulty in generating diverse, high-quality results.

In recent years, the introduction of diffusion models, including Stable Diffusion, has revolutionized the text-to-image generation process. Unlike GANs, diffusion models work by gradually transforming a random noise pattern into a coherent image through a series of denoising steps. This process has proven to generate more stable and diverse outputs, making it a significant leap forward in generative AI.

Stable Diffusion, in particular, has gained widespread attention due to its ability to produce high-quality images based on textual descriptions. This model has been trained on vast datasets and fine-tuned to generate realistic images that closely match the input prompts. The model's ability to create complex and coherent images from text prompts, combined with its open-source nature, makes it an attractive tool for developers and researchers in the field.

The user interface plays a crucial role in enabling accessibility to these powerful models. ComfyUI is a notable tool in this regard, offering an intuitive interface for interacting with Stable Diffusion. By simplifying the process of inputting text prompts and customizing image generation parameters, ComfyUI makes it easier for users to explore the full potential of Stable Diffusion without needing deep technical knowledge.

* 1. **Mention any existing models, techniques, or methodologies related to the problem**

Several AI models have been used for image generation, with GANs and Variational Autoencoders (VAEs) being among the most well-known. GANs, while groundbreaking in their time, are often criticized for their instability during training and their inability to generate truly diverse images. VAEs, on the other hand, are more stable but often produce less sharp or realistic images compared to GANs.

The introduction of **Denoising Diffusion Probabilistic Models (DDPMs)** and later **Stable Diffusion** marked a breakthrough in generative AI. Stable Diffusion uses a technique called latent diffusion to efficiently generate high-quality images. It works by operating on compressed representations of images, which makes it computationally efficient without compromising the quality of the generated outputs.

ComfyUI serves as a critical component in enhancing user experience. It provides an accessible interface for interacting with Stable Diffusion, allowing users to experiment with various parameters like image resolution, style, and level of detail, all through a straightforward graphical interface. This integration bridges the gap between powerful generative models and end users, democratizing access to AI-driven image generation.

* 1. **Highlight the gaps or limitations in existing solutions and how your project will address them.**

While models like GANs and VAEs laid the foundation for AI-based image generation, they come with certain limitations, such as mode collapse (where the model generates a limited set of outputs) and difficulty in producing fine details. Diffusion models, particularly Stable Diffusion, have addressed many of these issues, offering higher quality and more diverse image generation. However, they still have certain limitations, such as the need for significant computational resources to generate images at higher resolutions and the potential for slower processing times compared to traditional methods.

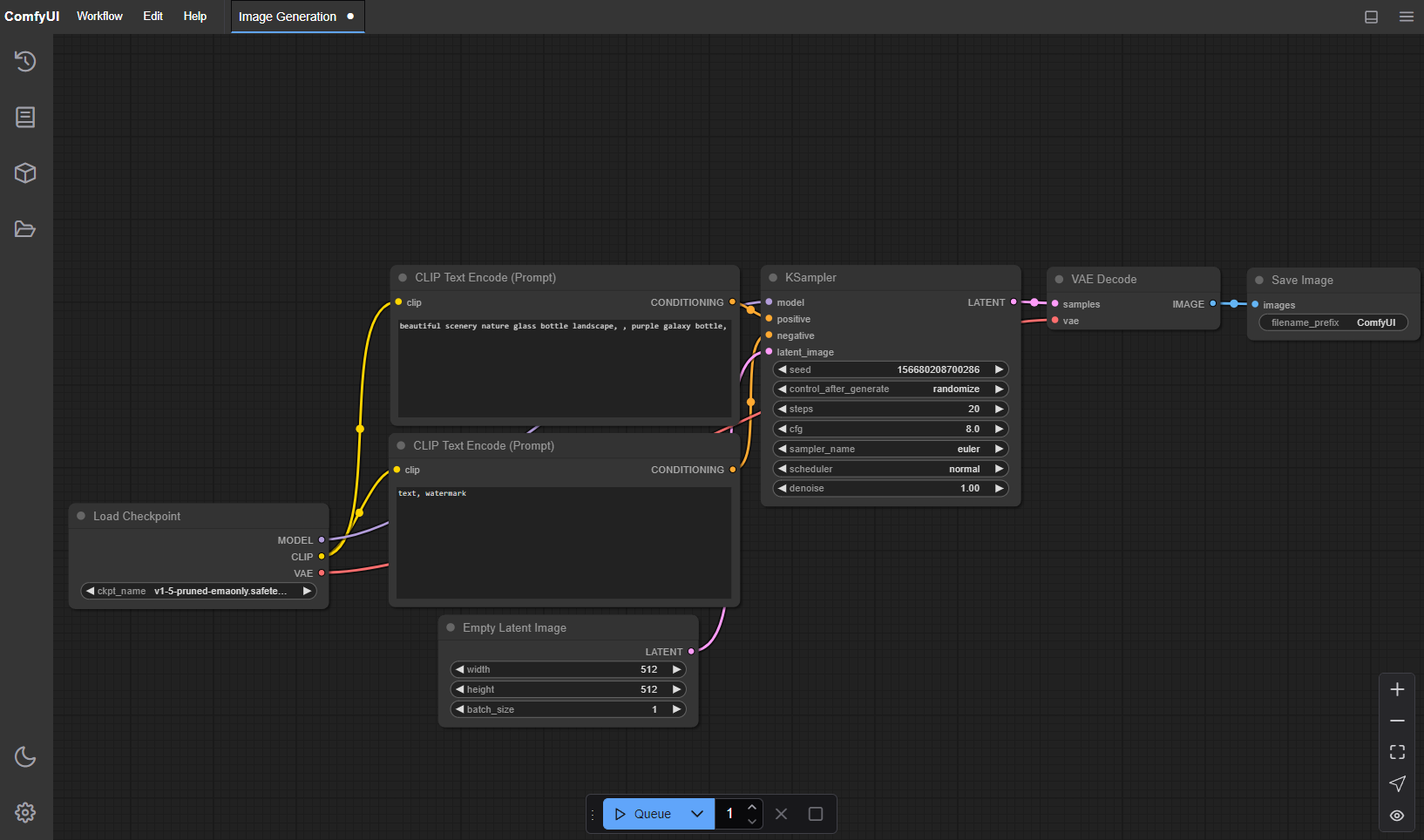
Another limitation of existing models is their accessibility to non-technical users. While Stable Diffusion is a powerful tool, many users find it challenging to interact with, especially without prior knowledge of AI or machine learning concepts. This is where the role of user-friendly interfaces like ComfyUI becomes essential. However, while ComfyUI has simplified some aspects of using Stable Diffusion, there are still improvements to be made in terms of enhancing the customization options and making the user experience even more intuitive.

My project aims to address these gaps by creating a more accessible and user-friendly solution for text-to-image generation using Stable Diffusion and ComfyUI. By simplifying the interaction process and exploring ways to optimize image generation, I hope to make this powerful AI tool available to a broader audience, including those without technical expertise.

**CHAPTER 3**

**Proposed Methodology**

* 1. **System Design**

****

1. **Loaded Checkpoint:** This refers to the pre-trained model (e.g., Stable Diffusion) that is loaded into the system for generating images. It provides the necessary weights and architecture for the model to work based on the input prompts.
2. **CLIP Text Encoder (Prompt Positive):** This component encodes the positive portion of the text prompt (the main description). It transforms the user’s input into a numerical format that the model can understand, helping the model generate an image that aligns with the prompt.
3. **CLIP Text Encoder (Prompt Negative):** Similar to the positive prompt encoder, the negative prompt encoder processes any negative descriptors. This is used to tell the model what elements to avoid or minimize in the generated image (e.g., "no animals" or "no bright colours").
4. **Empty Latent Image:** This is a placeholder image filled with random noise. The diffusion model starts the image generation process from this initial latent space and refines it through multiple iterations to produce the final image.
5. **K-Sampler:** The K-Sampler (K-LMS Sampler) is used to sample from the latent space. It controls the process of generating the image by guiding how the latent representation evolves into a more defined image, ensuring stable diffusion over multiple steps.
6. **VAE Decode:** The VAE (Variational Autoencoder) Decoder takes the processed latent image and decodes it back into a visible image. This step converts the low-dimensional latent space representation into the final, high-quality image that can be displayed to the user.
7. **Save Image:** After the image is generated, the “Save Image” component allows the user to save the final output as a file (such as a .png or .jpg) to their local system or elsewhere for use.
   1. **Requirement Specification**
      1. **Hardware Requirements:**
8. **CPU:** Any modern multi-core processor (e.g., Intel i5 or AMD Ryzen 5+).
9. **GPU:** NVIDIA GPU with CUDA support (6GB VRAM or more recommended).
10. **RAM:** Minimum 8GB, recommended 16GB for optimal performance.
11. **Storage:** At least 50GB of free space (SSD preferred).
    * 1. **Software Requirements:**
12. **OS:** Windows 10/11 or Linux (Ubuntu preferred).
13. **Python:** Version 3.8 or above.
14. CUDA and cuDNN (for NVIDIA GPUs).
15. **Required Libraries:** PyTorch, Pillow, etc.

**CHAPTER 4**

**Implementation and Result**

* 1. ** Snap Shots of Result:**

**Mystical Forest:**

**A serene, glowing forest illuminated by bioluminescent mushrooms and fireflies. Mist hovers above the ground, with tall, enchanted trees under a moonlit sky, creating an ethereal atmosphere.**

* **Include: "**Glowing mushrooms, fireflies, misty forest under a moonlit sky."
* **Exclude**: "No animals, no people, no daytime."

****

**Cyberpunk City**

A futuristic metropolis bathed in neon lights, with towering skyscrapers, high-tech billboards, and flying cars. The wet streets reflect the glow of advertisements, adding to the dystopian feel.

* **Include:** "Neon lights, skyscrapers, wet streets, futuristic cars, rain."
* **Exclude:** "No daytime, no nature, no rural settings."

****

**Majestic Dragon:**

A massive dragon with intricately detailed scales soars above snowy mountains, breathing fire as the sun sets behind dramatic, fiery clouds

* **Include: "**Fire-breathing dragon flying above snow-capped mountains, fiery sky."
* **Exclude:** "No people, no modern buildings."

****

**Steampunk Airship:**

A grand, vintage airship with brass and steam-powered mechanisms glides through golden clouds at sunrise, its intricate gears and sails catching the morning light.

* **Include:** "Vintage airship, flying through clouds, sunrise, steam-powered."
* **Exclude**: "No modern technology, no animals, no futuristic elements."

****

**New City:**

A stunning futuristic cityscape at sunset with towering skyscrapers, neon lights, and flying cars zipping through the sky. The streets glow with holographic billboards and illuminated pathways, creating a vibrant cyberpunk atmosphere without fog or blurriness.

* **Positive Prompt:** "A futuristic cityscape at sunset with flying cars, neon lights, and towering skyscrapers."
* **Negative Prompt:** "No animals, no blurry images, no excessive fog."

****

**Japanese Garden:**

A tranquil garden with a koi pond, cherry blossom trees in full bloom, and traditional stone lanterns. The reflection of the pink petals in the water enhances the peaceful ambiance.

* **Include:** "Traditional koi ponds, cherry blossoms, stone lanterns, serene atmosphere."
* **Exclude:** "No modern buildings, no crowds."

****

**Horror Asylum:**

A dark, decaying asylum with flickering lights, long shadows, and thick mist. A lone, eerie figure in a tattered hospital gown stands with their back turned, surrounded by ghostly, transparent figures. Strange symbols are etched on the walls, and the atmosphere feels heavy with dread.

**• Include:** "Flickering lights, decaying asylum, eerie figure, ghostly figures, strange symbols, thick mist, long shadows."

**• Exclude: "**No bright colours, no gore, no clear faces, no cliché monsters, no fast movements."

****

**Girl in Nature:**

A girl seated in a lotus position, adorned in flowing robes with gold and crimson patterns. Her hands are raised in a calm gesture, surrounded by soft, ethereal light. The tranquil background features distant mountains and delicate clouds**.**

**• Include:** "Serene girl, flowing robes, lotus position, blessing gesture, ethereal light, tranquil landscape, distant mountains." **• Exclude:** "No harsh contrasts, no aggressive elements, no dark tones, no ominous imagery."

****

**Post-Apocalyptic World:** An abandoned city overtaken by nature, with vines covering skyscrapers and rusting vehicles. Robots roam the streets as the setting sun casts a golden light over the ruined world.

* **Include**: "Abandoned cities, overgrown plants, abandoned vehicles, robots."
* **Exclude**: "No people, no wildlife, no bright skies."

**4.2GitHub Link for Code:**

[**https://github.com/data-manavpatil/ComfyDiffusion**](https://github.com/data-manavpatil/ComfyDiffusion)

**CHAPTER 5**

**Discussion and Conclusion**

During my internship project, I focused on utilizing Stable Diffusion in combination with ComfyUI for text-to-image generation. The integration of these tools enabled me to create high-quality images from textual prompts. Despite the success, there were limitations such as handling abstract prompts and refining the generated content based on negative prompts.

* 1. **Future Work:**

For future enhancements, I plan to improve model efficiency and speed for quicker image generation. Additionally, I would focus on refining the **negative prompt** handling for better control over image elements and explore advanced custom models for more specific applications. Lastly, improving the user interface and enabling real-time collaborative features could make the tool more accessible to a broader audience.

* 1. **Conclusion:**

This project has demonstrated the power of **Stable Diffusion** and **ComfyUI** in generating high-quality images from textual prompts. By utilizing these tools, I have gained hands-on experience in image generation, deep learning, and AI technologies. The project has not only deepened my understanding of model deployment but also showcased the potential of AI in creative applications. The successful completion of this project contributes to the growing field of AI-driven creativity and provides a foundation for further advancements in text-to-image generation.

**REFERENCES**

1. **Goodfellow, I., et al. (2014).** *Generative Adversarial Nets.*  
   You can find the original paper here: <https://arxiv.org/abs/1406.2661>
2. **Ho, J., et al. (2020).** *Denoising Diffusion Probabilistic Models.*  
   The paper is available here: <https://arxiv.org/abs/2006.11239>
3. **Rombach, R., et al. (2022).** *High-Resolution Image Synthesis with Latent Diffusion Models.*  
   You can find this paper here: <https://arxiv.org/abs/2112.10752>
4. **ComfyUI Documentation**  
   The official ComfyUI documentation can be found on their GitHub page or official website, depending on the specific tool or version you're using. Here’s a link to the GitHub repository (if applicable): <https://github.com/comfyanonymous/ComfyUI>