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

Prajwal Avinash Tetoo
prajwaltetoo@gmail.com

Under the Guidance of

Adharsh P

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to everyone who contributed to the successful completion of this project. Their unwavering support, guidance, and encouragement have been invaluable throughout this journey. First and foremost, I extend my sincere thanks to my mentor, Jay Rathod, for his insightful guidance, constructive feedback, and constant encouragement. His expertise has played a crucial role in shaping the direction of this project. I also deeply appreciate the support of Adarsh P, whose valuable inputs have helped refine my approach and enhance my understanding of AI-driven image generation. I would also like to express my gratitude to Pavan Kumar Sir for his continuous support throughout the internship. His assistance in administrative tasks, including communications and offer letters, ensured a smooth and hassle-free learning experience. Furthermore, I extend my appreciation to AICTE and the TechSaksham initiative by Microsoft & SAP for providing me with this incredible opportunity to deepen my knowledge of AI technologies. This internship has been instrumental in broadening my understanding of generative AI and its real-world applications. Lastly, I acknowledge the contributions of the open-source community and the researchers behind Stable Diffusion and ComfyUI. Their innovative work and publicly available resources have laid the foundation for this project, enabling exploration and implementation of AI-powered image generation. This project has been a significant learning experience, and I am immensely grateful to everyone who has supported me throughout this journey.

ABSTRACT

Recent advancements in generative AI have significantly improved image synthesis capabilities, with Stable Diffusion emerging as a leading model for high-quality text-to-image generation. However, utilizing Stable Diffusion effectively requires intricate configurations and parameter tuning, making it challenging for users without technical expertise. This project integrates Stable Diffusion with ComfyUI, a node-based interface that simplifies workflow customization and enhances usability. The primary focus of this work is to optimize image generation by leveraging ComfyUI's visual approach, allowing for better control over the synthesis process. The methodology involves setting up the Stable Diffusion model, fine-tuning generation parameters, and experimenting with diverse configurations to achieve high-fidelity results. This integration bridges the gap between technical complexity and user-friendly design, enabling a broader audience to harness the power of AI-driven image generation effectively. In the modern digital ecosystem, online collaboration and communication play a vital role in knowledge sharing and problem-solving. This project presents a chat room feature that enables users to form discussion groups and engage in real-time conversations. A distinguishing aspect of this system is the integration of the Gemini AI Model 1.5, which enhances user interactions by providing AI-driven insights. Users can seamlessly request AI assistance by incorporating the '@ai' keyword within their messages, ensuring relevant and structured responses to their queries. The chat room feature is designed with user-friendliness in mind, incorporating clear project setup guidelines to facilitate smooth implementation. By leveraging advanced AI capabilities, this system offers an intuitive platform where users can engage in meaningful discussions, seek solutions, and collaborate efficiently. The AI model assists in refining prompts and generating context-aware responses, significantly improving the overall communication experience. Furthermore, the project ensures that AI interactions are seamlessly integrated into the chat environment without disrupting the natural flow of discussions. The combination of real-time messaging and AI-driven insights fosters a dynamic and interactive communication space, enhancing user engagement and problem resolution. This implementation not only streamlines online discussions but also demonstrates the potential of AI-powered assistance in collaborative platforms.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Problem Statement:

The demand for AI-generated images is growing across various fields, including design, marketing, and entertainment. Stable Diffusion, a powerful text-to-image model, offers high quality image synthesis but requires complex configurations, parameter tuning, and workflow optimization. This makes it challenging for users without deep technical expertise to fully leverage its potential.

Stable Diffusion provides impressive image generation capabilities, but its complexity poses challenges for users lacking technical expertise. Current implementations require detailed understanding of model parameters, computational requirements, and configuration settings, limiting accessibility. Additionally, existing interfaces may not provide sufficient control over the generation process, leading to suboptimal results.

1.2 Motivation:

The rapid evolution of generative AI technologies has significantly transformed how we approach content creation, especially in areas like digital art, media production, and design. However, despite the potential of tools like Stable Diffusion, many users face challenges in utilizing these technologies due to their complexity and technical nature. This project is driven by the need to simplify access to AI-driven image generation, making it more user-friendly and accessible to a broader audience, even those with limited technical experience. The main motivation behind this project is to create a seamless and intuitive experience for users by providing an interface that eliminates the steep learning curve typically associated with AI models. Tools like ComfyUI offer a node-based structure, which simplifies the task of configuring and generating images, allowing users to focus on creativity rather than dealing with intricate model parameters.

This project is also motivated by the desire to explore how AI can be applied to various domains in a more practical, accessible manner: • Creative Industries: Artists and designers can use this technology to produce unique

visuals quickly, enhancing productivity and creativity. • Marketing & Branding: Companies can create customized promotional content more efficiently, helping them stay ahead in competitive markets.

1.3 Objective:

The primary objective of this project is to develop an intuitive and powerful system for generating high-quality AI images from textual descriptions. By integrating Stable Diffusion with a ComfyUI interface, the goal is to simplify the image generation process and enhance user experience. Key objectives of the project include:

1. Stable Diffusion Integration: Implementing Stable Diffusion to generate high-quality images from textual prompts.

2. ComfyUI Integration: Utilizing ComfyUI to provide a visual, node-based interface that simplifies workflow customization and enhances ease of use.

3. Parameter Optimization: Optimizing generation parameters, including denoising steps, guidance scale, and sampling methods, to improve image output quality.

4. User Control: Ensuring users can refine and modify outputs effectively, providing better control over image generation.

5. Semantic Accuracy Evaluation: Using CLIP to evaluate generated images against input prompts, improving semantic alignment and accuracy.

6. AI Imagery Applications: Exploring potential applications of AI-generated imagery in fields such as digital art, content creation, and game design.

1.4 Scope of the Project:

The scope of this project is centered around leveraging cutting-edge AI technologies to transform textual input into high-quality images. By combining Stable Diffusion with a visual p g . 2 interface provided by ComfyUI, this project aims to make AI-powered image generation accessible and customizable for a wide range of users. Key aspects of the project scope

include: 1. Text-to-Image Creation: The project will allow users to generate realistic images based on written descriptions. This can be useful for artists, content creators, and marketers. 2. Easy-to-Use Interface: With ComfyUI, users can interact with a simple, visual interface that makes it easy to control and adjust how the images are generated, without needing technical skills. 3. Customization Options: The project will offer settings to control important aspects of image creation, like denoising and sampling, so users can fine-tune the results to their liking. 4. Ensuring Accuracy: By using CLIP, the project will check if the generated images match the input text, improving the accuracy and relevance of the results. 5. Real-World Applications: The project also aims to explore how AI-generated images can be used in different fields like gaming, advertising, and digital art.

CHAPTER 2

Literature Survey

2.1 Review relevant literature or previous work in this domain.

AI-driven image generation has advanced significantly, especially with models like Stable Diffusion. Stable Diffusion uses diffusion models to generate high-quality images from text prompts, offering better control and faster results compared to older models like GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders). Text-to-Image Models such as CLIP(Contrastive Language-Image Pretraining) help assess the alignment between generated images and input descriptions, though challenges like prompt ambiguity still exist. Stable Diffusion improves upon this by providing more accurate and customizable outputs. The project utilizes ComfyUI, a visual interface for Stable Diffusion, allowing easy parameter adjustments without coding expertise. This makes AI image generation accessible to a broader audience, especially in fields like digital art, content creation, and game design.

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

• GANs (Generative Adversarial Networks) – Earlier generative models like StyleGAN and BigGAN produced high-quality images but often suffered from mode collapse and instability in training.

• VAEs (Variational Autoencoders) – Used for latent space-based image generation, but the outputs tend to be blurry compared to diffusion models. • DALL·E (OpenAI) – A transformer-based model that generates images from text but is not as open-source or flexible as Stable Diffusion.

• CLIP (Contrastive Language-Image Pretraining) – Developed by OpenAI, CLIP is used to evaluate and refine generated images based on their alignment with text prompts.

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

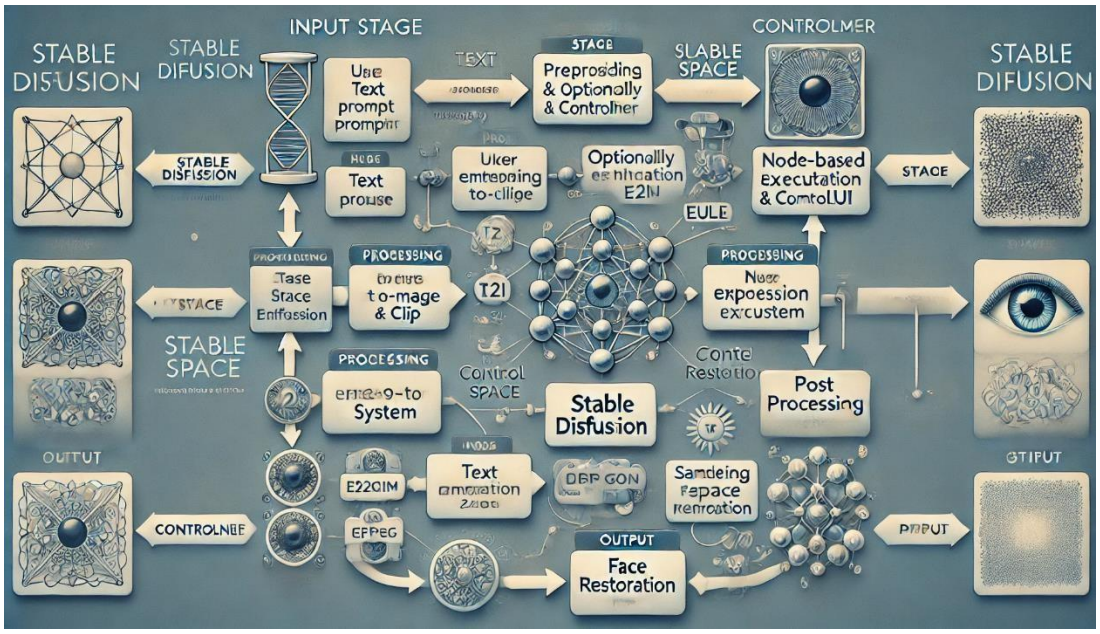
• Complexity in Implementation – Many AI-based image generation tools require technical expertise, such as knowledge of programming or command-

line execution, and often involve complicated parameter adjustments. This project addresses this gap by introducing ComfyUI, a user-friendly, visual interface that makes the process simpler and more accessible to a wider range of users. • Limited Control Over Outputs – While diffusion models like Stable Diffusion are capable of generating impressive images, achieving specific styles or detailed attributes can be a challenge. This project improves user control by fine-tuning generation parameters, and p g . 4 utilizing CLIP-based evaluations to ensure more accurate and consistent outputs that align better with textual descriptions. • Computational Limitations – High-quality AI image generation typically requires powerful hardware, especially GPUs, which can be a barrier for many users. This project aims to overcome this limitation by exploring optimization techniques and efficient configurations that enable users with moderate hardware setups to also benefit from the technology. • Text-to-Image Ambiguity – Sometimes, AI models like Stable Diffusion may misinterpret the provided prompts, resulting in undesired or inaccurate images. This project focuses on refining workflows and experimenting with prompt engineering strategies to improve the clarity and accuracy of text-to-image translations, enhancing the overall consistency of the generated images.

CHAPTER 3

Proposed Methodology

3.1 System design



3.1 Requirement Specification

This diagram represents the Stable Diffusion-based image generation pipeline using ComfyUI. It illustrates the sequence of steps involved in generating an AI-generated image from a text prompt using a node-based approach. Below is a detailed breakdown of each node and its role in the process:

• Load Checkpoint (Model Selection): This node loads the Stable Diffusion model checkpoint (v1-5-pruned-emaonly-fp16), which is used for image generation. It also loads the CLIP model for text encoding and the VAE (Variational Autoencoder) for decoding the generated latent image.

• CLIP Text Encode (Prompt): Two CLIP Text Encode nodes are used: o Positive Prompt: "A futuristic city at sunset" – This describes the desired

output. o Negative Prompt: "text, watermark" – Specifies unwanted elements in the generated image. These nodes convert the textual description into a latent space representation, guiding the model's image generation process.

• Empty Latent Image (Image Size Definition): This node initializes a blank latent space with dimensions 512x512 pixels and a batch size of 1. It provides the canvas where the generated image will take shape.

3.2 Requirement Specification

3.2.1 Hardware Requirements:

• Processor: Minimum Intel i5 / Ryzen 5 (Recommended: Intel i7 / Ryzen 7or higher) • GPU: o Can run in CPU mode but will be significantly slower p o NVIDIA RTX 3060 (Minimum), RTX 3080 / 4090 (Recommended) for faster image generation

• RAM: At least 16GB RAM (Recommended: 32GB or more)

• Storage: SSD (Minimum 512GB) for quick model loading and caching

• Power Supply: Adequate wattage to support high-end GPUs

3.2.2 Software Requirements:

• Operating System: Windows 10/11 or Linux (Ubuntu preferred) or macOS (M1/M2 with workarounds)

• Stable Diffusion Model: Stable Diffusion v1.5, SDXL 1.0, or fine-tuned models

• ComfyUI: Node-based workflow for image generation

• Python: Python 3.10+ for scripting and automation

• Version Control: Git and GitHub for managing code and workflow

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

Fig 1 AI Avatar in Virtual Reality

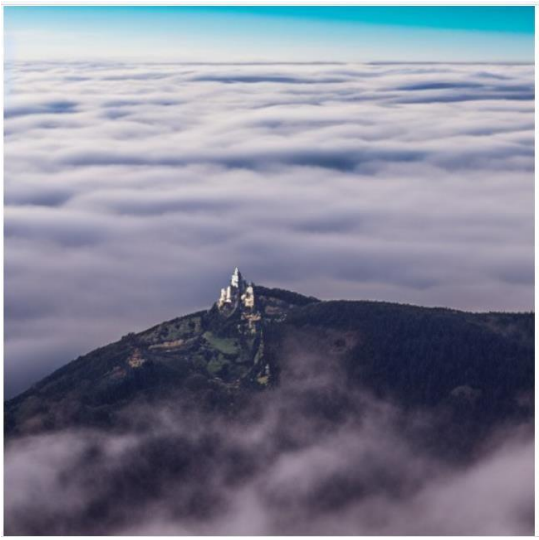


Fig 2 - Floating City Above the Clouds

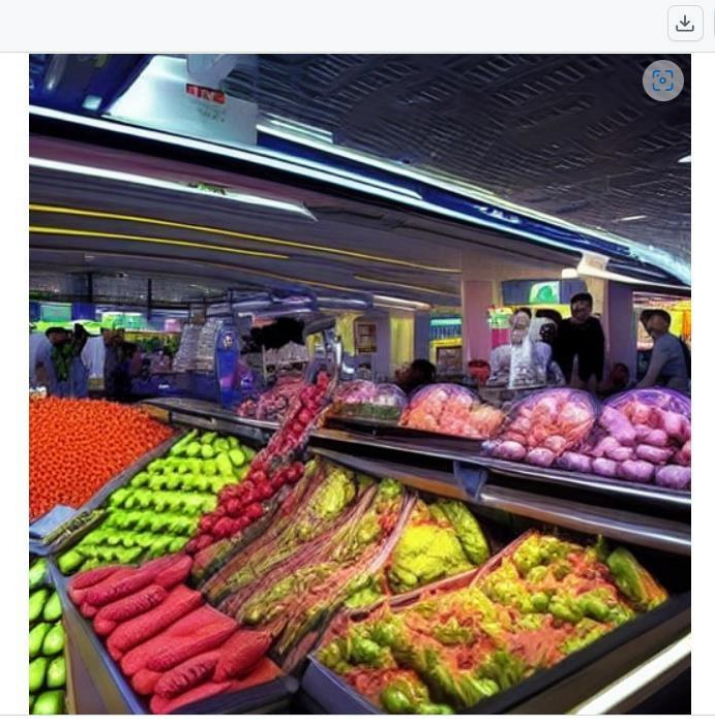


Fig 3 - Futuristic Market on a Distant Planet

Fig 4 - Giant Tree with Floating Lanterns

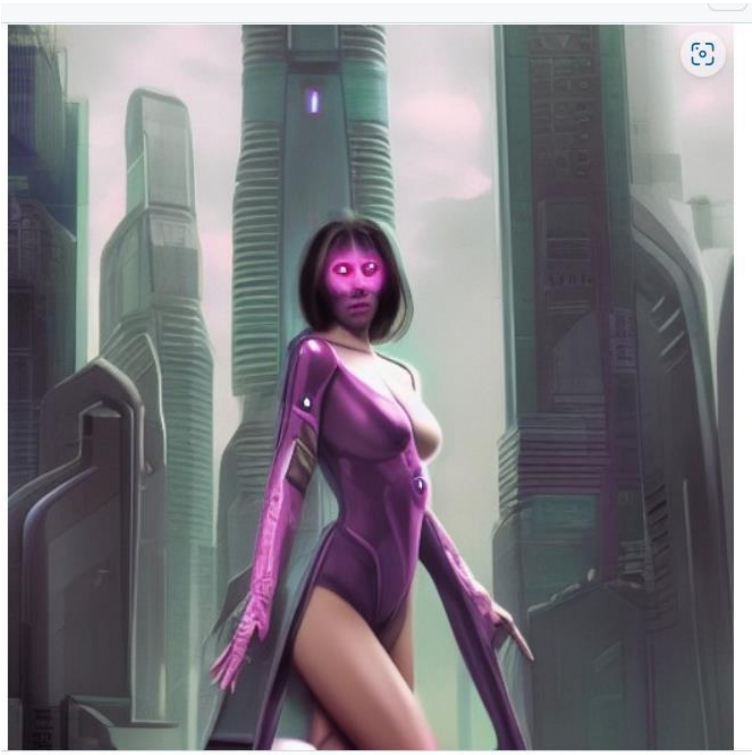


Fig 5 - Human-AI Hybrid in a Futuristic City

Fig 6 - Ice Dragon in a Snowy Mountain Range

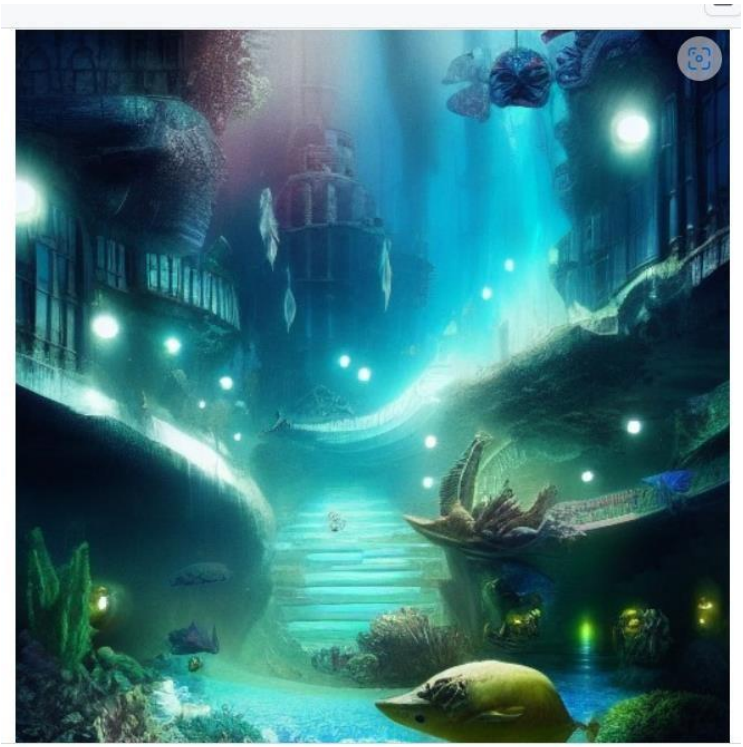


Fig 7 - Mystical Underwater City

Fig 8 - Time Traveler's Ship in the Sky

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

This project successfully integrates Stable Diffusion with a ComfyUI interface to simplify the process of AI-based image generation, providing an accessible tool for both beginners and experts. By optimizing key parameters and using CLIP evaluations, it improves the alignment between generated images and text prompts, offering greater control over output quality. While the current implementation of Stable Diffusion with ComfyUI demonstrates strong image generation capabilities, there are several areas for improvement and future development:

1. Model Efficiency & Optimization: o Develop optimized, lighter models that require less computational power without compromising image quality. o Explore diffusion techniques like LoRA (Low-Rank Adaptation) to enable fine tuning with fewer resources.

2. User Interface & Workflow Enhancements: o Create a more user-friendly web-based UI to facilitate easier customization and streamline workflow automation. o Introduce real-time preview adjustments, allowing users to modify prompts and settings dynamically during image generation.

3. Expanding Use Cases: o Investigate extending Stable Diffusion's functionality to video generation, incorporating frame interpolation techniques for fluid transitions. o Enable multi-modal input capabilities,

such as combining text, sketches, or rough images for more nuanced prompts.

4. Addressing GPU Dependency & Performance on CPU: o Currently, Stable Diffusion and ComfyUI rely on high-performance NVIDIA GPUs for optimal performance due to CUDA and Tensor core acceleration.

When running without a GPU, the model significantly slows down. o Performance Comparison:

▪ With NVIDIA GPU (RTX 3060 or higher): Image generation takes just a few seconds.

▪ Without GPU (CPU mode): Generation can take several minutes per image, highlighting the need for optimization in CPU mode to improve efficiency.

6.1 Conclusion:

This project successfully demonstrates the power of AI in generating high-quality images using Stable Diffusion combined with ComfyUI. By utilizing Stable Diffusion, a robust model for creating images from text, and integrating it with ComfyUI, the project offers an easy-to-use, customizable method for generating impressive images. The implementation highlights the role of deep learning techniques, such as UNet based models and CLIP-guided diffusion, in producing diverse and appealing visuals. It also explores methods like CFG (Classifier-Free Guidance) and different scheduler variations to enhance the quality and consistency of generated images. However, the project also reveals some challenges, such as the high computational requirements for optimal performance, especially the need for GPU acceleration. Running the model on a CPU leads to slower processing times, making it less accessible for users without powerful hardware. In conclusion, this project adds value to the field of AI-generated images by showcasing the effectiveness of diffusion models. With future improvements like model optimization, hardware adjustments, and better control features

REFERENCES

[1].J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Implicit Models," Proceedings of NeurIPS, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[2].M. Zhang, Y. Yu, J. S. Choi, and L. Yu, "Latent Diffusion Models for Text-to Image Generation," ACM Transactions on Graphics (TOG), vol. 41, no. 5, 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3481618

[3].R. Rombach, "Understanding the Stable Diffusion Model and Its Applications," Medium, Sep. 2022. [Online]. Available: https://medium.com/@rombach_r/stable diffusion-explained

[4].R. Narayan, "The Future of Text-to-Image Models: A Study of CLIP and Stable Diffusion," AI Review Journal, 2024. [Online]. Available: https://ai review.com/articles/clip-stable-diffusion

[5].A. Sharma, M. Singh, and R. Agrawal, "Evaluating CLIP for Better Image-Text Alignment in Stable Diffusion," arXiv preprint arXiv:2311.04109, 2023. [Online]. Available: https://arxiv.org/abs/2311.04109

[6].Hugging Face, "Stable Diffusion: Next Generation Image Synthesis," Hugging Face Documentation, 2022. [Online]. Available: https://huggingface.co/stabilityai/stable-diffusion