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

L Anunai Sai Goud, anunai6966l@gmail.com

Under the Guidance of

Jay Rathod

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during this thesis work.

Firstly, we would like to thank my supervisor, Jay Rathod, for being a great mentor and the best adviser I could ever have. His advice, encouragement and the critics are a source of innovative ideas, inspiration and causes behind the successful completion of this project. The confidence shown in me by him was the biggest source of inspiration for me. It has been a privilege working with him for the last one year. He always helped me during my project and many other aspects related to the program. His talks and lessons not only help in project work and other activities of the program but also make me a good and responsible professional.

I would like to express my sincere gratitude to Jay Rathod and the entire team at TechSaksham – A joint CSR initiative of Microsoft & SAP, focusing on AI Technologies. for their invaluable guidance and support throughout my internship on Image Generation using Stable Diffusion & ComfyUI.

I extend my heartfelt thanks to my professors and peers for their encouragement and insightful discussions that helped enhance my understanding of diffusion models, U-Net architecture, and AI-based image generation techniques.

Finally, I am grateful for the opportunity to work on this project, which has deepened my knowledge of generative AI and its practical applications.

ABSTRACT

AI-powered image generation has gained significant traction, with Stable Diffusion emerging as a leading model for creating high-quality images from textual descriptions. This project explores the implementation of *Stable Diffusion* using *ComfyUI*, focusing on the underlying diffusion models and the U-Net architecture.

The objective of this project was to understand the working principles of diffusion models, particularly how images are generated from noise through iterative denoising. The study also aimed to implement and test ComfyUI, an intuitive graphical interface for Stable Diffusion, to simplify model usage and customization.

The methodology involved setting up ComfyUI, integrating the Stable Diffusion model, and experimenting with different parameters to fine-tune image generation. Key concepts such as the forward and reverse diffusion process, the role of CLIP in text-to-image conversion, and the function of U-Net in feature extraction and reconstruction were analyzed.

The results demonstrated how diffusion models can progressively refine noise to generate coherent and detailed images. The use of skip connections in the U-Net architecture proved essential in preserving image details during the denoising process. By leveraging ComfyUI, the project successfully streamlined the image generation workflow, allowing for more efficient model experimentation.

In conclusion, this project provided a hands-on understanding of AI-based image generation, bridging theoretical knowledge with practical implementation. The insights gained will be beneficial for further exploration in generative AI applications.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1Problem Statement:

AI-generated images have transformed various fields, including digital art, design, and content creation. However, implementing advanced models like *Stable Diffusion* remains complex due to the need for in-depth knowledge of diffusion processes, U-Net architecture, and text-to-image conversion. Traditional generative models, such as GANs, often struggle with stability and image coherence, whereas diffusion models provide more reliable results. Despite their advantages, using diffusion models effectively requires a structured approach. This project aims to address these challenges by exploring *Stable Diffusion* with *ComfyUI*, a user-friendly interface that simplifies model interaction. By bridging theoretical concepts with practical implementation, the study enhances accessibility to AI-driven image generation.

1.2Motivation:

AI-driven image generation has revolutionized fields like digital art, game development, and content creation, with *Stable Diffusion* emerging as a powerful model for generating high-quality images from text. However, its complexity makes it challenging for beginners and developers to implement effectively. This project was chosen to simplify the understanding and application of *Stable Diffusion* using *ComfyUI*, a user-friendly interface that streamlines the process. By exploring the diffusion process, U-Net architecture, and CLIP’s role in text-to-image generation, this study enhances accessibility to AI-generated visuals. With applications in graphic design, advertising, and AI-assisted creativity, this project bridges the gap between theoretical knowledge and practical implementation, making AI-driven art more approachable and impactful.

1.3Objective:

The objective of this project is to understand and implement *Stable Diffusion* using *ComfyUI* for AI-driven image generation. It aims to explore the diffusion process, U-Net architecture, and the role of CLIP in text-to-image conversion. Additionally, the project seeks to simplify the workflow for generating high-quality images, making the technology more accessible to users. By bridging theory with practical application, the study enhances understanding of generative AI and its real-world applications.

1.4Scope of the Project:

This project focuses on implementing *Stable Diffusion* using *ComfyUI* to generate images from text prompts. It covers key concepts like diffusion models, U-Net architecture, and CLIP for text-to-image conversion. The scope is limited to exploring ComfyUI’s capabilities without modifying the core Stable Diffusion model or training custom datasets..

CHAPTER 2

Literature Survey

2.1 Review of Relevant Literature

AI-generated image synthesis has evolved significantly, with diffusion models gaining prominence over traditional approaches like GANs. *Stable Diffusion* is a leading diffusion model that progressively refines noise to generate high-quality images. The *U-Net architecture* plays a crucial role in this process by encoding and decoding image features, while *CLIP* enhances text-to-image alignment, improving output accuracy. Recent studies highlight how diffusion models provide more stability and diversity in image generation, making them a preferred choice for AI-driven creativity.

2.2 Existing Models, Techniques, and Methodologies

Traditional generative models like *GANs* and *VAEs* have been widely used for AI-based image generation, but they often struggle with instability and limited control over outputs. Diffusion models, particularly *Stable Diffusion*, overcome these limitations by using a step-by-step denoising process, allowing for more controlled and detailed image synthesis. The *U-Net architecture* enables efficient feature extraction, while *CLIP* helps interpret text prompts accurately. Tools like *ComfyUI* further simplify model interaction, making diffusion-based image generation more accessible.

2.3 Limitations in Existing Systems

Despite their advancements, diffusion models require significant computational power and a deep understanding of their architecture, making them challenging for beginners. GANs, while effective in some cases, often suffer from mode collapse and training instability. Additionally, most AI image-generation frameworks lack user-friendly interfaces, making parameter tuning difficult for non-experts. This project aims to address these issues by utilizing *ComfyUI*, which provides an intuitive interface for *Stable Diffusion*, allowing users to experiment with AI-generated images more efficiently.

CHAPTER 3

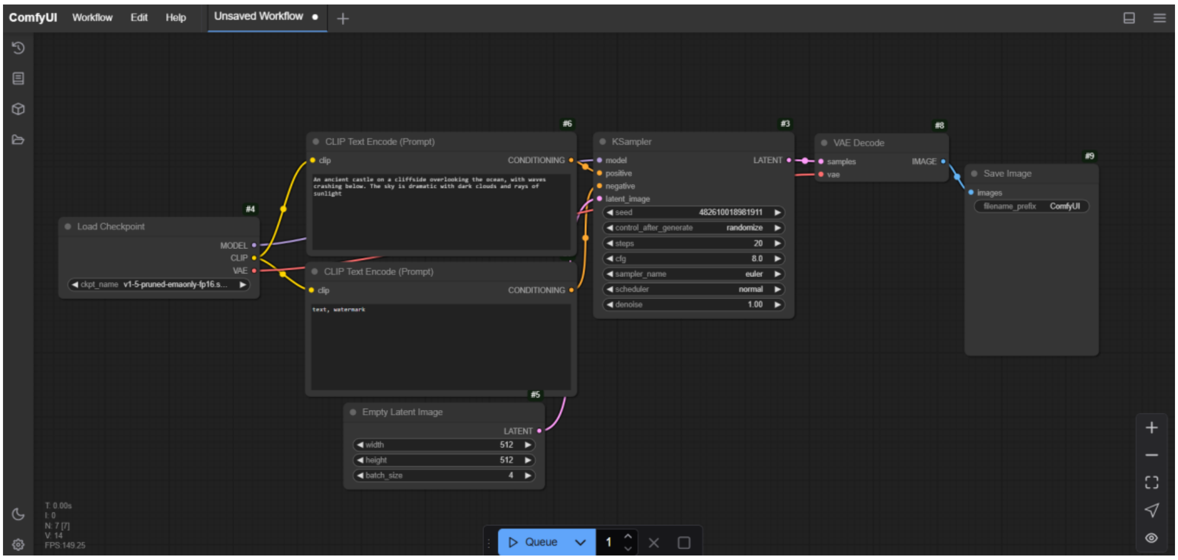
Proposed Methodology

3.1 System Design

The proposed system converts text prompts into images using *Stable Diffusion* and *ComfyUI*. First, the user inputs a prompt, which is processed by CLIP to generate a text embedding. A random noise image is then initialized, and the U-Net architecture progressively removes noise step by step to refine the image. Finally, the VAE decoder converts the processed data into a viewable image, which is displayed in ComfyUI, allowing users to fine-tune parameters for better results. This workflow ensures a seamless and efficient AI-driven image generation process. Requirement Specification

Mention the tools and technologies required to implement the solution.

The image shows a *ComfyUI* workflow for generating images using *Stable Diffusion*. It starts with loading a model, encoding text prompts, initializing latent space, applying the diffusion process via KSampler, decoding the latent image with VAE, and saving the final output.



CLIP Text Encode (Prompt) #6: An ancient castle on a cliffside overlooking the ocean, with waves crashing below. The sky is dramatic with dark clouds and rays of sunlight

CLIP Text Encode (Prompt): text, watermark

Load Checkpoint #4: ckpt_name v1-5-pruned-emaonly-fp16.s...

KSampler #3: seed 482610018981911; control_after_generate randomize; steps 20; cfg 8.0; sampler_name euler; scheduler normal; denoise 1.00

Empty Latent Image #5: width 512; height 512; batch_size 4

Save Image #9: filename_prefix ComfyUI

Fig1: explaining the workbench

3.1.1 Hardware Requirements:

To efficiently run *Stable Diffusion* using *ComfyUI*, a powerful GPU is essential. A NVIDIA GPU with at least 8GB VRAM (such as RTX 3060 or higher) is recommended to ensure smooth and fast image generation. The CPU should be at least an Intel i5/i7 **or**

AMD Ryzen 5/7, as it plays a crucial role in handling computational tasks alongside the GPU.

A minimum of 16GB RAM is necessary to support large model files and high-resolution image generation without system slowdowns. Additionally, at least 50GB of free storage is required to accommodate model checkpoints, dependencies, and generated images. The system should run on **Windows, Linux, or macOS** with GPU acceleration enabled for optimal performance. Higher-end hardware configurations will lead to faster processing and better-quality image outputs.

3.1.2 Software Requirements:

To implement Stable Diffusion with ComfyUI, several software components need to be installed. The system should run on Windows, Linux, or macOS with proper GPU drivers to support hardware acceleration. A Python environment, preferably Python 3.10 or higher, is required to execute the diffusion model and manage dependencies.

The core software includes Stable Diffusion models, which can be downloaded from sources like Hugging Face, and ComfyUI, which provides a user-friendly interface for managing the image-generation workflow. Additional libraries such as PyTorch, Torchvision, xFormers, and NumPy are essential for running the model efficiently. A package manager like pip or Conda is needed to install and manage these dependencies, ensuring smooth execution of the diffusion process.

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

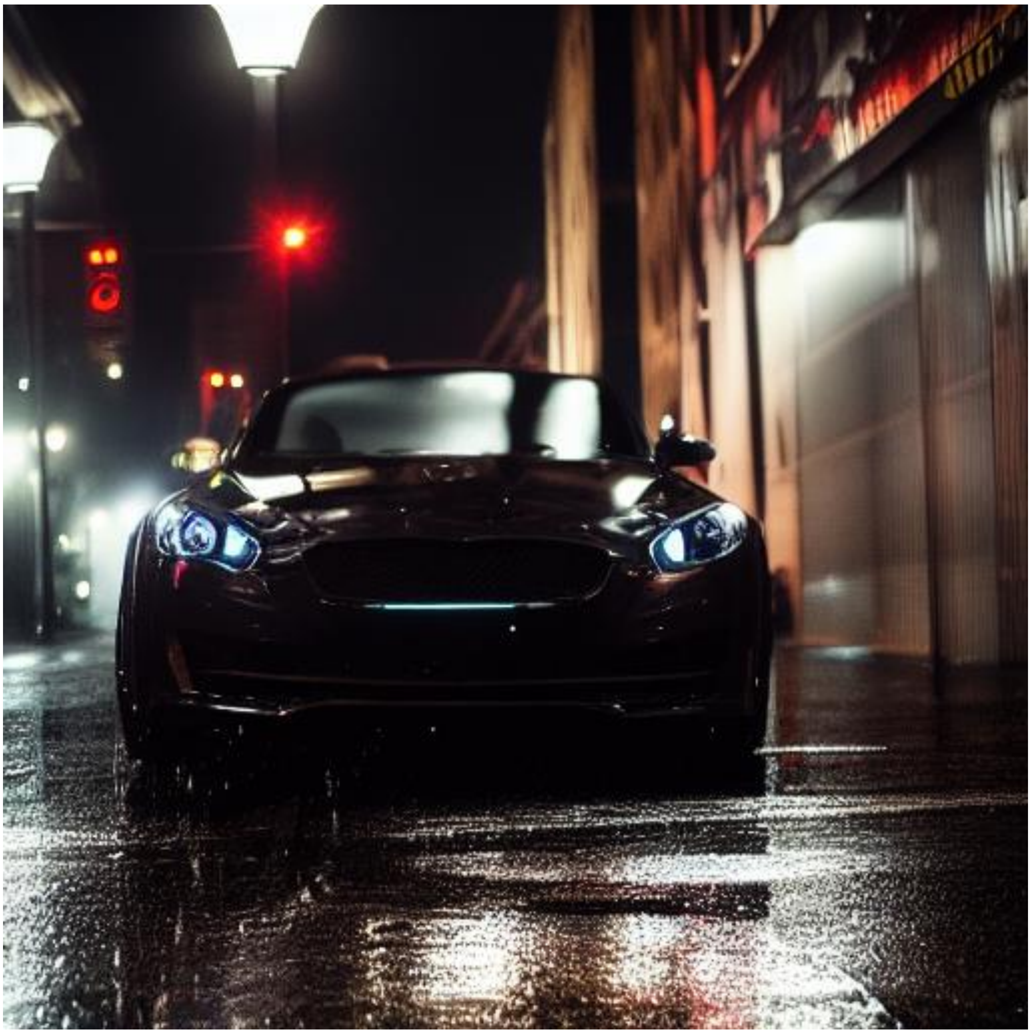


Fig2: A black sports car parked on a rainy city street at night, with neon lights reflecting on the wet road. The scene is cinematic and moody.

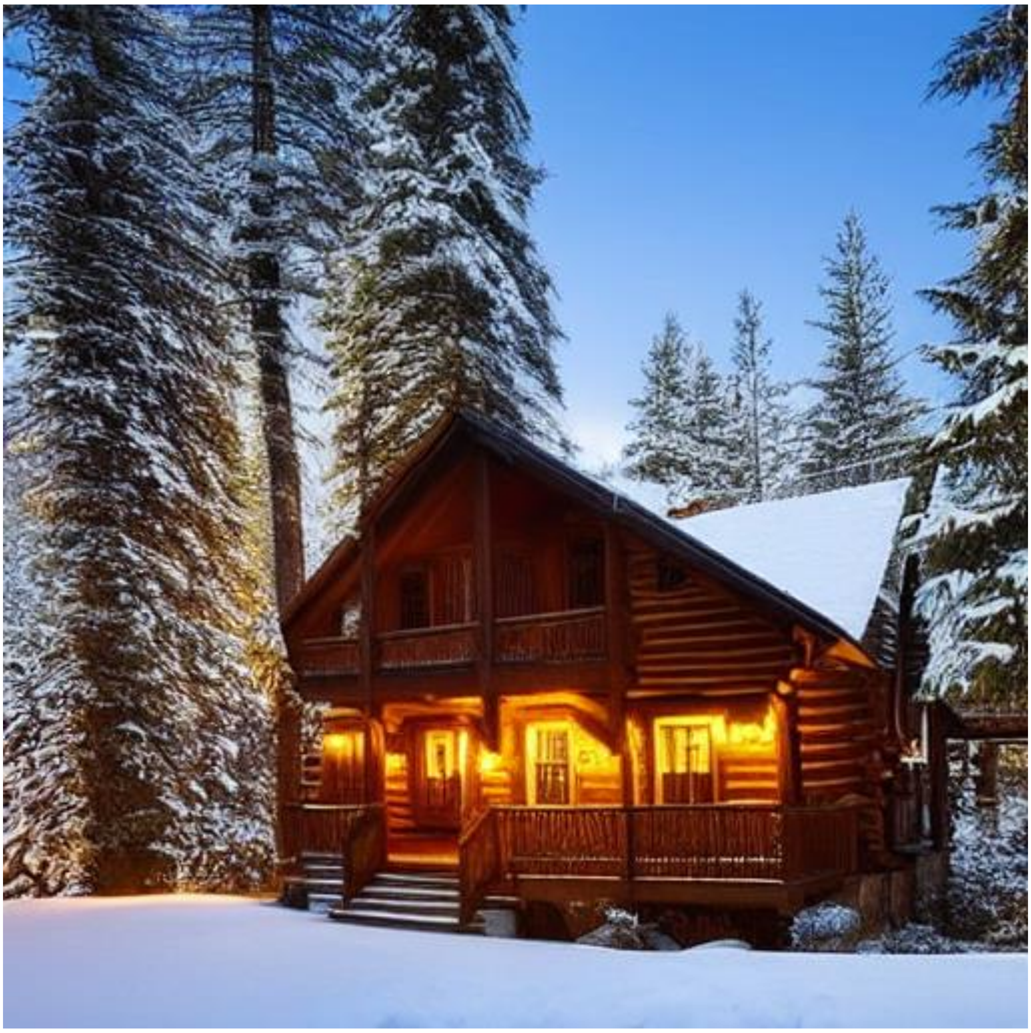


Fig3: A cozy wooden cabin in the mountains, surrounded by tall pine trees and covered in fresh snow. The warm glow of lights shines from the windows.

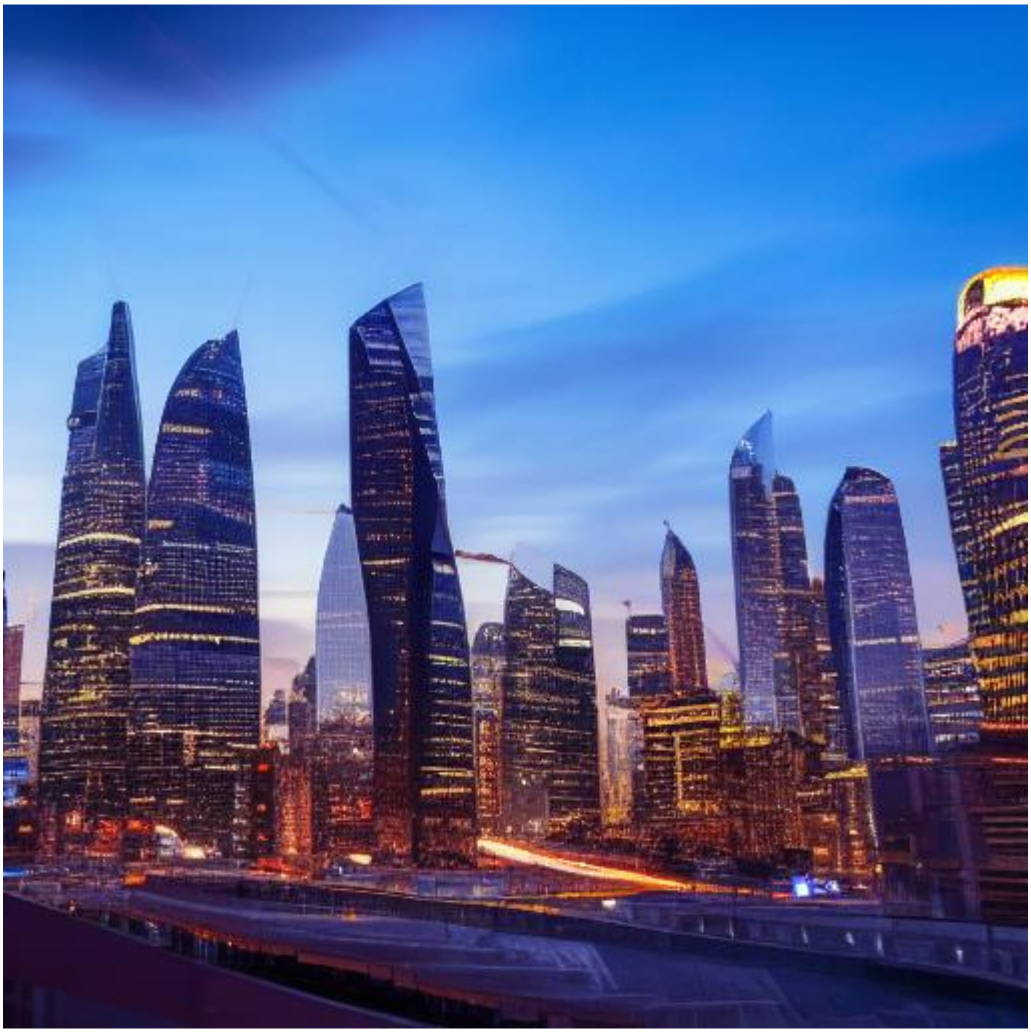


Fig4: A futuristic city skyline at dusk, with flying cars and neon-lit skyscrapers. The sky has a mix of deep blues and purples.

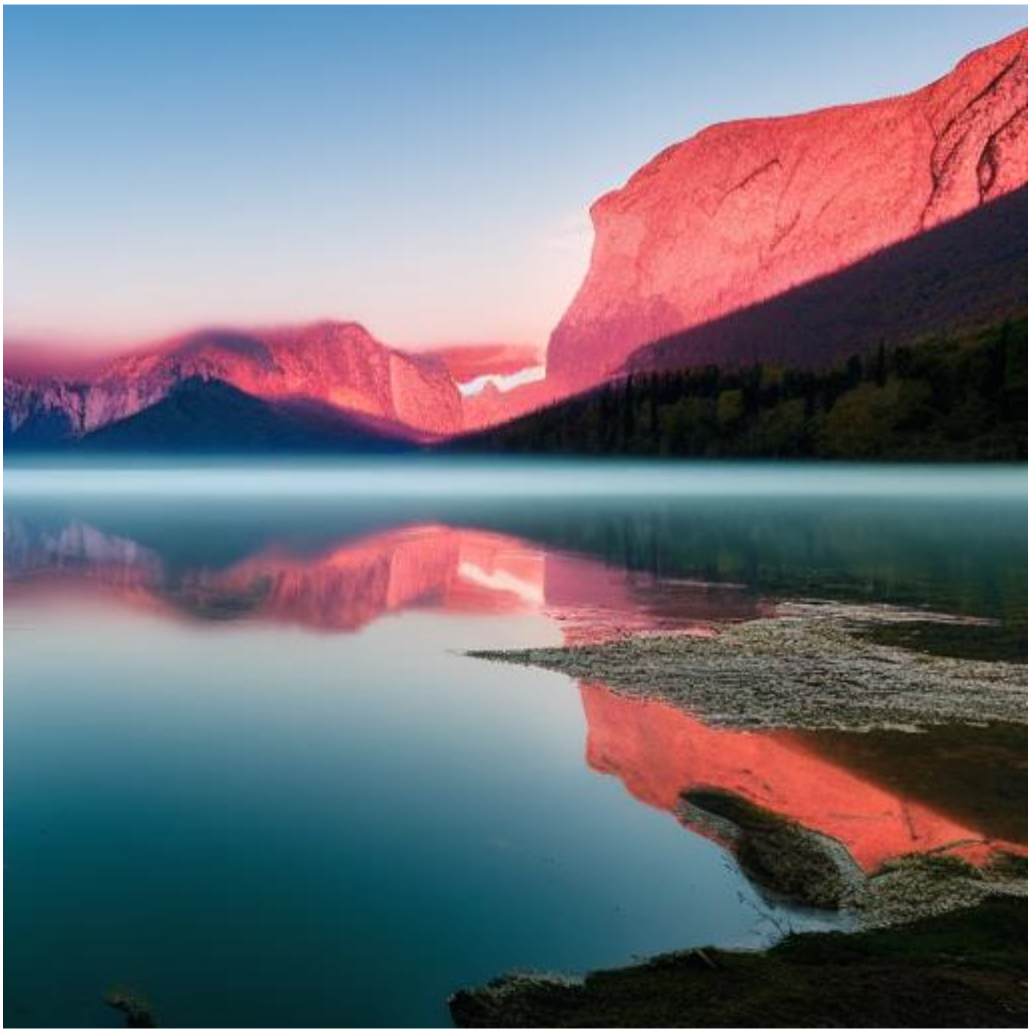


Fig5: A serene lake at sunrise, with mist rising from the water and mountains in the background. The scene is calm and peaceful.

4.2GitHub Link for Code:

https://github.com/Anunai6966/image-generation-using-comfyUI-

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

Future work can focus on optimizing the model for faster image generation by improving hardware utilization and refining sampling techniques. Enhancing user control over image attributes, integrating better negative prompt handling, and incorporating advanced upscaling methods can further improve output quality. Expanding support for more fine-tuned models and adding automated prompt enhancement features can also make the system more efficient and user-friendly.

Conclusion:

This project successfully demonstrated the process of generating high-quality images using Stable Diffusion and ComfyUI. By leveraging diffusion models, CLIP text encoding, and U-Net architecture, the system efficiently transforms text prompts into visually coherent images. The implementation provides a flexible and user-friendly approach to AI-generated art, making it accessible for both beginners and advanced users. The project highlights the potential of diffusion models in creative applications and sets the foundation for future improvements in efficiency, customization, and image quality.

REFERENCES

[1]. Official Stable Diffusion Documentation – Stability AI

[2]. ComfyUI GitHub Repository

[3]. Hugging Face – Stable Diffusion Model Downloads

[4]. Research Paper: "High-Resolution Image Synthesis with Latent Diffusion Models" – Rombach et al. (2022)

[5]. Research Paper: "Denoising Diffusion Probabilistic Models" – Ho et al. (2020)

[6]. OpenAI's CLIP Model Documentation

[7]. Tutorials and Blogs on AI Image Generation from Medium, Towards Data Science, and ArXiv