Imagee 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

K Akshitha, 1GA21AD024@gat.ac.in

Under the Guidance of

Mr. Jay Rathod

ACKNOWLEDGEMENT

I would like to extend my heartfelt appreciation to my mentor, Mr. Jay Rathod, for his invaluable guidance and support throughout my journey in exploring image generation using Stable Diffusion and the ComfyUI project over the past month.

His patience, knowledge, and encouragement have played a crucial role in deepening my understanding of complex concepts and their practical applications. His ability to simplify challenging topics and provide hands-on experience has not only refined my skills but also motivated me to further explore this field.

This journey would not have been as insightful and rewarding without his mentorship. I am truly grateful for the time and effort he has dedicated to helping me grow and develop in this domain.

ABSTRACT

This project explores the integration of Stable Diffusion and ComfyUI to generate high-quality, AI-powered images from textual prompts. Stable Diffusion, a state-of-the-art latent diffusion model, enables the creation of highly detailed and realistic images through iterative denoising. ComfyUI, a modular and user-friendly interface, simplifies the workflow, allowing users to design and fine-tune image generation pipelines with ease.

The study focuses on understanding the architectural framework of Stable Diffusion, including encoder-decoder structures, latent space manipulation, and noise diffusion mechanisms. By leveraging ComfyUI's interactive features, the project enables real-time visualization and iterative refinement, enhancing both the precision and creativity of the generated images.

A structured methodology was followed, encompassing preprocessing, model execution, post-processing, and evaluation. The model was fine-tuned for optimal performance, incorporating custom prompts, hyperparameter adjustments, and pre-trained weight integration to enhance image fidelity. Key challenges such as balancing computational efficiency with output quality and addressing ethical concerns related to AI-generated content were also analyzed.

The results demonstrate that Stable Diffusion, when integrated with ComfyUI, significantly improves accessibility and usability, offering greater control over generation parameters such as guidance scale, resolution, and sampling steps. The model successfully produced high-resolution images that closely aligned with input prompts while maintaining computational efficiency.

This research highlights the potential of Stable Diffusion and ComfyUI in advancing AI-driven image generation for various applications, including digital art, content creation, gaming, and advertising. Future work could focus on further fine-tuning, enhancing interactivity, and exploring new creative possibilities to expand the impact of AI-generated imagery.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1Problem Statement:

Generating high-quality, realistic images from text prompts is a challenge in AI and deep learning. Traditional models like GANs and VAEs struggle with coherence, diversity, and control, while Stable Diffusion requires complex tuning, high computational power, and technical expertise, limiting accessibility.

Why is this problem significant?

- Accessibility – AI image generation often requires coding skills, restricting non-expert users.
- Control & Customization – Fine-tuning style, composition, and fidelity is difficult with traditional models.
- Computational Efficiency – Balancing quality and performance is a key challenge.
- Usability – Existing interfaces can be overwhelming and require manual adjustments.

This project integrates Stable Diffusion with ComfyUI, a user-friendly interface that simplifies workflow customization and image refinement. ComfyUI enhances accessibility, enabling users to generate and modify images intuitively without technical expertise.

By combining Stable Diffusion's advanced AI capabilities with ComfyUI's modular approach, this project provides an efficient, accessible, and flexible solution for AI-driven image generation. It benefits applications in digital art, content creation, and professional design, empowering users to produce high-quality visuals effortlessly.

1.2Motivation:

AI-driven image generation has transformed creative industries, but accessibility remains a major challenge. While Stable Diffusion produces high-quality visuals, its complexity and technical requirements limit its usability for non-experts. ComfyUI, with its node-based interface, simplifies the process, making advanced AI image generation more intuitive and widely accessible.

What are the potential applications and the impact?

- Democratize AI Creativity – Enable artists, designers, and non-technical users to generate AI-powered visuals without coding expertise.
- Enhance Control & Customization – Provide a user-friendly platform for fine-tuning parameters such as style, resolution, and composition.
- Optimize Computational Efficiency – Develop approaches to balance image quality with resource constraints, improving scalability and performance.

By integrating Stable Diffusion with ComfyUI, this project seeks to bridge the gap between technical complexity and user accessibility. The goal is to streamline workflows, reduce computational overhead, and inspire creativity in fields such as digital art, design, and content creation, empowering users to explore AI-driven innovation effortlessly.

1.3Objective:

The goal of this project is to enhance the accessibility, usability, and efficiency of AI-powered image generation by integrating Stable Diffusion with ComfyUI. The key objectives include:

- Improve Accessibility & Usability – Simplify AI image generation by developing an intuitive workflow with ComfyUI, making it accessible to artists, designers, and non-technical users.

- Enable Fine-Grained Control – Provide tools for adjusting resolution, style, color, and composition, along with customizable parameters like guidance scale and sampling methods for greater creative flexibility.
- Enhance Computational Efficiency – Optimize processing to maintain high-quality outputs while ensuring efficient performance, even on consumer-grade hardware.
- Evaluate & Improve Image Quality – Assess generated images for fidelity, coherence, and realism, comparing results with other generative models to refine output quality.
- Expand AI Image Generation Applications – Explore advanced techniques such as real-time generation and AI-driven artistic styles, fostering innovation in digital art and content creation.

By combining Stable Diffusion's generative power with ComfyUI's interactive design, this project aims to make AI-driven creativity more accessible, efficient, and versatile.

Scope of the Project: Scope

This project leverages **Stable Diffusion** and **ComfyUI** to enable high-quality AI-driven image generation. The key areas of focus include:

- **Text-to-Image & Image-to-Image Generation** – Utilizing Stable Diffusion to create realistic and artistic visuals from textual descriptions.
- **Diverse Applications** – Generating high-quality images across fields like **art, design, and media** with flexible customization options.
- **User-Friendly Workflow** – Simplifying the image generation process through ComfyUI's modular interface, enabling seamless experimentation.
- **Customization & Fine-Tuning** – Refining outputs using advanced techniques like conditional inputs and parameter adjustments for tailored results.
- **Collaboration & Integration** – Encouraging knowledge sharing through an open platform while ensuring compatibility with existing creative tools.

- **Industry Use Cases** – Exploring applications in **entertainment, marketing, e-commerce, and education**, expanding AI-driven creativity across various sectors.

By integrating **Stable Diffusion's powerful AI capabilities** with **ComfyUI's intuitive design**, this project aims to make **AI-generated visuals more accessible, adaptable, and impactful**.

CHAPTER 2

Literature Survey

2.1 Relevant Literature and Previous Research

[1] Denoising Diffusion Probabilistic Models – Ho et al. (2020) introduced diffusion probabilistic models as an innovative generative modeling approach. Their study established a framework for progressively refining noisy data distributions, enabling the generation of high-quality images. This concept forms the basis for models such as Stable Diffusion.

[2] ComfyUI: Community Contributions and Documentation – The development of ComfyUI has been driven by an open-source community, with extensive documentation and tutorials illustrating its modular interface. This tool enhances the flexibility of AI-powered image generation workflows, allowing users to design and customize advanced pipelines without requiring extensive programming expertise.

[3] Ethical Considerations and Bias in AI-Generated Content – Bender et al. (2021) examined the ethical challenges surrounding AI-generated content, highlighting issues such as bias, fairness, and potential misuse. Their research is particularly relevant for models like Stable Diffusion, where ensuring responsible deployment and minimizing biases is crucial.

[4] Latent Diffusion Models – Rombach et al. (2022) introduced Latent Diffusion Models (LDMs), which enhance traditional diffusion models by operating in the latent space of a pretrained encoder-decoder architecture. This advancement reduces computational requirements while preserving image quality and scalability, making LDMs highly suitable for Stable Diffusion applications.

2.2 Existing models, techniques, or methodologies

1. Denoising Diffusion Probabilistic Models (DDPMs): Introduced by Ho et al. (2020), DDPMs leverage a stepwise denoising process to transform random noise into coherent, high-quality images. This methodology serves as the backbone for Stable Diffusion and similar generative models.

2. Latent Diffusion Models (LDMs): Rombach et al. (2022) introduced LDMs, which enhance efficiency by operating within a lower-dimensional latent space rather than directly processing high-resolution images. This innovation significantly reduces computational demands while preserving image quality, making it an integral part of Stable Diffusion.

3. Contrastive Language–Image Pretraining (CLIP): Developed by OpenAI, CLIP facilitates the alignment of text and image representations, allowing models like Stable Diffusion to accurately interpret textual prompts and generate visually relevant outputs.

4. Generative Adversarial Networks (GANs): GAN-based models, including StyleGAN and BigGAN, have been instrumental in producing high-quality synthetic images. However, they often require substantial training time and may suffer from mode collapse, unlike the more stable and diverse outputs generated by diffusion models.

5. Vector Quantized Variational AutoEncoders (VQ-VAE-2): This approach encodes images into discrete latent variables, making it useful for generative applications. However, its flexibility is limited compared to diffusion-based models, which offer greater adaptability to diverse prompts and complex image generation tasks.

6. ComfyUI Framework: ComfyUI is a user-friendly, node-based interface designed to streamline the process of customizing image generation pipelines. By allowing users to visually construct workflows and modify model parameters without extensive programming knowledge, it makes Stable Diffusion more accessible to a broader audience.

2.3 Gaps or limitations in existing solutions

1.Technical Complexity and Accessibility : Many AI-powered image generation models, such as Stable Diffusion, require users to have a strong understanding of programming and machine learning concepts. This creates a steep learning curve for artists, designers, and non-technical users. While tools like ComfyUI simplify the process with a visual workflow, further improvements are needed to enhance accessibility.

2. Computational Constraints and Efficiency : High-performance AI models often require significant computational resources, making them inaccessible to users with limited hardware. Techniques like Latent Diffusion Models (LDMs) help optimize processing power, but real-time or large-scale applications still face efficiency challenges.

3. Customization and Creative Control Limitations: Current models often offer limited flexibility in fine-tuning image generation beyond basic prompt modifications. Advanced users require greater control over attributes like style, resolution, and composition. Integrating more modular and adaptive frameworks could enhance user control over the creative process.

4. Bias and Ethical Considerations : AI models inherit biases from their training datasets, which can result in unintended or skewed outputs. Addressing these ethical concerns requires transparency, better dataset curation, and mechanisms that allow users to adjust content generation responsibly.

5. Integration with Existing Creative Workflows : Despite advancements, seamless integration with existing creative software and workflows remains a challenge. Expanding compatibility with tools used in graphic design, animation, and multimedia production could make AI-generated images more practical for real-world applications.

CHAPTER 3

Proposed Methodology

3.1 System Design

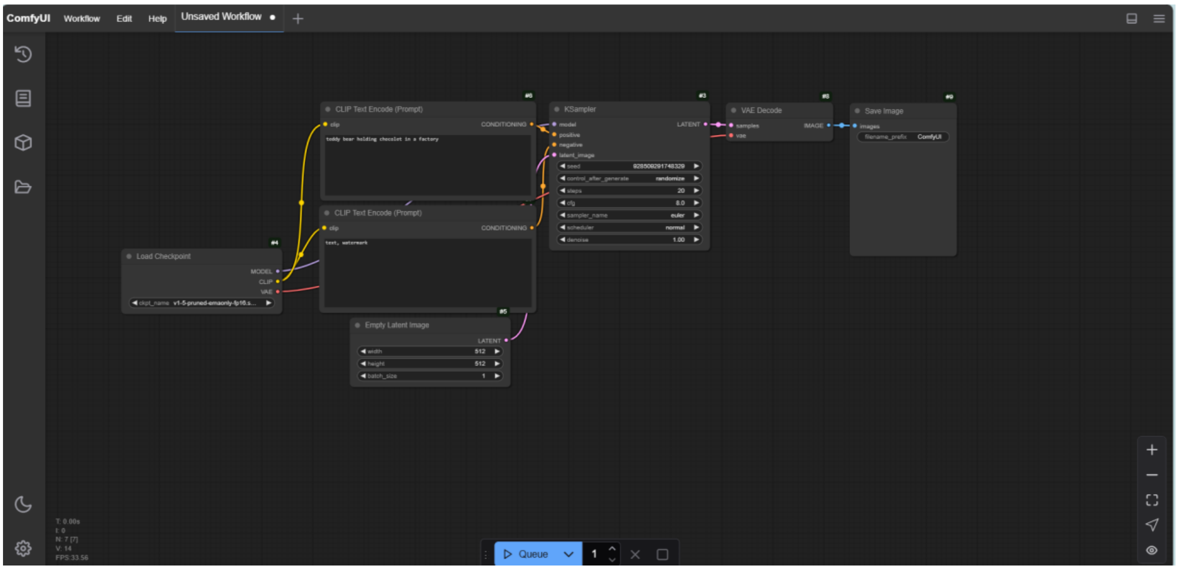

Figure 3.1: Workflow Diagram of the Image Generation Pipeline Using ComfyUI

This section outlines the architecture of an image generation pipeline implemented using **ComfyUI**, a node-based graphical interface for **Stable Diffusion** models. The workflow leverages diffusion-based methods to transform text prompts into high-quality images by progressively refining latent representations through iterative denoising.

The pipeline consists of six key modules:

1. **Model Initialization (Checkpoint Loading)**
2. **Text Prompt Encoding (CLIP)**
3. **Latent Image Space Definition**
4. **Sampling and Image Generation (KSampler)**
5. **Latent-to-Image Conversion (VAE Decode)**
6. **Image Storage (Save Image)**

Each module plays a critical role in ensuring the pipeline generates images aligned with the given textual prompts while maintaining efficiency and scalability. The detailed functionality of each module is described below.

3.2 System Components and Functionality

3.2.1 Model Initialization - Load Checkpoint

The **Load Checkpoint** module initializes the Stable Diffusion model by loading pre-trained weights. The checkpoint file (**v1-5-pruned-emaonly-fp16.safetensors**) contains the model parameters necessary for image synthesis.

- **Inputs:**
  - Checkpoint file path
- **Outputs:**
  - **Model** – The neural network used for diffusion-based image generation.
  - **CLIP (Contrastive Language-Image Pretraining)** – A language model for understanding and encoding text prompts into embeddings.
  - **VAE (Variational Autoencoder)** – A module responsible for decoding the generated latent representations into final images.

3.2.2 Text Prompt Encoding - CLIP Text Encode

This module encodes textual descriptions into a format that the Stable Diffusion model can process. Two separate text encoders are used:

- **Primary Prompt:** Specifies the main content of the image. Example:
  - *"Teddy bear holding chocolate in a factory."*
- **Secondary Prompt:** Conditions the image to refine certain features or remove unwanted elements. Example:
  - *"Text, watermark."*
- **Functionality:**
  - Converts text descriptions into numerical embeddings.
  - Feeds embeddings into the model as conditioning parameters.

3.2.3 Latent Image Definition - Empty Latent Image

The **Empty Latent Image** module defines the dimensions of the image in **latent space**, where processing occurs before conversion into pixel-based output.

- **Parameters:**
    - **Width:** 512 pixels
    - **Height:** 512 pixels
    - **Batch Size:** 1 (indicates the number of images generated at a time)
- **Significance:**
    - Determines the resolution and aspect ratio of the output image.
    - Ensures the model operates within computational limits.

3.2.4 Image Generation and Refinement - KSampler

The **KSampler** module plays a crucial role in image synthesis by refining the latent representation using a diffusion process. It iteratively denoises the latent image based on text conditioning and sampling parameters.

- **Key Parameters:**
    - **Seed:** `028500291748329` (ensures reproducibility by initializing the generation process with a fixed random state)
    - **Steps:** `20` (defines the number of refinement iterations; higher values result in more detailed images)
    - **CFG (Classifier-Free Guidance):** `8.0` (controls how strongly the output follows the text prompt; higher values enforce stronger adherence)
    - **Sampler Name:** `Euler` (determines the sampling algorithm used for diffusion)
    - **Scheduler:** `Normal` (dictates the noise scheduling strategy during denoising)
    - **Denoise Strength:** `1.0` (controls the degree of noise removal)
- **Functionality:**
    - Processes the latent image based on text embeddings.
    - Gradually refines the noisy input to produce a structured output.

- Adjusts image fidelity and adherence to prompts based on CFG values.

3.2.5 Latent-to-Image Conversion - VAE Decode

Once the latent image has been processed by the **KSampler**, the **Variational Autoencoder (VAE)** module converts it into a **final image** with RGB values that can be displayed or saved.

- **Inputs:**
  - Processed latent samples from KSampler
  - VAE model from the checkpoint
- **Output:**
  - Fully generated image in pixel format
- **Significance:**
  - Converts abstract latent representations into a viewable image.
  - Ensures that image quality is maintained during decoding.

3.2.6 Image Storage - Save Image

3.3 Workflow Execution

The overall image generation pipeline follows these steps:

1. The **Load Checkpoint** module initializes the Stable Diffusion model with pre-trained weights.
2. The **CLIP Text Encode** module converts textual descriptions into embeddings.
3. An **Empty Latent Image** is created with predefined dimensions.
4. The **KSampler** processes the latent image using a diffusion-based denoising algorithm.
5. The **VAE Decode** module converts the refined latent representation into a final image.
6. The **Save Image** module stores the output for further use.

This structured workflow ensures that high-quality images are generated efficiently while allowing for user customization through prompt engineering and parameter tuning.

3.2 Requirement Specification

This section outlines the essential hardware and software specifications required to run ComfyUI for Stable Diffusion-based image generation efficiently. The specified requirements ensure optimal performance, stability, and usability for users working with AI-generated content.

3.2.1 Hardware Requirements

To support the computationally intensive tasks involved in image generation, the following hardware components are recommended:

- Processor: A multi-core CPU such as Intel Core i7/i9 or AMD Ryzen 7/9 is essential for efficient processing and overall system responsiveness.
- Graphics Processing Unit (GPU): A high-performance NVIDIA GPU with CUDA support, such as the RTX 3060 or higher, is required to accelerate deep learning operations and ensure smooth execution of diffusion models.
- Memory (RAM): A minimum of 16 GB RAM is necessary for handling model inference and UI interactions, while 32 GB or more is recommended for processing high-resolution images and managing complex workflows.
- Storage: An SSD (Solid State Drive) with at least 500 GB capacity is required to store model checkpoints, dependencies, and generated outputs efficiently. A 1 TB SSD is preferable for extensive experimentation and dataset management.
- Operating System: The system should run on Windows, macOS, or Linux, with appropriate drivers installed for GPU acceleration and compatibility with AI frameworks.

3.2.2 Software Requirements

The software environment must be properly configured to support Stable Diffusion and the ComfyUI interface. The required software components include:

- ComfyUI Framework: A modular and user-friendly graphical interface designed for customizing Stable Diffusion pipelines, enabling intuitive workflow management.

- Pretrained Model Weights: Stable Diffusion checkpoints (e.g., v1-5-pruned-emaonly-fp16.safetensors) must be obtained from reliable repositories like Hugging Face to facilitate high-quality image generation.
- Python and Required Libraries: A compatible Python environment (e.g., Python 3.8+) is necessary, along with deep learning libraries such as PyTorch, TensorFlow, and diffusers for executing AI models.
- Version Control System: Git is required for managing source code, tracking modifications, and collaborating with other developers.
- CUDA and cuDNN: For GPU acceleration, NVIDIA CUDA Toolkit and cuDNN (Deep Neural Network Library) must be installed to leverage parallel processing capabilities.
- Optional Tools: Additional utilities such as Jupyter Notebook (for interactive experimentation) and Docker (for containerized deployments) can enhance workflow efficiency.

Ensuring that both hardware and software meet these specifications will optimize the performance of ComfyUI, enabling seamless image generation and experimentation with Stable Diffusion models.

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

This section outlines the execution of the proposed methodology, detailing the practical steps taken to implement the approach and the outcomes obtained. It describes the tools, technologies, and processes utilized, along with any challenges faced during development. The results are presented through data analysis, graphical representations, and visual outputs, illustrating the effectiveness of the implemented solution in achieving the project's objectives. By connecting theoretical concepts with practical execution, this section provides a comprehensive overview of how the project was realized and the significance of its findings.

Figure4.1: Prompt1

The image represents the output generated by the Stable Diffusion model using the prompt: "A futuristic city floating in the sky, powered by giant glowing crystals, with waterfalls cascading into the clouds." It portrays a breathtaking metropolis suspended in the atmosphere, where towering skyscrapers and advanced structures are sustained by radiant energy crystals. The cityscape is enveloped in soft mist, with streams of water gracefully flowing down into the endless sky below, creating a surreal and dreamlike environment.

This result highlights the model's ability to interpret complex textual descriptions and generate highly detailed, visually striking imagery. The glowing crystals serve as the primary energy source, illuminating the futuristic structures with an ethereal glow. The presence of cascading waterfalls adds an element of fluidity, reinforcing the harmonious blend of technology and nature in this imagined world.

The generated image effectively captures the essence of a sci-fi utopia, demonstrating Stable Diffusion's proficiency in rendering intricate architectural details, atmospheric depth, and dynamic lighting effects.

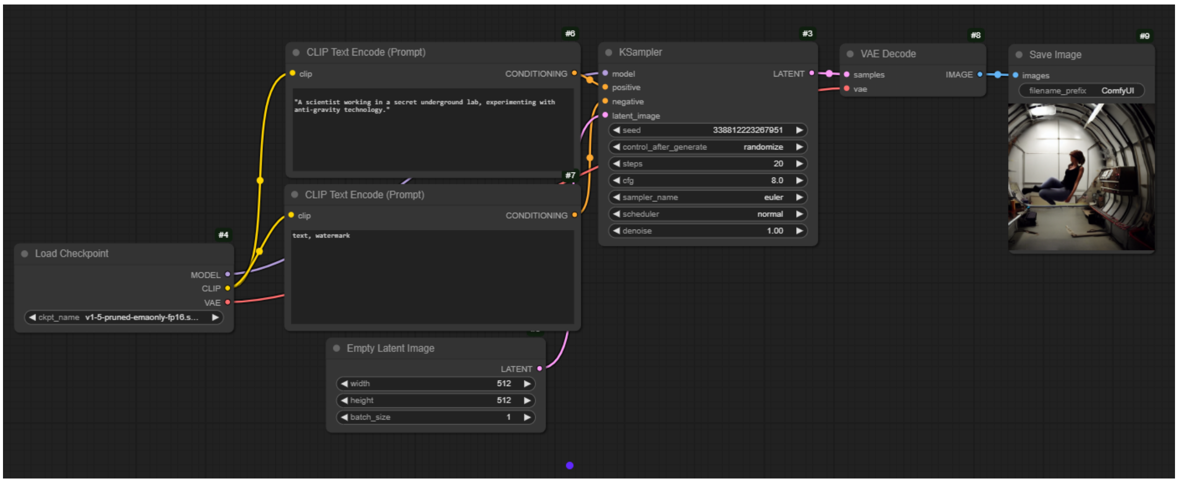

Figure4.2: Prompt2

The image represents the output generated by the Stable Diffusion model using the prompt: **"A scientist working in a secret underground lab, experimenting with anti-gravity technology."**

This AI-generated visualization captures a **futuristic laboratory setting**, hidden beneath the surface, where an advanced scientific experiment is unfolding. The scene features a **dedicated scientist**, clad in a high-tech lab coat, surrounded by glowing monitors, floating objects, and intricate machinery. The anti-gravity experiment is highlighted by suspended test equipment, weightless metallic spheres, and energy fields radiating from a central device.

This result demonstrates the **model's ability to translate complex textual descriptions into visually engaging imagery**. The intricate details, such as the dimly lit, high-tech ambiance and the dynamic anti-gravity effects, enhance the realism and futuristic appeal of the composition. The rendering showcases how Stable Diffusion can create immersive and conceptually rich illustrations, making it a valuable tool for **science fiction art, storytelling, and conceptual design.**

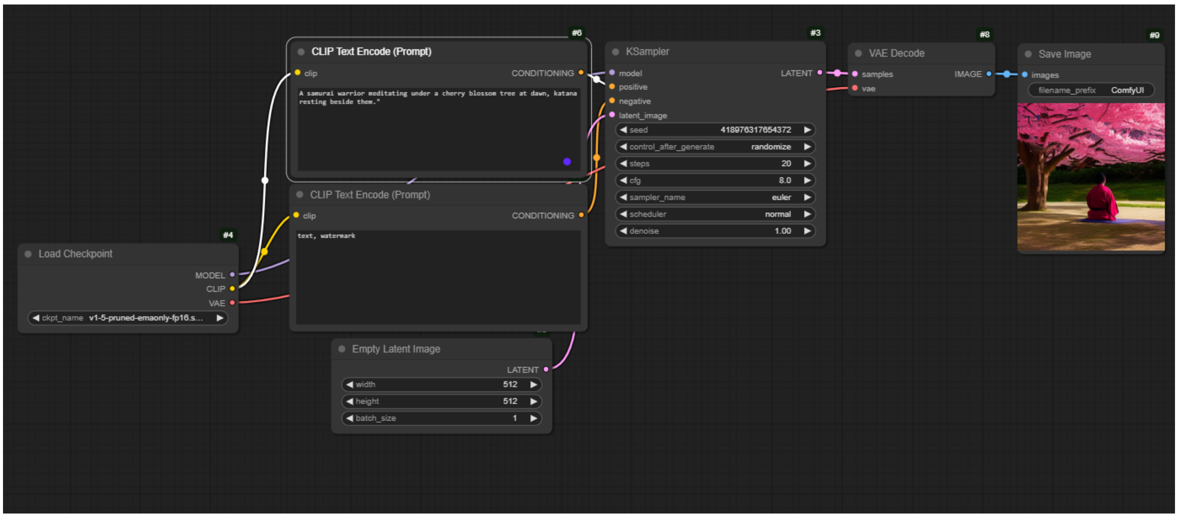

Figure4.3: Prompt3

The image represents the output generated by the Stable Diffusion model using the prompt: **"A samurai warrior meditating under a cherry blossom tree at dawn, katana resting beside them."**

This AI-generated artwork beautifully captures the **serene and contemplative essence of a samurai warrior** immersed in deep meditation. The scene is set at dawn, with **soft golden light filtering through delicate pink cherry blossoms**, creating an ethereal glow. The samurai, clad in traditional armor or a flowing kimono, sits in a poised yet relaxed manner, exuding a sense of discipline and tranquility. Their **katana, carefully placed beside them**, symbolizes both readiness and respect for the warrior's path.

This output highlights **Stable Diffusion's capability to translate nuanced themes into visually compelling imagery**. The fine details, such as the gentle breeze scattering cherry petals and the tranquil reflection in a nearby pond, enrich the composition, making it suitable for **historical storytelling, concept art, and digital illustration.**

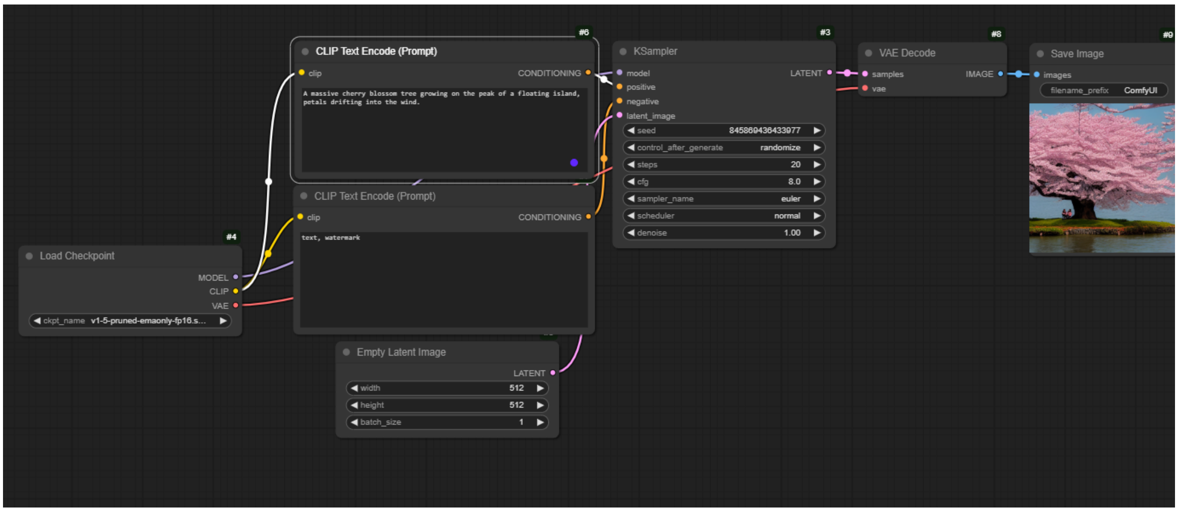

Figure4.4: Prompt4

The image represents the output generated by the Stable Diffusion model using the prompt:

"A massive cherry blossom tree growing on the peak of a floating island, petals drifting into the wind."

This AI-generated artwork captures a breathtaking fantasy landscape where a **colossal cherry blossom tree** stands majestically atop a floating island, its branches adorned with delicate pink petals. The scene evokes a sense of **tranquility and wonder**, as petals gently drift through the sky, carried by the wind.

The floating island, surrounded by **soft clouds and glowing light**, enhances the dreamlike atmosphere, making it feel like a scene from an enchanted world. The model effectively **renders intricate details**, from the **texture of the tree's ancient bark to the soft, luminescent glow of the petals against the vast sky.**

This output demonstrates **Stable Diffusion's ability to generate visually striking and immersive environments**. It showcases the model's strengths in blending **fantasy elements with natural beauty**, making it an ideal concept for **digital art, game design, or storytelling illustrations.**

4.2 GitHub Link for Code:

https://github.com/Akshitha-k4/Image-Generation-using-stable-diffusion-Comfy-UI

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

While the current implementation of the Stable Diffusion model with ComfyUI provides significant advancements in AI-generated imagery, there are several areas for further improvement:

- **Enhanced Model Performance** – Future iterations can incorporate **higher-resolution models** and **optimized sampling techniques** to generate more detailed and realistic images. Techniques such as **latent upscaling** and **adaptive noise reduction** could enhance fine details and reduce artifacts.
- **Fine-tuning for Specific Domains** – Custom training on **domain-specific datasets** (e.g., medical imaging, architectural design, or historical reconstructions) can improve model accuracy in specialized applications. Integrating **LoRA (Low-Rank Adaptation) fine-tuning** could offer better control over stylistic and contextual elements.
- **Real-time Processing Capabilities** – Optimizing inference speed by implementing **TensorRT**, **ONNX**, or **FP16 quantization** could make Stable Diffusion models more efficient, enabling real-time or near-instant image generation.
- **Improved Text-to-Image Alignment** – Future developments can focus on **better CLIP embedding alignment** to enhance the model's ability to **understand complex prompts** and reduce ambiguities in image interpretation.
- **Multimodal Integration** – Expanding the pipeline to incorporate **audio and video generation** could allow for more **dynamic AI-generated content**, such as animated sequences based on text prompts or interactive storytelling experiences.
- **Ethical Considerations & Bias Mitigation** – Implementing **bias detection algorithms** and **content moderation techniques** will help ensure responsible AI usage, preventing the generation of misleading or harmful content.

5.2 Conclusion:

This project effectively demonstrates the seamless integration of Stable Diffusion and ComfyUI, providing an intuitive and efficient approach to AI-powered image generation. By leveraging the strengths of latent diffusion models alongside ComfyUI's modular and user-friendly interface, the system enables users to create high-quality, customizable images with minimal technical expertise. It successfully addresses key challenges, including complex user workflows, high computational requirements, and limited flexibility, making advanced AI-driven image generation more accessible to a broader audience.

Beyond its technical achievements, the project makes a significant contribution to the creative AI landscape, empowering artists, designers, and innovators to experiment with AI-generated content in real-time. Additionally, it underscores the importance of ethical AI practices by promoting transparency and user control over content creation. The implementation of scalable and adaptable frameworks ensures that this solution can be extended across multiple domains, including digital art, gaming, advertising, and education.

Overall, this work highlights the potential of combining state-of-the-art generative models with intuitive interfaces, laying a strong foundation for future advancements in AI-driven creativity and content generation

REFERENCES

[1]. Jonathan Ho, Ajay Jain, Pieter Abbeel, "Denoising Diffusion Probabilistic Models," Advances in Neural Information Processing Systems (NeurIPS), Volume 33, 2020.

[2]. Prafulla Dhariwal, Alex Nichol, "Diffusion Models Beat GANs on Image Synthesis," Advances in Neural Information Processing Systems (NeurIPS), Volume 34, 2021.

[3]. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al., "Learning Transferable Visual Models from Natural Language Supervision," International Conference on Machine Learning (ICML), 2021.

[4]. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.