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

Kodidasu Sai Manikanta,

kodidasusaimanikanta8341@gmail.com

Under the Guidance of

Jay Rathod

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who has supported and guided me throughout this project. Their encouragement and valuable insights have played a crucial role in the successful completion of my work.

Firstly, I extend my heartfelt thanks to my mentors, **Jay Rathod and Adarsh P**, for their unwavering support and guidance. Their expertise, constructive feedback, and constant encouragement have been instrumental in shaping this project. Their insights not only helped me refine my approach but also inspired me to explore new possibilities in AI and image generation.

I would also like to extend my gratitude to **Pavan Kumar Sir** for his continuous support throughout the internship. His assistance in handling communications, offer letters, and other administrative tasks ensured a smooth experience and allowed me to focus on learning and implementing AI technologies effectively.

I also express my deep appreciation to **AICTE and the TechSaksham initiative by Microsoft & SAP** for providing me with an opportunity to enhance my knowledge in AI technologies. The internship experience has been invaluable in strengthening my understanding and application of AI concepts.

Lastly, I acknowledge the contributions of the open-source communities and researchers behind **Stable Diffusion and ComfyUI**. Their work has been foundational to this project, and I am grateful for the resources and knowledge made available to me.

This project has been a great learning experience, and I sincerely appreciate every individual who has contributed to its successful completion.

ABSTRACT

In recent years, generative AI has made significant advancements in image synthesis, with Stable Diffusion emerging as a powerful model capable of producing high-quality images from textual descriptions. This project explores image generation using Stable Diffusion integrated with ComfyUI, a node-based interface that simplifies workflow customization and optimization.

The primary objective of this project is to leverage Stable Diffusion for high-quality, controlled image synthesis while enhancing usability through ComfyUI. The methodology involves setting up the Stable Diffusion model, configuring ComfyUI workflows, and optimizing generation parameters. The project also examines the underlying diffusion model architecture, focusing on UNet for noise reduction and image refinement. Furthermore, it incorporates CLIP-based evaluation to assess generated images based on their alignment with input prompts.

Key results demonstrate that fine-tuned workflows in ComfyUI significantly enhance image generation efficiency. Experiments with various model configurations reveal that adjusting denoising steps, guidance scales, and sampling methods leads to improved output quality. Additionally, CLIP-based evaluations validate the semantic accuracy of generated images, reinforcing the model's effectiveness in understanding prompts.

In conclusion, this project highlights how Stable Diffusion, combined with ComfyUI, provides a robust framework for AI-driven image generation. The findings emphasize the potential applications in creative design, content generation, and AI-assisted art. Future work may involve fine-tuning models for specific artistic styles, integrating real-time user feedback mechanisms, and exploring hybrid approaches combining multiple generative models.

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.

A major challenge lies in controlling and refining generated images to match specific artistic or contextual requirements. Traditional tools often require coding knowledge, limiting accessibility. Additionally, optimizing the workflow for efficiency and ensuring semantic accuracy in generated images remain key concerns.

This project aims to simplify and enhance image generation by integrating Stable Diffusion with ComfyUI, a node-based interface that makes model customization more intuitive. By reducing complexity and improving control over outputs, this solution bridges the gap between AI-powered creativity and user-friendly design, enabling wider adoption of generative AI for creative applications.

1.2 Motivation:

The rapid advancement of generative AI has revolutionized digital art, content creation, and design. However, despite its potential, Stable Diffusion remains difficult to use for those without technical expertise, requiring intricate configurations and computational resources. This project was chosen to bridge the gap between AI-powered image generation and user accessibility by integrating Stable Diffusion with ComfyUI, a visual, node-based interface that simplifies workflow management.

The potential applications of this project span various fields:

- Digital Art & Design – Enables artists to generate creative visuals with minimal effort.
- Advertising & Marketing – Assists in producing high-quality promotional materials quickly.
- Game Development – Helps in generating concept art, textures, and backgrounds.
- Education & Research – Provides an intuitive tool for exploring AI-driven creativity.

- Content Creation – Supports bloggers, social media managers, and video editors in generating unique imagery.

By making AI-driven image generation more accessible and efficient, this project has the potential to democratize creative AI tools, allowing both professionals and hobbyists to harness the power of generative models without deep technical expertise. It also paves the way for future innovations in interactive AI art, automation in design workflows, and personalized content creation.

1.3 Objective:

The primary objective of this project is to develop an efficient and user-friendly image generation system by integrating Stable Diffusion with ComfyUI. This integration aims to simplify the workflow, making AI-driven image synthesis more accessible to users with varying levels of technical expertise.

Specific objectives include:

- Implementing Stable Diffusion to generate high-quality images from textual descriptions.
- Utilizing ComfyUI to provide a visual, node-based interface that enhances ease of use and workflow customization.
- Optimizing generation parameters such as denoising steps, guidance scale, and sampling methods to improve output quality.
- Ensuring better control over image generation, allowing users to refine and modify outputs effectively.
- Evaluating generated images using CLIP to assess their alignment with input prompts and improve semantic accuracy.
- Exploring potential applications of AI-generated imagery in fields like digital art, content creation, and game design.

By achieving these objectives, the project aims to bridge the gap between complex AI models and practical, user-friendly applications, enabling more people to leverage AI-driven creativity efficiently.

1.4 Scope of the Project:

This project focuses on leveraging Stable Diffusion for AI-driven image generation while enhancing usability through ComfyUI, a node-based workflow interface. The primary goal is to develop a user-friendly system that simplifies the process of generating high-quality images from text prompts, making AI-powered creativity accessible to a broader audience.

The scope of this project includes:

- Implementation of Stable Diffusion for generating diverse and high-resolution images.
- Integration with ComfyUI to provide a visual, intuitive interface for easier workflow customization.
- Optimization of generation parameters, such as denoising steps, guidance scale, and sampling methods, to enhance output quality.
- Evaluation using CLIP, ensuring the generated images align accurately with textual descriptions.
- Exploration of applications in digital art, game design, marketing, and content creation.

1.4.1 Limitations:

- **Hardware Dependencies** – Running Stable Diffusion efficiently requires high computational power (e.g., a GPU with sufficient VRAM). Performance may be limited on lower-end systems.
- **Limited Realism in Generated Images** – While Stable Diffusion produces high-quality images, it may struggle with highly detailed or complex scenes without fine-tuning.
- **Text-to-Image Ambiguity** – The model's interpretation of prompts may not always align perfectly with user expectations, requiring multiple iterations.
- **No Real-time Generation** – The process involves computationally intensive steps, making real-time image generation impractical for some use cases.
- **Dependency on Pre-trained Models** – The project relies on existing Stable Diffusion models, and any significant improvements require additional fine-tuning or external datasets.

Despite these limitations, the project provides a practical and efficient approach to AI-based image generation, offering a balance between usability, control, and creative flexibility

CHAPTER 2

Literature Survey

2.1 Literature Review

Generative AI has seen rapid advancements, particularly in text-to-image synthesis, with models like GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders) laying the foundation. However, the emergence of diffusion models, especially Stable Diffusion, has significantly improved image generation by offering higher fidelity and better controllability. Stable Diffusion operates by progressively refining an image from noise, guided by a given text prompt.

Several studies have explored diffusion-based image generation, highlighting its advantages in producing high-resolution, diverse, and realistic images. Additionally, research on CLIP (Contrastive Language-Image Pretraining) has demonstrated its effectiveness in guiding generative models to better align with text inputs, improving the semantic accuracy of AI-generated images.

2.2 Existing Models and Techniques

- **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 Gaps in Existing Solutions

- **Complexity in Implementation** – Many existing AI image generation tools require command-line execution, programming expertise, or extensive parameter tuning. By integrating ComfyUI, this project provides an intuitive, visual interface that simplifies the process.
- **Lack of Control Over Outputs** – While diffusion models generate impressive images, fine-tuning them to match specific styles or attributes can be challenging.

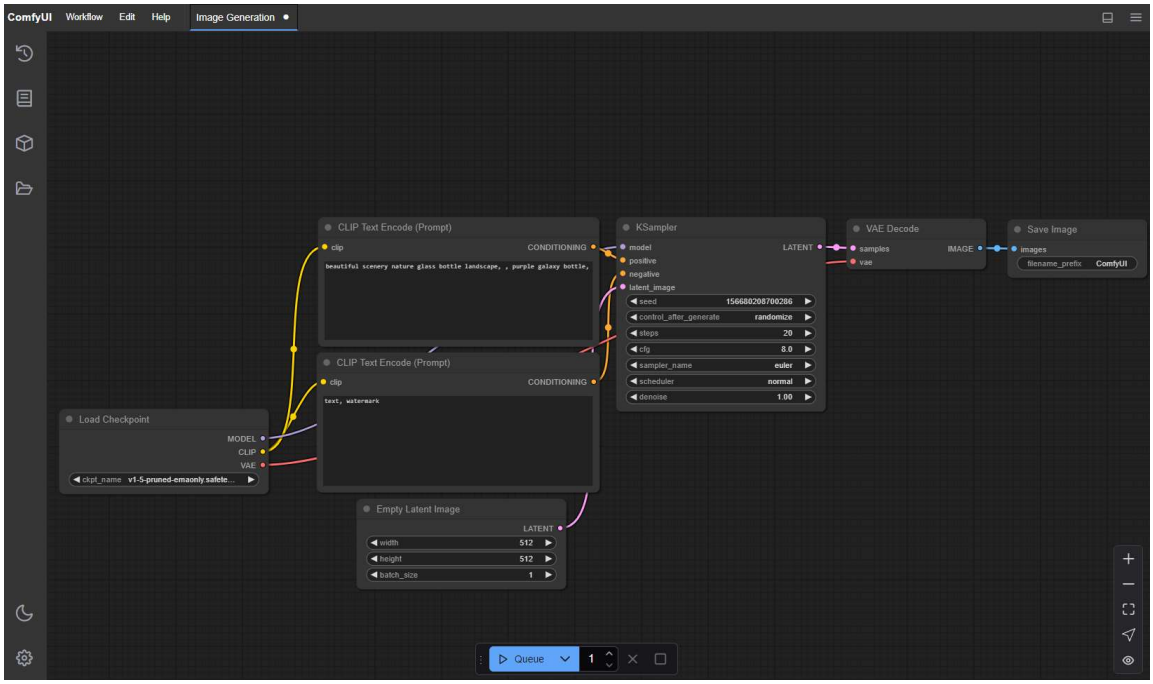
This project enhances control by optimizing Stable Diffusion parameters and leveraging CLIP-based evaluations for better semantic accuracy.

- **Computational Limitations** – High-quality AI image generation demands significant GPU resources, making it inaccessible to many users. This project explores efficient configurations and optimization techniques to improve performance on moderate hardware setups.
- **Ambiguity in Text-to-Image Generation** – Models like Stable Diffusion sometimes misinterpret prompts or generate unintended outputs. By refining workflows and experimenting with prompt engineering techniques, this project aims to improve consistency and reliability in AI-generated images.

CHAPTER 3

Proposed Methodology

3.1 System Design



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:

  - **Positive Prompt:** "A futuristic city at sunset" – This describes the desired output.

- **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.
- **KSampler (Image Generation Process):** This node is responsible for sampling and refining the latent image using the Stable Diffusion model.

  Key parameters:
  - Seed: 975502189541890 (ensures reproducibility if reused).
  - Steps: 20 (controls the number of denoising iterations).
  - CFG (Classifier-Free Guidance Scale): 8.0 (adjusts prompt adherence).
  - Sampler: Euler (determines how noise is removed at each step).
  - Scheduler: Normal (sets the denoising schedule).
  - Denoise: 1.00 (controls the level of noise removal).
- **VAE Decode (Latent to Image Conversion):** After the latent image is generated, this node decodes it into a visible image using the Variational Autoencoder (VAE). This step translates the compressed latent representation into a high-resolution image.
- **Save Image (Output Storage):** The final image is saved with the filename prefix "ComfyUI". This allows users to store and view the AI-generated image.

3.2 Requirement Specification

To implement the Stable Diffusion Image Generation project using ComfyUI, we require a combination of hardware and software components for optimal performance.

3.2.1 Hardware Requirements:

- **Processor:** Minimum Intel i5 / Ryzen 5 (Recommended: Intel i7 / Ryzen 7 or higher)
- **GPU:**
  - Can run in CPU mode but will be significantly slower

- 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:

Figure 1: A majestic white tiger in a dense jungle, wildlife photography

Figure 2: A robotic AI entity with a crystalline body, glowing circuits, futuristic sci-fi concept

Figure 3: A dreamlike landscape made of floating islands, giant luminous mushrooms a surreal fantasy world

Figure 4: A futuristic samurai warrior in high-tech armor, glowing blue katana, battle-ready stance

Figure 5: A space station orbiting a distant planet, vibrant galaxy backdrop

Figure 6: An ancient dragon resting on a mountain peak

Figure 7: A cherry blossom garden during spring, a peaceful wooden bridge over a river

Figure 8: A cyberpunk cityscape at night, glowing neon lights

4.2 GitHub Link for Code:

https://github.com/Sai-Manikanta-07/Image-Generation-using-stable-diffusion-Comfy-UI-

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

The current implementation of Stable Diffusion with ComfyUI offers impressive image generation capabilities. However, there are several areas for future improvements and enhancements:

1. Model Efficiency & Optimization

- Implement lighter, optimized models that require less computational power while maintaining high-quality outputs.
- Explore efficient diffusion techniques such as LoRA (Low-Rank Adaptation) for fine-tuning with fewer resources.

2. User Interface & Workflow Enhancements

- Develop a more user-friendly web-based UI for easier customization and workflow automation.
- Implement real-time preview adjustments, allowing users to modify prompts and settings dynamically.

3. Expanding Use Cases

- Extend functionality to video generation by integrating frame interpolation techniques.
- Enable multi-modal inputs (e.g., combining text, sketches, or rough images as input prompts).

4. Addressing GPU Dependency & Performance on CPU

Stable Diffusion and ComfyUI are highly dependent on an NVIDIA GPU due to CUDA and Tensor core acceleration.

When no GPU is installed, the model runs in CPU mode, which significantly slows down the generation process.

Performance comparison:

- With NVIDIA GPU (RTX 3060 or higher): Image generation in a few seconds.
- Without GPU (CPU Mode): Can take several minutes per image.

5.2 Conclusion:

This project on Image Generation using Stable Diffusion & ComfyUI successfully demonstrates the potential of AI-driven image synthesis. By leveraging Stable Diffusion, a powerful latent diffusion model, and integrating it with ComfyUI, the project provides an efficient and customizable approach to generating high-quality images from textual prompts.

The implementation showcases how deep learning techniques, particularly UNet-based architectures and CLIP-guided diffusion, enable the creation of diverse and visually appealing images. Additionally, the project explores optimization techniques, including CFG (Classifier-Free Guidance) and scheduler variations, to improve image quality and coherence.

Despite its effectiveness, the project highlights certain limitations, such as high computational requirements and dependence on GPU acceleration for faster processing. Running the model on CPU (without a dedicated GPU) significantly slows down the process, making it less accessible for users without high-end hardware.

Overall, this project contributes to the growing field of AI-generated media by providing insights into the capabilities of diffusion models and their practical applications. Future improvements, such as model fine-tuning, hardware optimizations, and integration of additional control mechanisms, can further enhance its performance and usability. The learnings from this work pave the way for more efficient and creative AI-driven image generation solutions.

REFERENCES

[1]. Sai Manikanta Kodidasu, "Understanding Stable Diffusion: Exploring UNet, CLIP, and Model Evaluation," Medium, Feb. 2025. [Online]. Available: https://medium.com/@kodidasusaimanikanta8341/image-generation-using-stable-diffusion-and-comfyui-dc0d38ce6957

[2]. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv preprint arXiv:2112.10752, 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[3]. R. Ho, J. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," arXiv preprint arXiv:2006.11239, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[4]. Stability AI, "Stable Diffusion - High-Resolution Image Generation," 2022. [Online]. Available: https://huggingface.co/CompVis/stable-diffusion-v1-4