Image Generation using Stable diffusion and ComfyUI

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

Charantej Manchikanti, 214g1a3316@srit.ac.in

Under the Guidance of

Jay Rathod

ACKNOWLEDGEMENT

First and foremost, we are incredibly grateful to our supervisor, **Jay Rathod**, for being an exceptional mentor and guide. His insights, patience, and constant encouragement have been invaluable. Whether it was offering constructive feedback, pushing us to think beyond our limits, or simply motivating us during challenging times, his support has been a driving force behind the successful completion of this project. Working under his guidance has been an absolute privilege, and the lessons we have learned from him will stay with us far beyond this project.

We also want to extend our heartfelt thanks to our faculty members, friends, and family, whose support and encouragement kept us going. Their valuable advice and willingness to help at every step made a world of difference.

Lastly, we deeply appreciate all the resources, discussions, and individuals who contributed in any way, big or small. This project wouldn’t have been possible without their help, and we are truly grateful.

ABSTRACT

Image Generation using **Stable Diffusion and ComfyUI** explores AI-driven creativity by generating high-quality, lifelike images from textual descriptions. The project integrates **Stable Diffusion**, a deep learning text-to-image model, with **ComfyUI**, a node-based workflow tool, to enhance usability and flexibility.

The **problem statement** focuses on making AI-generated images more accessible and controllable, as existing methods often require extensive coding knowledge and manual fine-tuning. The goal is to simplify image generation while maintaining high customizability and quality.

The **methodology** follows a modular, visual approach:

1. **Model Selection & Setup** – Implementing Stable Diffusion with pre-trained weights.
2. **ComfyUI Workflow Design** – Creating a node-based interface for intuitive control.
3. **Prompt Engineering & Customization** – Refining outputs using text prompts, sampling techniques, and image conditioning.
4. **Performance Evaluation** – Assessing coherence, realism, and prompt adherence.

Key results demonstrate that ComfyUI simplifies workflow customization, offering a more intuitive alternative to traditional command-line or API-based implementations. It enhances creative control while reducing complexity for users.

In **conclusion**, this project bridges the gap between AI-generated art and user accessibility by combining **Stable Diffusion's generative power with ComfyUI's modular design**. This approach empowers both beginners and experts to experiment with AI-driven creativity seamlessly. Future improvements may focus on real-time efficiency and enhanced model adaptability.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Problem Statement:

AI-generated images are becoming a powerful tool in creative fields like digital art, design, gaming, and marketing. However, most existing solutions require users to have advanced coding skills and manually adjust multiple parameters to get the desired output. This makes it difficult for beginners and non-technical users to fully explore the potential of AI-powered image generation.

The challenge lies in making these tools more accessible while still allowing for customization and creative control. Stable Diffusion, one of the most popular AI image-generation models, produces stunning visuals from text prompts but often requires complex configurations. Meanwhile, traditional tools rely heavily on command-line interactions, making them intimidating for new users. The significance of this problem is clear—without an intuitive and easy-to-use interface, many creative individuals miss out on the benefits of AI in their work.

1.2 Motivation:

This project was inspired by the growing need for a user-friendly, flexible, and powerful AI image generation tool. Instead of requiring technical expertise, we wanted to build a system that allows users to visually design their workflows while maintaining the high-quality results of Stable Diffusion.

By integrating Stable Diffusion with ComfyUI, a node-based workflow tool, we aim to make AI-generated art more approachable. ComfyUI offers a drag-and-drop system where users can easily modify and experiment with their prompts and settings, making AI creativity more intuitive and enjoyable.

This has real-world applications in multiple industries:

- Digital artists and designers can create unique visuals without deep programming knowledge.
- Game developers can use AI to generate concept art or textures.
- Marketing teams can produce high-quality advertisements quickly.
- Content creators can experiment with AI-generated graphics for social media or storytelling.

- The impact of this project extends beyond just making AI tools easier to use—it empowers more people to explore AI-driven creativity.

1.3 Objective:

The main goal of this project is to make AI-powered image generation more accessible and customizable. To achieve this, we focus on:

- Creating a seamless workflow that integrates Stable Diffusion with ComfyUI.
- Developing an intuitive, drag-and-drop interface for easy customization.
- Allowing users to fine-tune parameters without requiring advanced technical skills.
- Evaluating the system's usability and efficiency compared to traditional AI image generation methods.
- Enhancing creative control through better prompt engineering and modular design.
- Ultimately, we want to build a tool that both beginners and experts can use to experiment with AI-generated images in a more engaging and efficient way.

1.4 Scope of the Project:

This project is designed to enhance the usability of Stable Diffusion rather than creating a new AI model from scratch. Our focus is on making the process smoother and more interactive using ComfyUI's visual workflow system.

The project has some limitations, including performance variations based on hardware and slower real-time processing due to the computational demands of AI models. While customization is extensive, it still relies on the capabilities of the underlying Stable Diffusion model. Despite these challenges, the goal is to make AI-generated creativity more accessible and user-friendly.

CHAPTER 2

Literature Survey

2.1 Review relevant literature or previous work in this domain.

The field of image and video generation has evolved significantly with models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models. GANs, introduced by Goodfellow et al. (2014), are powerful but suffer from unstable training and mode collapse. VAEs, proposed by Kingma and Welling (2013), are more stable but often produce blurry images. Diffusion models, particularly Rombach et al.'s (2022) work on Latent Diffusion Models, have emerged as a robust alternative, offering stable training and high-quality outputs by iteratively denoising images from noise.

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

- Stable Diffusion: This model operates in a latent space, reducing computational complexity while maintaining image quality. It uses a UNet architecture for iterative denoising, guided by a noise schedule. Conditioning mechanisms, such as textual prompts and time embeddings, enable controlled image generation based on specific descriptions.
- Stable Video Diffusion: Extending stable diffusion to videos, this model incorporates temporal convolutions to handle temporal consistency. It is trained on a large dataset of video clips, annotated using tools like CoCa and V-BLIP, and refined using optical flow analysis and CLIP scores for quality.
- Evaluation Metrics: CLIP (Contrastive Language-Image Pretraining) is used to assess alignment between text and images, while directional CLIP loss evaluates rotational adjustments between reference and generated outputs.

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

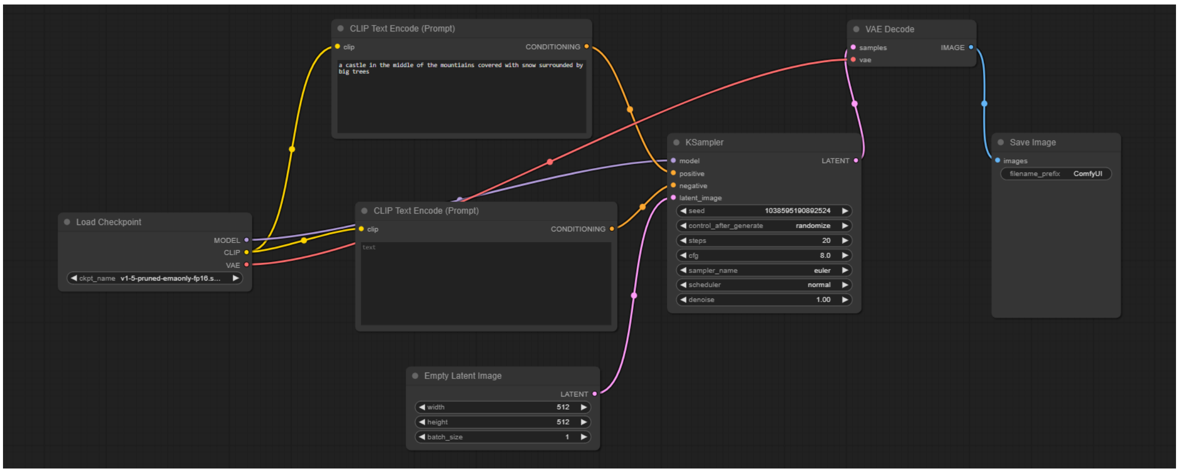
Existing models have limitations: GANs are unstable, VAEs produce blurry images, and diffusion models are computationally intensive. Stable Diffusion addresses these by operating in a latent space, reducing computational load, and enabling controlled generation through conditioning. Stable Video Diffusion extends this to videos, addressing the lack of temporal consistency in existing models. The use of CLIP for evaluation provides an objective measure of model performance, overcoming the subjective nature of qualitative assessments.

In summary, stable diffusion and stable video diffusion models advance generative AI by addressing key limitations in existing models, offering efficient, high-quality, and controllable image and video generation.

CHAPTER 3

Proposed Methodology

3.1 System Design



a castle in the middle of the mountains covered with snow surrounded by big trees

ckpt_name v1-5-pruned-emaonly-fp16.s...

seed 1038595190892524
control_after_generate randomize
steps 20
cfg 8.0
sampler_name euler
scheduler normal
denoise 1.00

filename_prefix ComfyUI

width 512
height 512
batch_size 1

Fig 3.1.1 Components of ComfyUI

Core Workflow Elements:

- **Load Checkpoint:** This node (the purple one on the left) is crucial. It loads a pre-trained model (like Stable Diffusion) along with its associated VAE (Variational Autoencoder). This model is the engine for generating images. The "clip_skip" parameter suggests it's also handling CLIP (Contrastive Language–Image Pre-training) model loading, which is used for understanding text prompts.
- **CLIP Text Encode (Prompt):** This node (the yellow one at the top) takes your text prompt as input (e.g., "a castle in the middle of the mountains covered with snow surrounded by big trees") and converts it into a format the model can understand. This is done using the CLIP text encoder.
- **CLIP Text Encode (Prompt):** This second yellow node (below the first) seems to be taking a shorter text input "dip". It likely serves as a secondary or conditioning prompt, potentially influencing the generation in a more subtle way.
- **WAF Decode:** This node (the purple one on the right) is responsible for decoding the latent space representation (the intermediate data the model works with) into an actual image. WAF likely stands for a specific decoding method or model being used.

- **KSampler:** This is the large central node with multiple parameters. It's the core of the image generation process. It takes the encoded text prompts (from the CLIP nodes), the loaded model, and settings like the number of steps, CFG scale, and sampler type. It iteratively refines the image in the latent space.

3.2 Requirement Specification

Mention the tools and technologies required to implement the solution.

3.2.1 Hardware Requirements:

- NVIDIA GPUs: Minimum 8 GB VRAM (12 GB+ recommended). Use PyTorch with CUDA.
- AMD GPUs (Linux only): Minimum 8 GB VRAM, use ROCm and PyTorch.
- Intel GPUs: Minimum 8 GB VRAM, use PyTorch with XPU support.
- CPU-Only: Modern multi-core CPU (i7/i9, Ryzen 7/9), 16 GB RAM (32 GB+ recommended).
- Apple Silicon: M1/M2 chip, 16 GB RAM, use PyTorch nightly.
- DirectML (Windows): For older AMD cards, use PyTorch with DirectML.
- Storage: 10 GB+ free space.

3.2.2 Software Requirements:

- ComfyUI
- Stable Diffusion

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

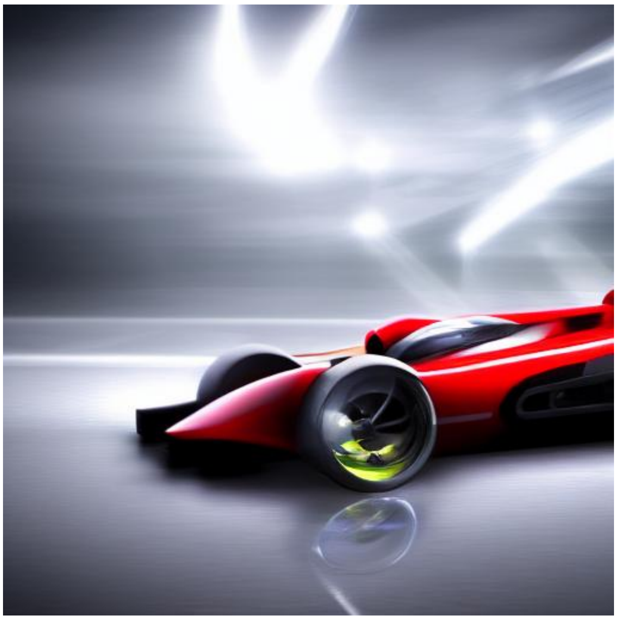


Fig 4.1.1 : A futuristic racing car speeding through

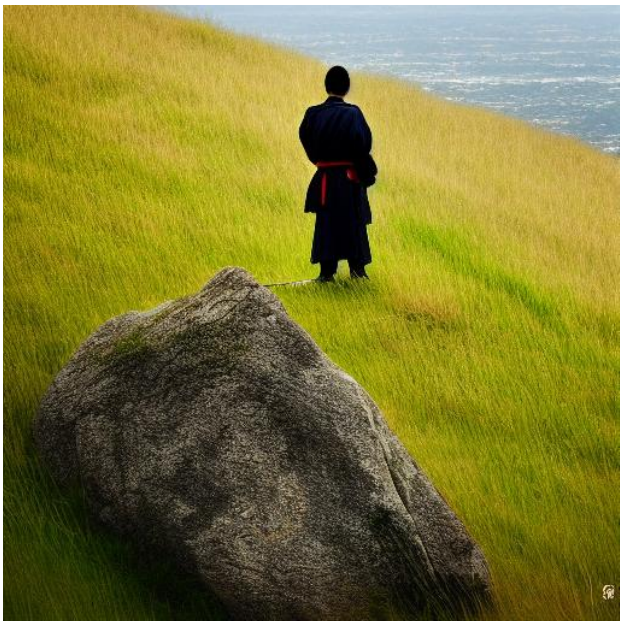


Fig 4.1.2 : A lone samurai meditating on a cliff, the wind blowing through the grass

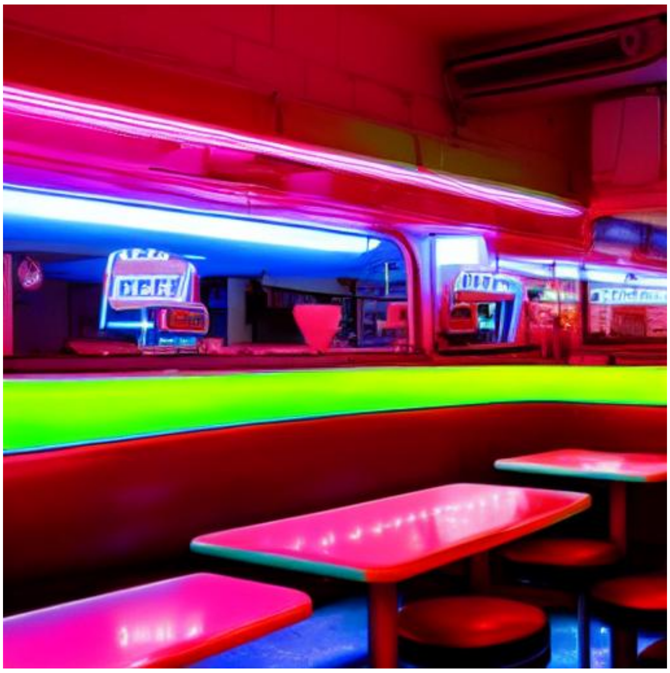


Fig 4.1.3 : A retro 1950s diner glowing with neon lights, classic cars parked outside.

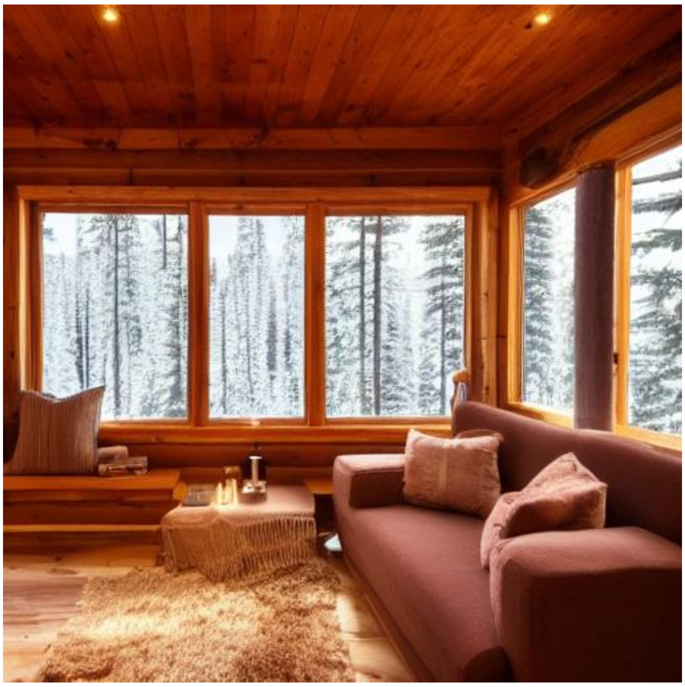


Fig 4.1.4 : A cozy wooden cabin in a snowy forest, warm lights glowing inside

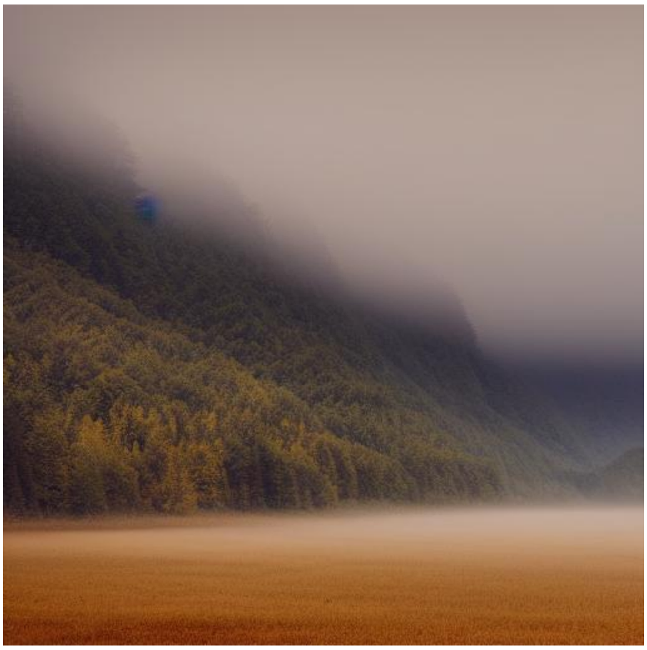


Fig 4.1.5 : A misty mountain

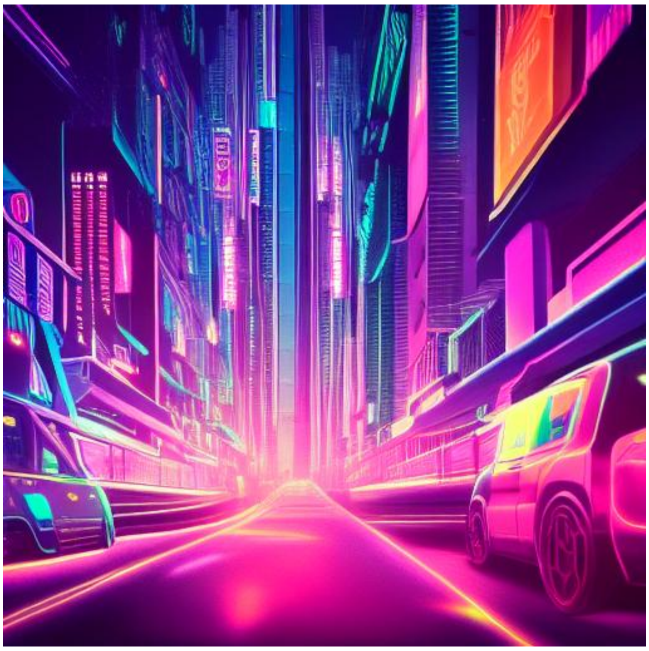


Fig 4.1.6 : Futuristic cityscape at night, neon lights, cyberpunk aesthetic

4.2GitHub Link for Code:

https://github.com/CharanTejManchikanti/Image_generation-using-Stable-Diffusion-and-ComfyUI

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

- **Temporal Consistency**: Improve transitions in video generation using refined temporal layers.
- **Optimization for Low-End Hardware**: Explore quantization, pruning, or distillation for better efficiency.
- **Advanced Conditioning**: Integrate multi-modal inputs (text, audio, images) for richer outputs.
- **Dataset Improvement**: Expand diverse, high-quality datasets for better generalization.
- **Real-Time Generation**: Reduce denoising steps for faster synthesis.
- **Cross-Platform Support**: Enhance performance on AMD (Windows) and Intel GPUs.
- **Ethical Considerations**: Implement safeguards against misuse (e.g., deepfakes, bias).

5.2 Conclusion:

Stable Diffusion and Stable Video Diffusion enable high-quality image and video generation, leveraging latent space optimization and iterative denoising. Key contributions include realistic image synthesis, controlled text-based generation, improved video consistency, and cross-hardware support. The project advances generative AI and will continue evolving for efficiency, accessibility, and ethical responsibility

REFERENCES

[1] Rombach, Robin, et al. “High-resolution image synthesis with latent diffusion models.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. *2022*.