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

MOHAMMED ZAIDAAN SHIRAZ, zaidaanshiraz8@gmail.com

Under the Guidance of

Jay Rathod ,P Adarsh

ACKNOWLEDGEMENT

I sincerely thank my mentors, **Jay Rathod Sir and Adarsh P Sir**, for their invaluable guidance and support throughout this Project, **“Image Generation Ussing Stable Diffusion and ComfyUI”.** Their encouragement and insightful feedback have helped me improve my understanding and skills. Their belief in my abilities has been a great source of motivation, pushing me to think critically and grow with confidence.

Though our time was short, their mentorship has had a lasting impact. I have not only gained technical knowledge but also learned the importance of problem-solving, professionalism, and continuous learning. Their patience and willingness to help have truly inspired me.

I am also grateful to **Pavan Kumar Sir** for conducting the sessions smoothly and ensuring an effective learning experience. His dedication made the learning process more engaging and meaningful.

Special thanks to **TechSaksham, a joint initiative by Microsoft and SAP**, for providing this incredible opportunity. This platform has allowed me to explore and apply my knowledge practically, and develop skills that will shape my future.

Once again, my heartfelt gratitude to **Jay Rathod Sir, Adarsh P Sir, Pavan Kumar Sir, and the entire TechSaksham team** for their support and mentorship. This experience has been truly enriching and will always be a valuable part of my learning journey.

MOHAMMED ZAIDAAN SHIRAZ

ABSTRACT

This project improves how AI turns text into images using two key ideas: a simpler way to train image-generating models and a smarter design for understanding text prompts. While current AI tools like Stable Diffusion (which creates images from text) work well, They can struggle with complex prompts or fine details. Our goal was to make these tools more reliable, user-friendly, and better at matching what users imagine.

We focused on two improvements:

1. **Better Training Method:** Instead of randomly adding noise to images during training, we guided the AI to focus on noise patterns that matter most to human eyes. Think of it like teaching someone to clean a messy room by first showing them which areas people notice most.
2. **Smarter Architecture:** We designed a new AI model that lets text and images "talk" to each other during generation. For example, if you type "a cat wearing sunglasses" the model ensures the sunglasses actually fit the cat's face and aren't just pasted randomly. This uses Stable Diffusion's core system (which turns text into images step-by-step) but adds a fresh way for words and visuals to interact.

Using ComfyUI—a visual, drag-and-drop tool for Stable Diffusion we tested our approach with users.

The results showed:

Images looked sharper and matched text prompts more accurately (e.g., text in images was readable, objects were placed logically).

Users preferred our model over existing tools, especially for tricky prompts like "*a steampunk owl reading a newspaper*."

The system became faster and easier to customize, thanks to ComfyUI's flexible design.

In simple terms, this project makes AI image generators more like a helpful artist: they listen better to your ideas, make fewer mistakes, and work smoothly with tools people already enjoy. We're sharing our code and models openly so everyone can build on these improvements. This work is a step toward AI that bridges the gap between imagination and reality, giving users more control and creativity.

.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Problem Statement:

Current diffusion models for text-to-image synthesis face limitations in (1) inefficient noise sampling during training, neglecting perceptually critical scales, and (2) poor bidirectional text-image alignment, causing weak typography and text comprehension. While rectified flow offers theoretical advantages, its adoption is hindered by unoptimized noise strategies and scalability gaps.

This project proposes:

1. **Perceptually Guided Noise Sampling**: Enhancing rectified flow training with noise schedules prioritizing human-visually relevant scales to boost high-resolution image quality.
2. **Bidirectional Transformer Architecture**: A novel model with modality-specific weights and cross-attention mechanisms for dynamic text-image interaction, improving text fidelity.

Leveraging **Stable Diffusion**'s VAE, U-Net, and CLIP components via **ComfyUI**'s modular interface, the work integrates these innovations into reproducible workflows. The approach combines ComfyUI's node-based customization with optimized training pipelines to refine noise prediction and text conditioning.

1.2 Motivation:

AI image generation tools like Stable Diffusion hold vast potential for creative industries, education, and accessible content creation. However, limitations in high-resolution output quality and text-image alignment hinder their reliability for professional use. Current models often produce incoherent typography or struggle with complex prompts, while rectified flow—a promising, simpler framework—remains underdeveloped. This project bridges these gaps by optimizing noise sampling for perceptual relevance and introducing a bidirectional text-image architecture to enhance semantic accuracy. By integrating these innovations with ComfyUI's accessible interface, the work democratizes advanced synthesis for non-experts while addressing ethical risks through open-source transparency, fostering equitable, creative AI tools.

1.3 Objective:

1. **Optimize Noise Sampling**: Develop perceptually guided noise schedules for rectifying flow models to enhance high-resolution image quality and training efficiency.
2. **Design Bidirectional Architecture**: Build a transformer-based model with modality-specific weights and cross-attention mechanisms to improve text-image alignment, typography, and prompt fidelity.
3. **Integrate with ComfyUI**: Implement the enhanced framework into ComfyUI's node-based workflows, enabling accessible experimentation and customization for diverse users.
4. **Benchmark Performance**: Validate improvements via quantitative metrics (FID, CLIP score) and human evaluations against state-of-the-art diffusion models.

1.4 Scope of the Project:

Scope:

1. **Enhanced Noise Sampling**: **Develop** perceptually guided noise schedules for Stable Diffusion's rectified flow models to improve high-resolution image generation and training efficiency.
2. **Bidirectional Text-Image Model**: Design a transformer-based architecture with cross-modal attention to refine text comprehension, typography, and prompt adherence.
3. **ComfyUI Integration**: Implement user-friendly workflows in ComfyUI for customizable Image generation, including nodes for noise adjustment, text conditioning, and output evaluation.

Limitations:

1. **Hardware Requirements**: High-resolution generation and training demand powerful GPUs, restricting access for users with limited computational resources.
2. **Inference Speed**: Generating **detailed** images (e.g., 4K) may incur longer processing times compared to standard diffusion pipelines.
3. **Training Data Bias**: Model performance relies on existing datasets, potentially propagating societal biases in generated outputs.
4. **Ethical Risks**: Open-source availability of advanced models raises concerns about misuse, such as deepfakes or unauthorized content creation.

CHAPTER 2

Literature Survey

2.1 Review relevant literature

1. **Evolution of Image Generation Techniques**

   - Early methods relied on adversarial networks that used a competition between two networks to generate images.
   - These approaches introduced innovative ideas but often faced challenges like instability and difficulty in capturing fine details.

2. **Transition to Diffusion Models**

   - Diffusion models start with random noise and gradually refine it into a coherent image through an iterative process.
   - This method improves image quality by mimicking the gradual process of refining a sketch into a complete picture.

3. **Advances in Text-to-Image Synthesis**

   - Recent models can combine detailed textual instructions with the image generation process.
   - Researchers have developed techniques to convert text prompts into clear guidance for generating images, ensuring that the visual output aligns closely with the given description.

4. **Optimization of Prompt Engineering**

   - Effective prompt design is crucial for steering the generation process toward the desired output.

2.2 Existing Models, Techniques, and Methodologies

1. **Stable Diffusion Models**
   - These models operate by progressively removing noise from an image until a final, high-quality output is achieved.
   - Working in a compressed latent space reduces computational requirements while maintaining image detail.
2. **Guided Diffusion Techniques**
   - Additional information is used during the diffusion process to steer the output, ensuring the generated images better match user intent.
   - Conditional control techniques have been developed to offer more precise influence over the style and content of the final image.
3. **User-Centric Prompt Engineering**
   - Detailed guidelines and strategies help users craft effective prompts for better image generation.
   - Negative prompting techniques are used to suppress unwanted elements in the final image.
4. **ComfyUI: A User-Friendly Interface**
   - ComfyUI provides an intuitive visual workflow that allows users to adjust Parameters easily.
   - The interface democratizes advanced image generation, making it accessible to non-experts and encouraging creative experimentation.

2.3 Limitations in Existing Systems

1. **Limitations of Current Systems**

   - **Dependency on Prompt Engineering:** Small changes in wording can lead to significantly different outcomes, making consistency a challenge.
   - **Computational Intensity:** The iterative process of diffusion models requires Substantial computing power, which can slow down real-time applications.
   - **Evaluation Difficulties:** There is no standard metric that fully captures both Technical quality and artistic appeal, complicating comparisons between different models.

2. **How This Project Addresses the Gaps**

   - **Enhanced Prompt Optimization:** Incorporating advanced techniques to refine text prompts for more robust and consistent image generation.
   - **User-Friendly Tools:** Leveraging interfaces like ComfyUI to simplify parameter adjustments and encourage broader experimentation.
   - **Adaptive Workflows:** Developing feedback-driven mechanisms that automatically fine-tune the image generation process based on user preferences and evaluation outcomes.

CHAPTER 3

Proposed Methodology

Below is the proposed methodology for the project, structured to clearly outline the system design and requirement specifications, along with additional components that support the overall workflow.

3.1 System Design

3.1.1 Overall System Workflow

The system design integrates several interconnected modules to ensure smooth functionality:

1. **Overall Architecture:**
   - Handles user-provided text prompts and any additional control signals.
   - Pre-processes the input to ensure compatibility with the generation engine.
2. **Core Generation Engine:**
   - Uses a stable diffusion model that begins with random noise and refines it through iterative steps to produce high-quality images.
   - Operates in a latent space to reduce computational demands while maintaining fine details.
3. **User Interface Layer (ComfyUI):**
   - Provides an intuitive, visual workflow that allows users to easily modify parameters and monitor progress in real time.
   - Designed to be accessible for both experts and novices, facilitating experimentation and creative exploration.
4. **Feedback and Adaptation Module:**
   - Captures user feedback on generated images and adjusts parameters automatically to improve future outputs.
5. **Evaluation and Logging:**
   - Monitors performance metrics and logs data to support iterative enhancements and ensure the consistency of image quality.
6. **Data Flow and Integration:**
   - The system accepts textual input, processes and optimizes the prompt, and then feeds into the core engine.

- The generated image is rendered through the user interface, where real-time adjustments and user feedback are integrated into the adaptive workflow.
- Each module is designed to interact seamlessly, ensuring that improvements in one component can be independently updated without affecting the overall system

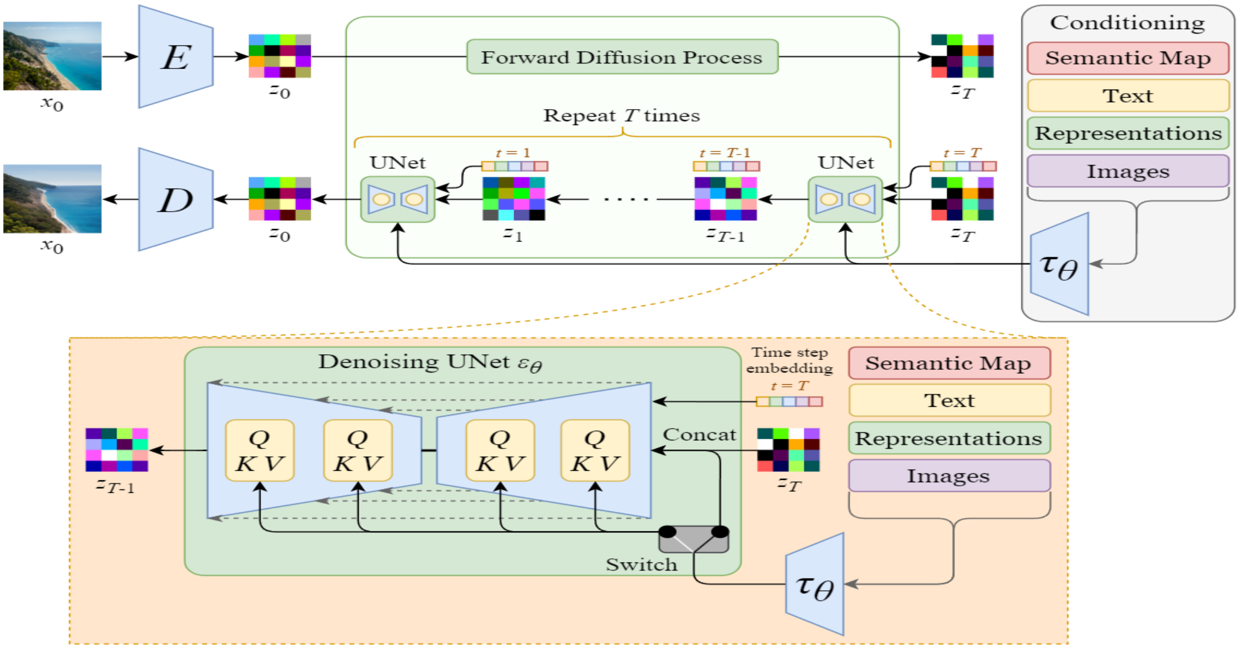


Figure 1: System Workflow for image generation using stable diffusion

3.1.2 Stable Diffusion WorkFlow

This diagram illustrates the diffusion process for generating images and how a denoising model, such as a UNet, is used in this framework. Here's an explanation:

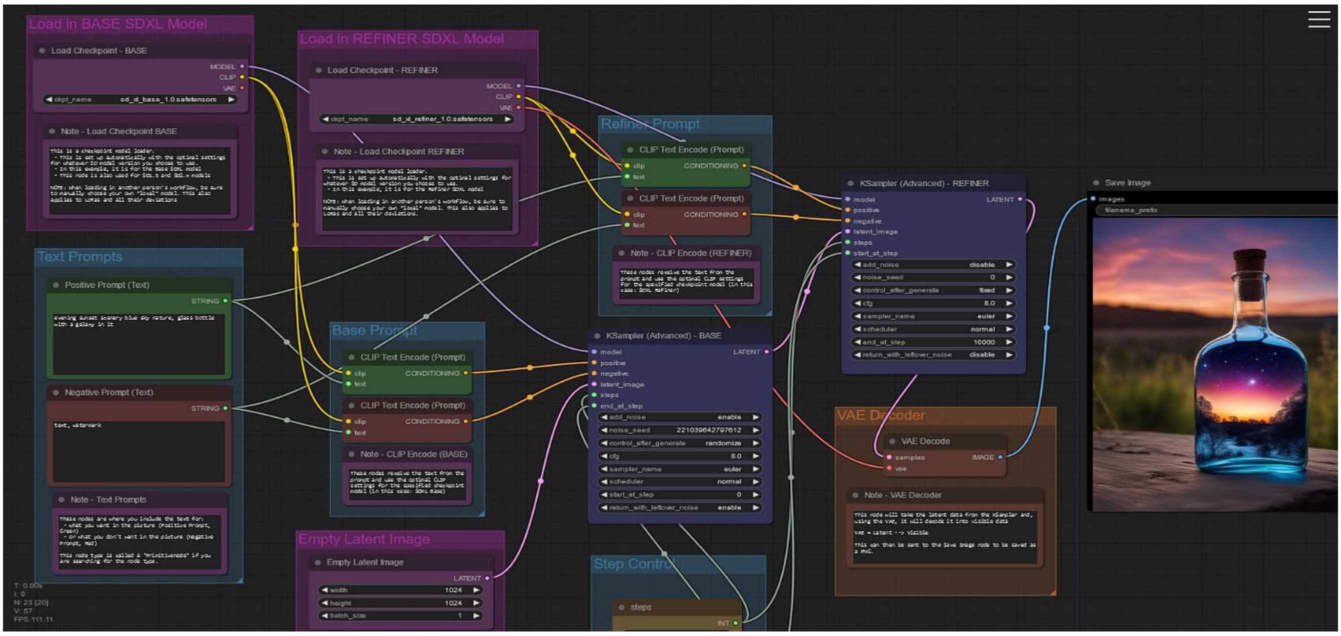
Top Section: Diffusion Process Overview

1. **Forward Diffusion Process (Left Path):**
   - Starts with a clean image x0_x0 (e.g., a beach image).
   - Gradually adds noise over multiple time steps (t=0 to t=T) to transform the original image into a noisy representation (zT).
2. **Reverse Diffusion Process (Right Path):**
   - The noisy representation zT is passed through a model (UNet), which iteratively denoises it.
   - At each step t, the model refines the noisy image until it closely resembles the original image x0_x0.
3. **Conditioning Information:**
   - The model is guided using conditioning inputs like semantic maps, text

descriptions, representations, and prior image context. These inputs help steer the generation process to align with the desired output.

Bottom Section: Denoising UNet Architecture

1. **Denoising UNet (θz):**
   - The core of the reverse process is the UNet model, which is responsible for predicting the noise and refining the image.
   - Attention Mechanism (Q, K, V):
   - "Query (Q), Key (K), and Value (V)" modules highlight critical features for each step, enabling the model to focus on relevant details.
2. **Concatenation of Inputs:**
   - Inputs like semantic maps, text embeddings, and noisy image representations are combined and fed into the UNet for better alignment and refinement.
3. **Switch and Time Step Embedding:**
   - The time step embedding incorporates the diffusion step (ttt) into the model, helping it understand which stage of the denoising process it's working on.
   - The "Switch" mechanism adjusts pathways to handle specific processing tasks dynamically.



ckpt_name sd_xl_base_1.0.safetensors

This is a checkpoint model loader.
 - This is set up automatically with the optimal settings for whatever SD model version you choose to use.
 - In this example, it is for the Base SDXL model
 - This node is also used for SD1.5 and SD2.x models

NOTE: When loading in another person's workflow, be sure to manually choose your own "local" model. This also applies to LoRAs and all their deviations

ckpt_name sd_xl_refiner_1.0.safetensors

This is a checkpoint model loader.
 - This is set up automatically with the optimal settings for whatever SD model version you choose to use.
 - In this example, it is for the Refiner SDXL model

NOTE: When loading in another person's workflow, be sure to manually choose your own "local" model. This also applies to LoRAs and all their deviations.

evening sunset scenery blue sky nature, glass bottle with a galaxy in it

text, watermark

These nodes are where you include the text for:
 - what you want in the picture (Positive Prompt, Green)
 - or what you don't want in the picture (Negative Prompt, Red)

This node type is called a "PrimitiveNode" if you are searching for the node type.

These nodes receive the text from the prompt and use the optimal CLIP settings for the specified checkpoint model (in this case: SDXL Base)

width 1024
height 1024
batch_size 1

These nodes receive the text from the prompt and use the optimal CLIP settings for the specified checkpoint model (in this case: SDXL Refiner)

add_noise enable
noise_seed 221039642797612
control_after_generate randomize
cfg 8.0
sampler_name euler
scheduler normal
start_at_step 0
return_with_leftover_noise enable

add_noise disable
noise_seed 0
control_after_generate fixed
cfg 8.0
sampler_name euler
scheduler normal
end_at_step 10000
return_with_leftover_noise disable

This node will take the latent data from the KSampler and, using the VAE, it will decode it into visible data

VAE = Latent --> Visible

This can then be sent to the Save Image node to be saved as a PNG.

Figure 2: ComfyUI Workflow for text-to-Image Generation

3.1.3 ComfyUI WorkFlow

This image represents a ComfyUI workflow diagram for generating an AI-generated image using the SDXL (Stable Diffusion XL) Model along with a refiner model.

1. Base SDXL Model (Load Checkpoint - BASE)

- **Purpose**: Loads the base SDXL model checkpoint, which is the foundational model used for generating the initial image.
- **Details:** This model handles the bulk of the image generation process, focusing on capturing the structure and overall details of the scene.

2. Refiner SDXL Model (Load Checkpoint - REFINER)

- **Purpose:** Loads the refiner checkpoint, a complementary model used to enhance the details, sharpness, and quality of the base-generated image.
- **Details:** The refiner model improves subtle details, textures, and finer elements in the image, providing a more polished output.

3. Text Prompts

- **Positive Prompt:** Specifies the desired features or elements in the image. For example, it might contain phrases like “a magical potion bottle with glowing liquid, sunset reflection.”
- **Negative Prompt:** Specifies elements to avoid in the image. For instance, it could include “blurry, low quality, distorted.”

4. CLIP Text Encoding

- **Base Prompt Encoding:** Converts the positive and negative prompts into embeddings that guides the base model in generating the image.
- **Refiner Prompt Encoding:** Similarly encodes prompts to guide the refiner model in improving the image details.

5. Empty Latent Image

- **Purpose:** Initializes a blank latent space (a placeholder) where the model begins the process of image generation. This latent space evolves into the final image through multiple iterations.

6. K-Sampler (Advanced)

- **Base Sampler:** Processes the latent image using the base SDXL model to create an initial image based on the prompts.

- **Refiner Sampler:** Further processes the latent image using the refiner model to enhance details, textures, and overall quality.

7. VAE Decoder

- **Purpose:** Decodes the latent representation of the image into a visible, high-resolution image format.
- **Details:** This is the final step where the encoded image is transformed into It’s RGB pixel representation.

8. Save Image

- **Purpose:** The final generated image is saved in the specified format and location. In this case, the result is an artistic image of a potion bottle with glowing liquid, set against a backdrop of a serene sunset.

Output Image:

The generated image is a beautifully rendered artistic composition:

- A glass bottle with a glowing, vibrant liquid inside.
- A sunset scene reflected within the bottle, blending realism and fantasy.
- Detailed textures and high-quality rendering due to the collaboration of the base and refiner SDXL models.

3.2 Requirement Specification

3.2.1 Hardware Requirements:

- **Processor (CPU):**
  - A multi-core processor (e.g., Intel i7 or AMD Ryzen 7) is recommended to handle the intensive computations during image processing.
- **Graphics Processing Unit (GPU):**
  - A dedicated GPU (e.g., NVIDIA RTX 3070 or higher) is essential for accelerating the diffusion processes and ensuring efficient model inference.
- **Memory (RAM):**
  - At least 16GB of RAM is necessary; however, 32GB or more is preferred to manage large model parameters and support parallel processing.
- **Storage:**
  - A fast SSD with a minimum of 500GB is recommended to facilitate quick data access, loading of models, and smooth operation.
- **Network:**
  - A stable and fast internet connection is required, especially if the system is to support cloud-based updates or remote feedback synchronization.

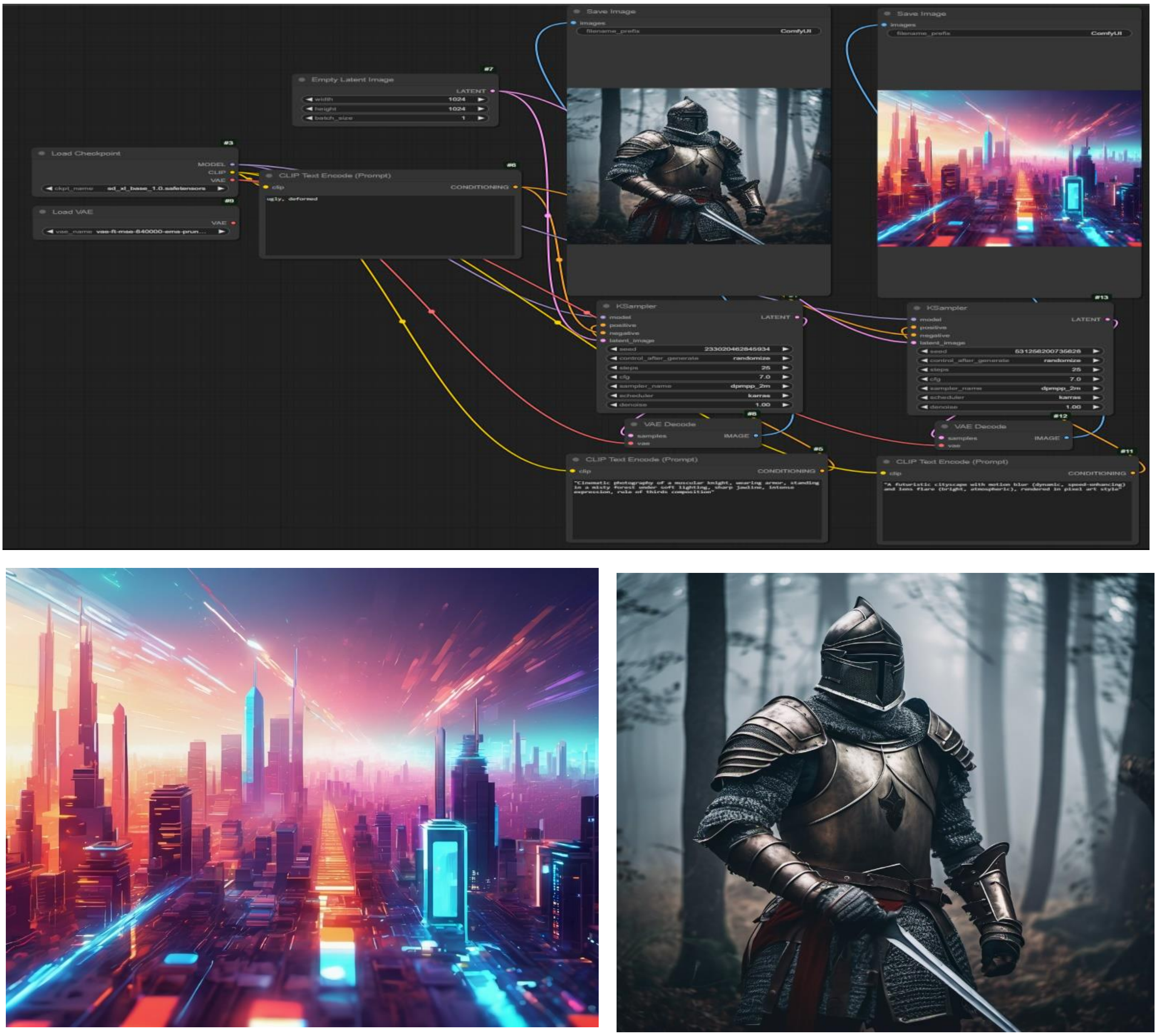
3.2.2 Software Requirements:

- **Operating System:** Windows/Linux/MacOS.
- **Programming Language:** Python 3.x.
- **Libraries/Frameworks:**
  - libraries for machine learning (such as PyTorch or TensorFlow) and diffusion models.
  - A version control system (like Git) for collaborative development and tracking changes.
- **User Interface Framework:**
  - ComfyUI is required as the front-end framework, providing an intuitive interface that allows users to adjust parameters and visualize real-time outputs.

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:



Save Image
images
filename_prefix ComfyUI

Save Image
images
filename_prefix ComfyUI

#7 Empty Latent Image
LATENT
width 1024
height 1024
batch_size 1

#3 Load Checkpoint
MODEL
CLIP
VAE
ckpt_name sd_xl_base_1.0.safetensors

#9 Load VAE
VAE
vae_name vae-ft-mse-840000-ema-prun...

#6 CLIP Text Encode (Prompt)
clip
CONDITIONING
ugly, deformed

KSampler
model
positive
negative
latent_image
LATENT
seed 233020462845934
control_after_generate randomize
steps 25
cfg 7.0
sampler_name dpmpp_2m
scheduler karras
denoise 1.00

#13 KSampler
model
positive
negative
latent_image
LATENT
seed 531256200735628
control_after_generate randomize
steps 25
cfg 7.0
sampler_name dpmpp_2m
scheduler karras
denoise 1.00

#8 VAE Decode
samples
vae
IMAGE

#12 VAE Decode
samples
vae
IMAGE

#5 CLIP Text Encode (Prompt)
clip
CONDITIONING
"Cinematic photography of a muscular knight, wearing armor, standing in a misty forest under soft lighting, sharp jawline, intense expression, rule of thirds composition"

#11 CLIP Text Encode (Prompt)
clip
CONDITIONING
"A futuristic cityscape with motion blur (dynamic, speed-enhancing) and lens flare (bright, atmospheric), rendered in pixel art style"

Figure 3: Snapshots of the ComfyUI Workflow: SDXL Text-to-Image Generation with Diffusion Models.

Prompts:

1. *"Cinematic photography of a muscular knight, wearing armor, standing in a misty forest under soft lighting"*
2. *"A futuristic cityscape with motion blur (dynamic, speed-enhancing) and lens flare (bright, atmospheric), rendered in pixel art style"*

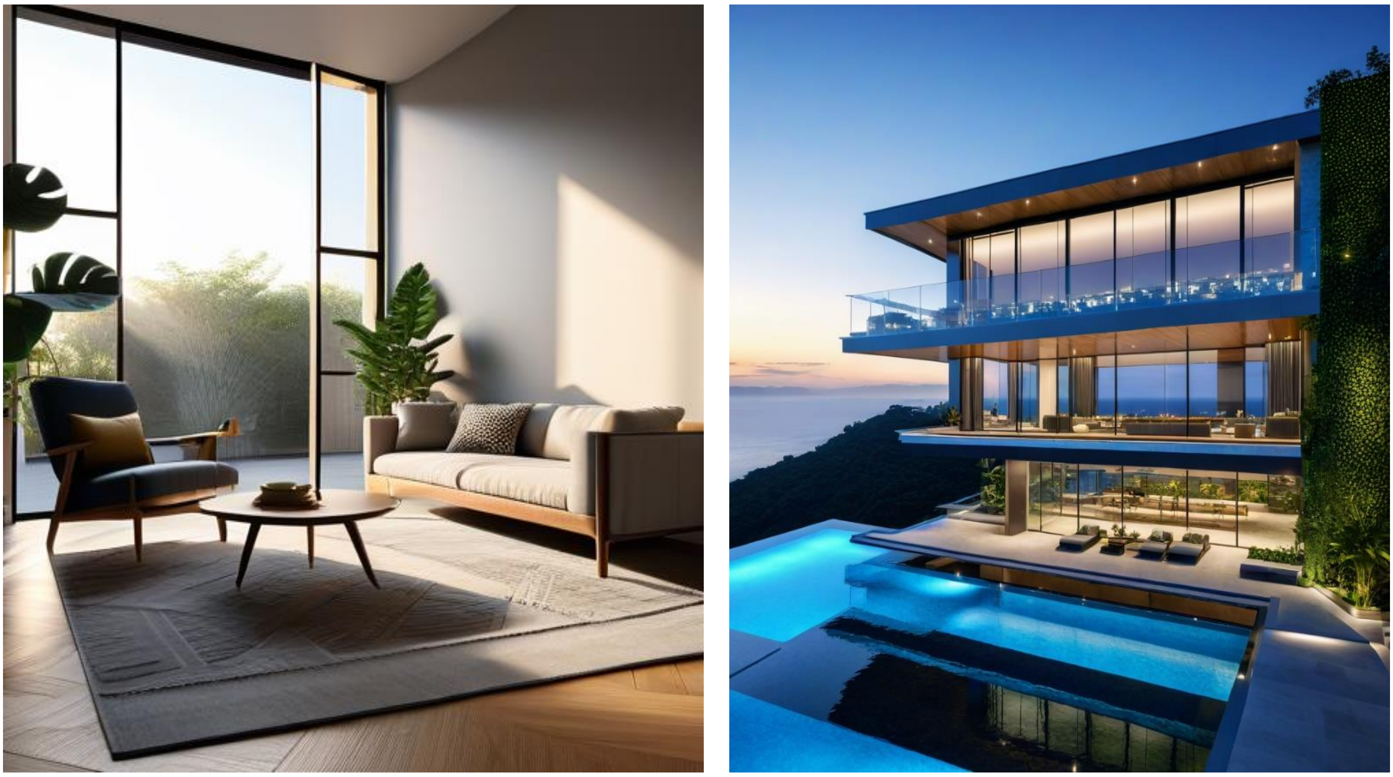


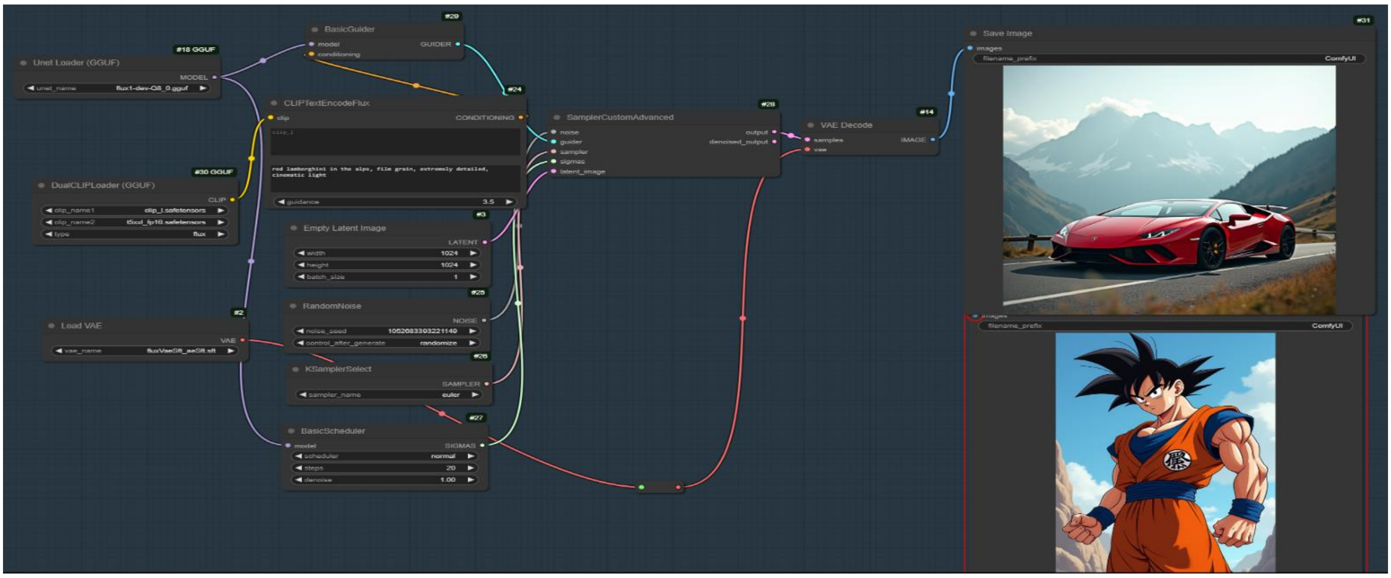
Figure 4: Snapshots of the ComfyUI Workflow: RealArchVisXL Model for Text-to-Image Synthesis.

Description:

- This snapshot illustrates a ComfyUI workflow utilizing the RealArchVisXL model for generating high-quality, realistic architectural and interior design images from text prompts. The nodes depict the seamless integration of text encoding, model checkpoints, and rendering processes to create visually stunning outputs.
- The RealArchVisXL model specializes in generating photorealistic architectural and environmental visualizations. It is ideal for creating detailed and immersive scenes, including urban landscapes, interiors, and exteriors, with remarkable realism and precision.

Prompts:

1. *“Ultra-detailed modern living room, minimalist Scandinavian design, floor-to-ceiling windows with golden hour sunlight, soft shadows, concrete walls, oak wood flooring”*
2. *"A Futuristic eco-luxury villa at twilight, glass and steel architecture, solar-panelled roof, terraced landscaping with vertical gardens, infinity pool reflecting city skyline”*



unet_name flux1-dev-Q8_0.gguf

red lamborghini in the alps, film grain, extremely detailed, cinematic light

guidance 3.5

clip_name1 clip_l.safetensors
clip_name2 t5xxl_fp16.safetensors
type flux

width 1024
height 1024
batch_size 1

noise_seed 1052083303221140
control_after_generate randomize

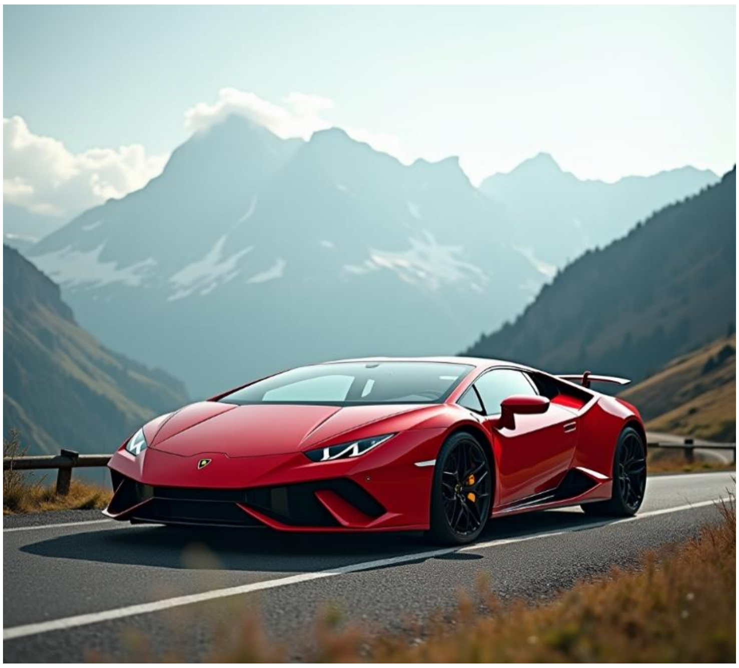
vae_name fluxVaeSft_aeSft.sft

sampler_name euler

scheduler normal
steps 20
denoise 1.00

filename_prefix ComfyUI

filename_prefix ComfyUI



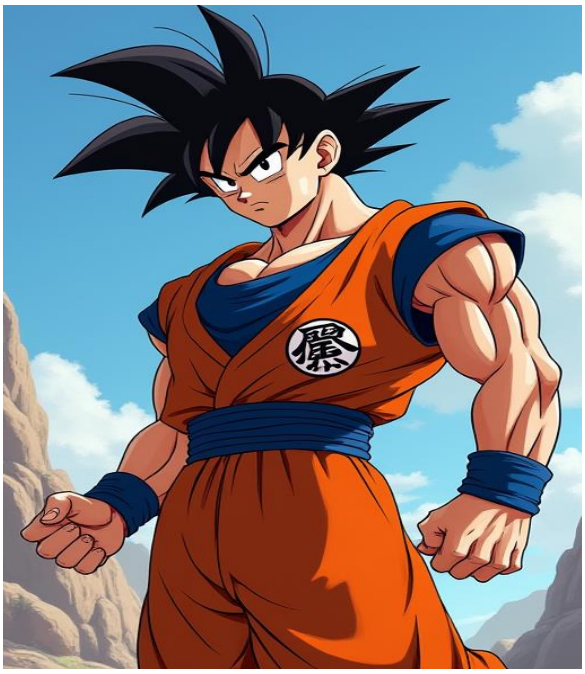


Figure 5: Snapshots of the ComfyUI Workflow: Flux Q8 GGUF Diffusion Model for text to image synthesis

Flux Model: The Flux model is a powerful text-to-image generator designed for artistic and conceptual visuals. It excels in producing vibrant, stylized, and imaginative imagery, making it ideal for creative projects and abstract designs.

Flux GGUF Model: The Flux GGUF model is an optimized version of the Flux model, designed for better performance on lower-resource systems. It maintains the same artistic capabilities while being more efficient and accessible for broader use cases.

Prompts:

1. *“Red lamborghini in the alps, film grain, extremely detailed, cinematic light”*
2. *"A realistic Dragon Ball Z Character Goku”*

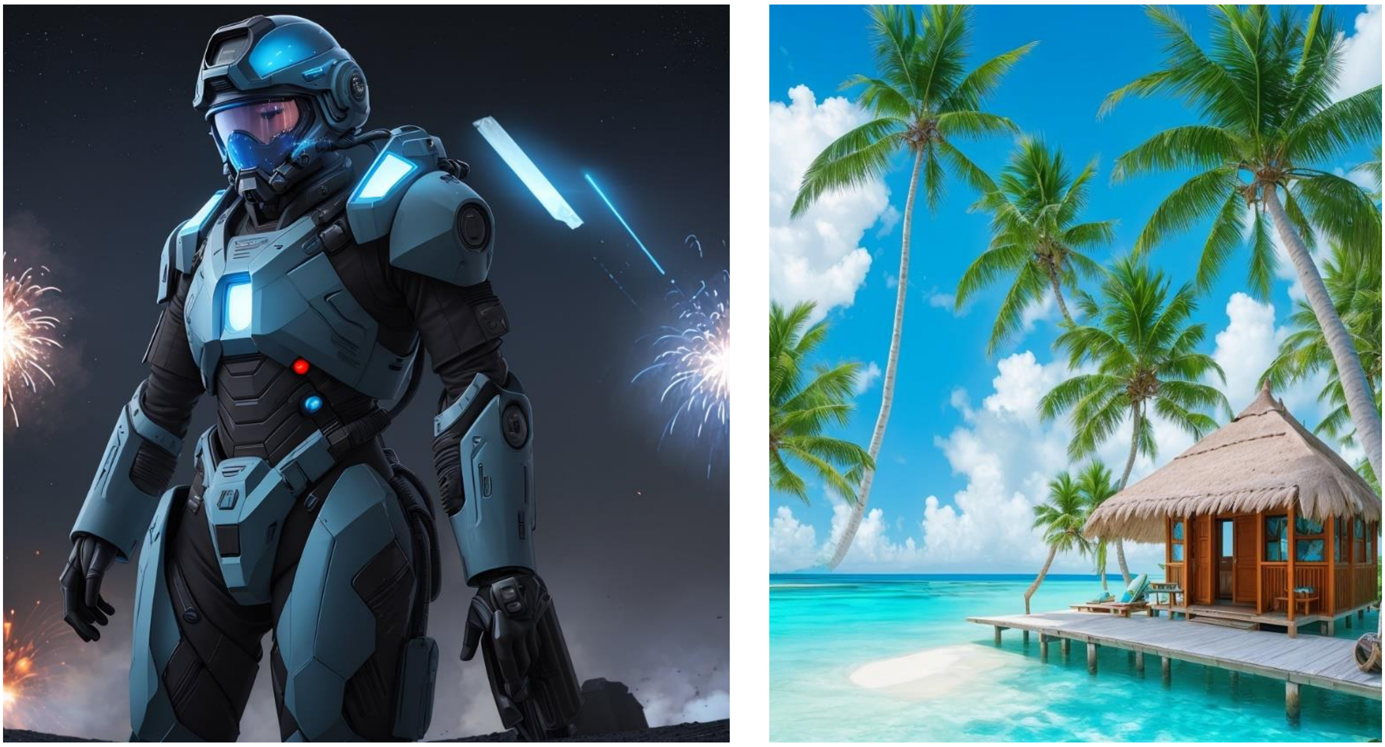


Figure 6: Snapshots of the ComfyUI Workflow: DreamShapers_8 Diffusion Model for text to image synthesis

Description:

DreamShaper 8 Diffusion Model is an advanced text-to-image generative AI model designed for producing highly creative, artistic, and visually stunning images. It blends the strengths of realistic rendering with imaginative artistry, making it ideal for creating fantasy landscapes, detailed portraits, sci-fi concepts, and abstract designs. DreamShaper 8 excels at capturing intricate details, vibrant colors, and dynamic compositions, catering to both hyper-realistic and surreal aesthetics. This model is particularly popular for its versatility and ability to adapt to different artistic styles, ranging from photorealism to dreamy, painterly effects.

Prompts:

1. *"A futuristic soldier in a high-tech exosuit, detailed helmet with glowing blue visors, standing in a battlefield with explosions in the background, cinematic and intense."*
2. *" A tropical island paradise with crystal-clear turquoise water, palm trees swaying in the breeze, and a small hut on the beach, vibrant and peaceful.”*

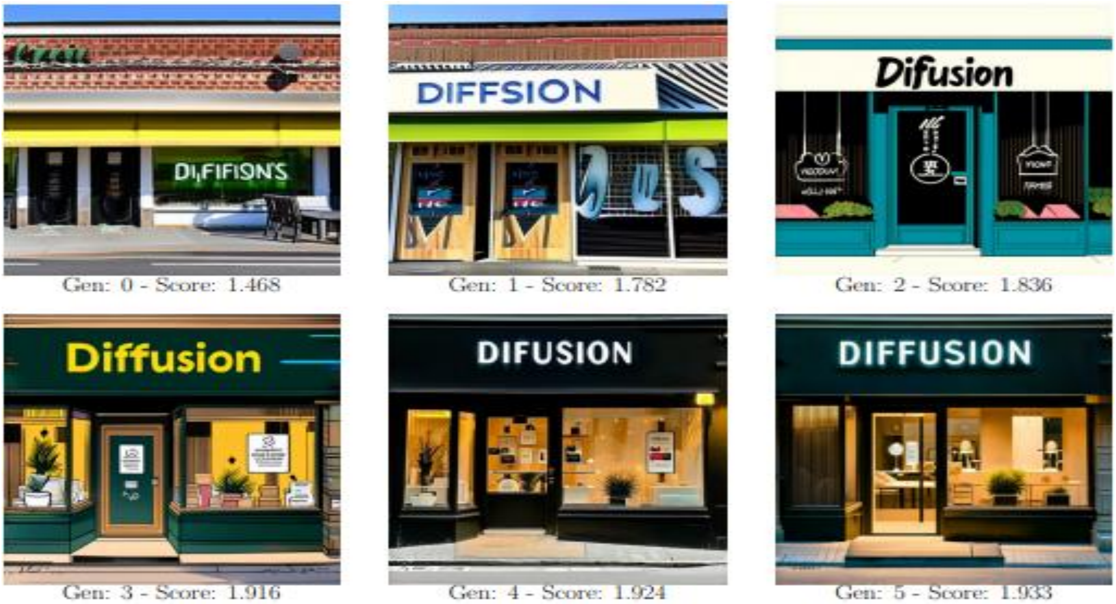
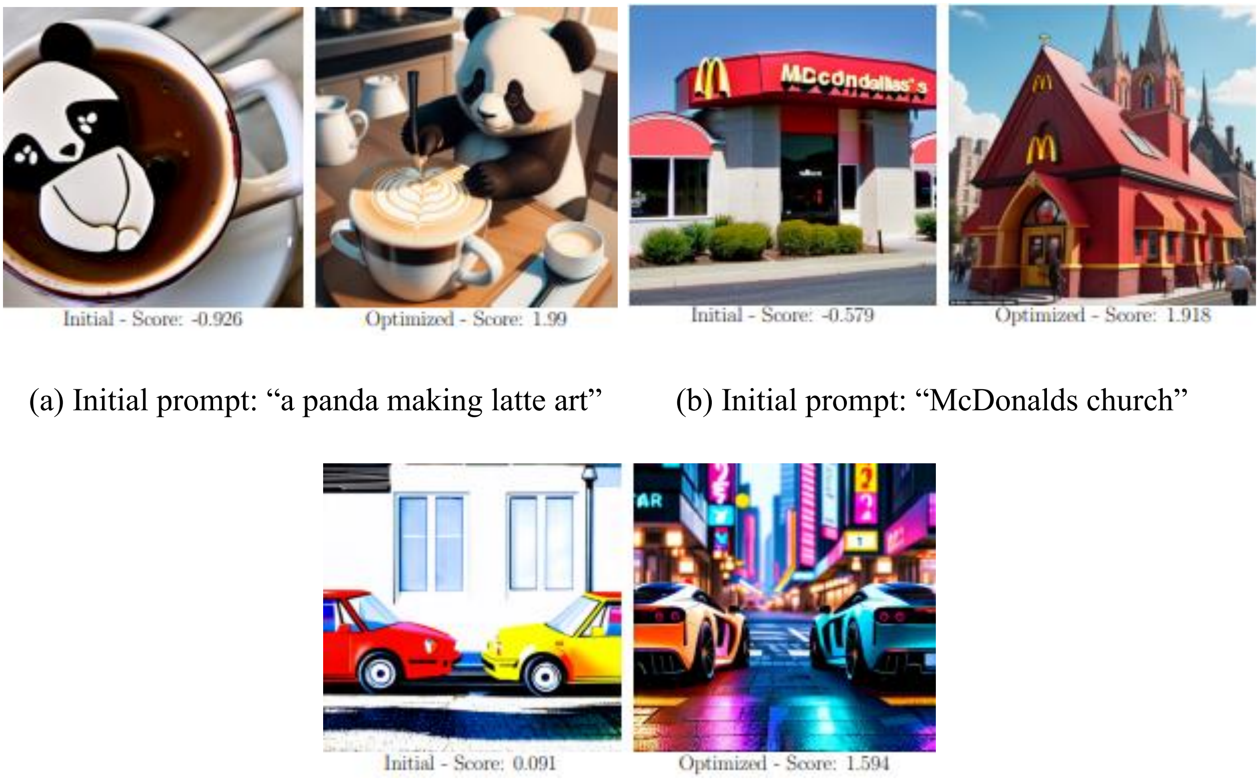


Figure 7: An example for image improvement with ComfyGI over several generations for the prompt “storefront with ‘diffusion’ written on it”. For every generation, we show the image and the score for the best-found patch so far.



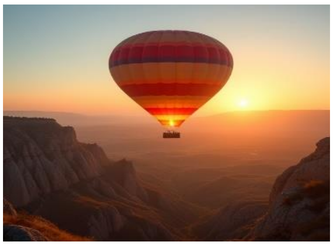
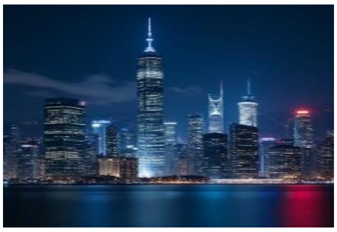
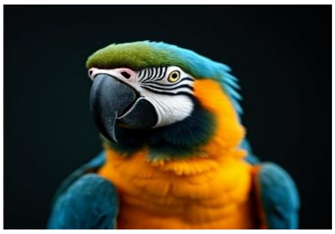
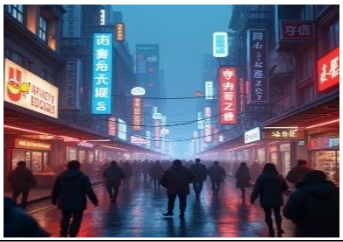
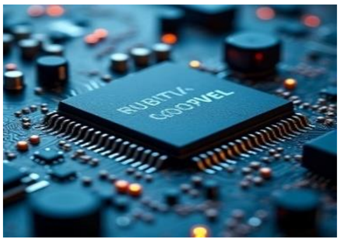
(a) Initial prompt: “a panda making latte art”

(b) Initial prompt: “McDonalds church”

(c) Initial prompt: “two cars on the street”

Figure 8: Three examples for image improvement with ComfyUI. The left image shows the initial image and the right one the optimized counterpart.

Our evaluation using Stable Diffusion and ComfyUI shows that iterative optimization Greatly enhances image quality. For instance, as illustrated in **Figure 7**, an initial error-prone image (with an Image Reward score of 1.468) improved over successive generations, eventually achieving a score of 1.933 with clearer details and accurate text. Similarly, **Figure 8** demonstrates that the optimized images are not only visually more appealing with brighter colors and sharper structures but also more closely aligned with the intended prompts. Overall, our analysis indicates a median improvement of about 50% in image quality, with most gains occurring within the first three generations.

Task Instruction	**Output**
Generate an image of a hot air balloon floating over a scenic valley at sunrise. The result should be a high-quality image.	
Generate an image of a modern city skyline at night with illuminated skyscrapers. The result should be a high-quality image.	
Generate an image of a Highly Detailed Macau Parrot with vibrant colors. The result should be a high-quality image.	
Generate an image of a bustling futuristic marketplace, featuring neon signs, holographic advertisements, and vibrant street life. The result should be a high-quality image.	
Generate an image of an electronics microchip circuit board. The result should be a high-quality image.	

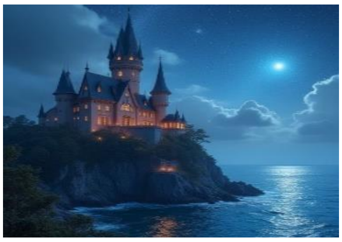
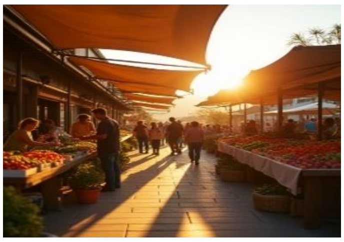
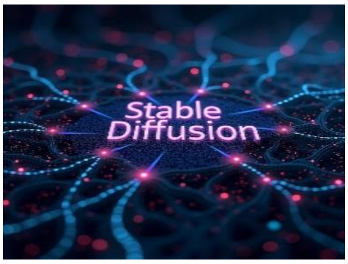
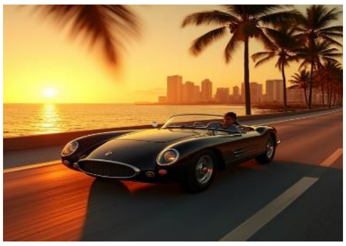
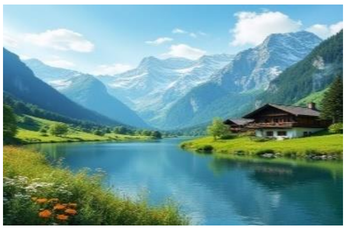
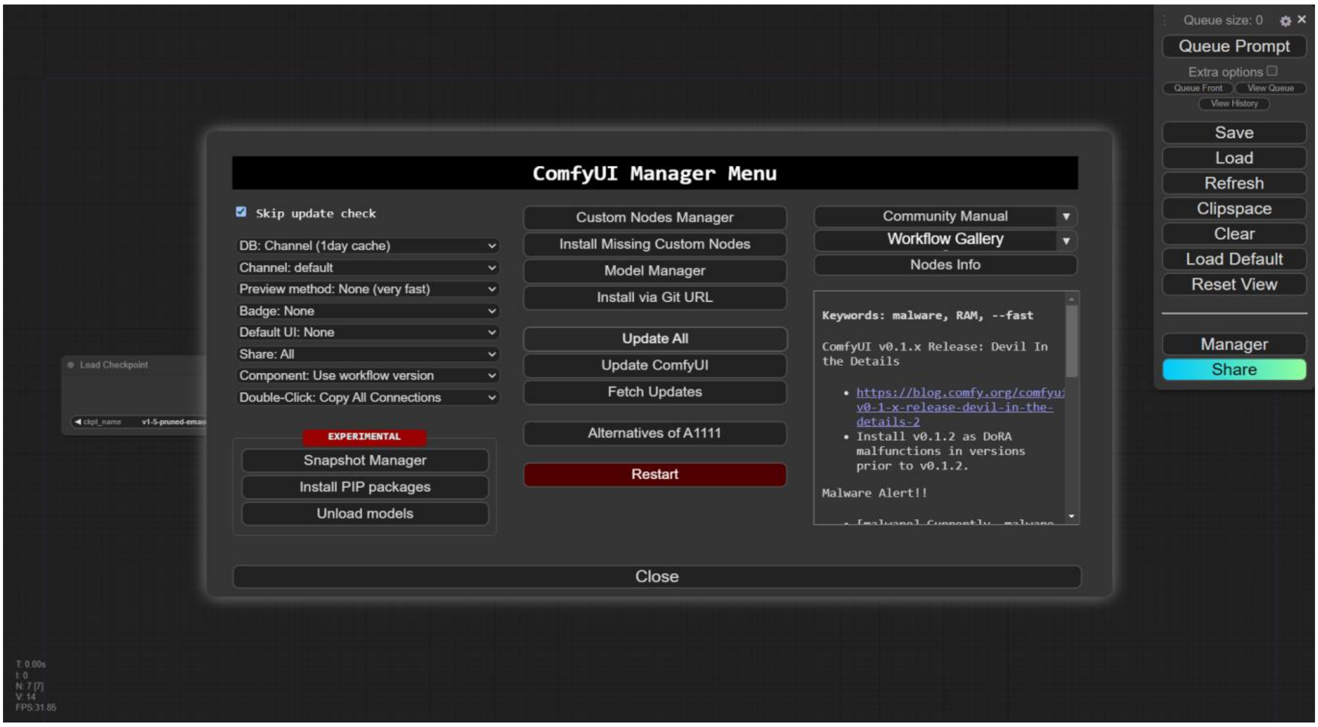
Task Instruction	Output
Generate an image of a whimsical fairytale castle perched on a cliff, overlooking a sparkling sea under a starry sky. The result should be a high-quality image.	
Generate an image of a bustling outdoor farmer's market at sunset, with rows of fresh produce, people shopping, and warm golden sunlight casting long shadows. The result should capture a vibrant, lively atmosphere with warm tones and fine details	
Network of interconnected nodes representing "Stable Diffusion" logo prominently displayed, abstract U-Net architecture diagram, CLIP text encoder visualization, diffusion process illustration, digital landscape with flowing data streams, vibrant neon colors, matrix-style code in background, neural network patterns, tensor flow diagrams, hyper-realistic render, highly Detailed, 4k resolution	
A sweeping shot of a sleek vintage sports car, its glossy black finish radiating golden highlights as it cruises along the Miami Beach coastal highway at sunset.	
Generate an image of a serene Swiss alpine landscape featuring snow-capped mountains, lush green meadows dotted with wildflowers, a crystal-clear lake reflecting the sky, and a small wooden chalet with smoke rising from the chimney. Include a backdrop of clear blue skies with soft clouds for a peaceful, picturesque ambiance	

Figure 9: illustrates the outputs generated by Stable Diffusion for various task instructions, showcasing the model's versatility in image synthesis.



☑ Skip update check

DB: Channel (1day cache)
Channel: default
Preview method: None (very fast)
Badge: None
Default UI: None
Share: All
Component: Use workflow version
Double-Click: Copy All Connections

Keywords: malware, RAM, --fast

ComfyUI v0.1.x Release: Devil In the Details

- https://blog.comfy.org/comfyui-v0-1-x-release-devil-in-the-details-2
- Install v0.1.2 as DoRA malfunctions in versions prior to v0.1.2.

Malware Alert!!

- [malware] Currently, malware

Figure 10: ComfyUI Workflow Manager Interface for Stable Diffusion

ComfyUI Manager:

While ComfyUI comes with a set of pre-built modules, its real power comes from its extensibility through custom modules.

ComfyUI Manager is a plugin that lets users manage and install custom modules directly within the ComfyUI interface. It provides an easy-to-use interface for browsing available modules, installing them with a single click, and integrating them into your workflows.

4.2 GitHub Link for Code:

https://github.com/zaidaanshiraz/Image_Generation_using_Stable_diffusion_and_ComfyUI

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

Prompt Optimization and Refinement

1. **Enhanced Prompt Engineering:**
   - Develop more robust algorithms for refining user prompts.
   - Improve the model's ability to interpret nuanced textual inputs.
2. **Adaptive Prompt Adjustment:**
   - Implement mechanisms that adjust prompt parameters based on real-time feedback.

Computational Efficiency Improvements

1. **Model Pruning and Quantization:**
   - Explore techniques to reduce the model's size without compromising quality.
   - Optimize the diffusion process to lower computational overhead.
2. **Hybrid Architectures:**
   - Investigate the integration of alternative model architectures to accelerate inference times.

Multi-Modal and Interactive Inputs

1. **Integration of Additional Inputs:**
   - Allow incorporation of sketches, voice commands, or other creative inputs alongside text.
   - Broaden the system's applications beyond text-to-image synthesis.
2. **Real-Time User Feedback Loop:**
   - Develop an interactive module that adapts the image generation process based on immediate user evaluations.

Advanced Evaluation and Adaptation

1. **New Evaluation Metrics:**
   - Establish both automated and user-centric metrics to better assess image quality.
   - Use these metrics to continually refine the generation process.
2. **Learning from User Interactions:**
   - Implement adaptive workflows that learn from accumulated feedback to personalize and improve output quality over time.

5.2 Conclusion

1. **Project Overview**
   - The project successfully integrates stable diffusion techniques with a user-friendly interface (ComfyUI) to generate high-quality images from text prompts.
   - The use of both a base and a refiner model ensures that the images are both structurally sound and rich in fine details.
2. **Key Strengths**

   2.1 High-Quality Image Generation:
   - The dual-model approach produces visually appealing and detailed outputs.

   2.2 User Accessibility:
   - ComfyUI makes the advanced technology accessible to users with varying levels of expertise.

   2.3 Creative Flexibility:
   - The system supports a wide range of artistic styles, from photorealism to abstract and surreal imagery.
3. **Challenges and Considerations**

   3.1 Prompt Sensitivity:
   - Small variations in text inputs can lead to significantly different outputs, highlighting the need for further prompt optimization.

   3.2 Computational Demands:
   - The iterative nature of diffusion models requires significant computational resources, which could limit real-time applications.
4. **Overall Impact**
   - This project lays a strong foundation for further innovations in generative art and image synthesis.
   - With future enhancements, the system promises to expand its versatility and efficiency, paving the way for new applications in art, design, and creative industries.

REFERENCES

[1] **Dhariwal, P., & Nichol, A. (2021).** *Diffusion models beat GANs on image synthesis*. Advances in Neural Information Processing Systems, 34, 8780–8794.

[2] **Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022).** *High-resolution image synthesis with latent diffusion models*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10695).S. Kakarla, P. Gangula, M. S. Rahul, C. S. C. Singh, and T. H. Sarma, "Smart Attendance Management System Based on Face Recognition Using CNN," in *2020 IEEE-HYDCON*, IEEE, Sep. 2020, pp. 1–5. doi: 10.1109/HYDCON48903.2020.9242847.

[3] **Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021).** *Zero-shot text-to-image generation*. In International Conference on Machine Learning (pp. 8821–8831). PMLR.

[4] **Liu, V., & Chilton, L. B. (2022).** *Design guidelines for prompt engineering text-to-image generative models*. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1–23).K. Painuly, Y. Bisht, H. Vaidya, A. Kapruwan, and R. Gupta, "Efficient Real-Time Face Recognition-Based Attendance System with Deep Learning Algorithms," in *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, IEEE, Jan. 2024, pp. 1–5. doi: 10.1109/IITCEE59897.2024.10467743.

[5] **Santana, G. (2022).** *Stable-diffusion-prompts*. Retrieved from https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts/blob/main/data/train.parquet

[6] **ComfyUI GitHub Repository. (n.d.).** *ComfyUI*. Retrieved from https://github.com/comfyanonymous/ComfyUI

[7] ***Ku, Sobania, D., Briesch, M., & Rothlauf, F. (2024).*** *ComfyGI: Automatic Improvement of Image Generation Workflows. arXiv preprint arXiv:2411.14193.*

[8] **Xue, X., Lu, Z., Huang, D., Wang, Z., Ouyang, W., & Bai, L. (n.d.).** *ComfyBench: Benchmarking LLM-based Agents in ComfyUI for Autonomously Designing Collaborative AI Systems. Shanghai Artificial Intelligence Laboratory.*

[9] **CivitAI. (n.d.). Retrieved February 9, 2025,** *from https://civitai.com/*

[10] **Zergtant. (n.d.). The complete guide to ComfyUI and Stable Diffusion: 1. Introduction & navigation. Medium. Retrieved February 9, 2025,** *from https://medium.com/@zergtant/the-complete-guide-to-comfyui-and-stable-diffusion-1-introduction-navigation-306796afa8a2*