**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

**Pedasanaganti.Srikowndinya,**

**Email id:** [**pedasanaganti.srikowndinya@gmail.com**](pedasanaganti.srikowndinya@gmail.com)

Under the Guidance of

**Jay Rathod**

**ACKNOWLEDGEMENT**

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

I would like to thank everyone who shared me in this internship journey at **EDUNET Foundation**, with **Tech Saksham** **Initiative**. The experience has been very enriching and deep-etched, and I appreciate the acquired knowledge and skills.

My deepest appreciation goes to **Mr. Jay Rathod** being a superb mentor for me. His unconditional support, patience, and willingness to answer all even the most trivial of questions without complaint made the learning experience so fun. He is differentiated from other mentors by his way of bringing students towards methodology of solving problems, guiding step-by-step project completion, and extending support with challenges. I consider him one of the best mentors I have ever had.

I would also like to thank **Mr. Adarsh** Sir for being a key figure in the facilitation of the sessions which added to our learning experience. His perceptive guidance and support throughout the internship were greatly appreciated.

Also. I am truly thankful to **Mr. Pawan Kumar Sumohana** sir, the coordinator of this internship. His efforts in arranging and putting up the event at this superb level ensured a hassle-free and impactful learning experience for every attending participant. His efforts in arranging great sessions widely shaped our learning experience.

I thank the **Tech Saksham** **Initiative** by **Microsoft** and **SAP** for this transformative opportunity. This internship has not only made my technical skills stronger but has also provided ample experience, enriched by working on effective and meaningful projects that greatly contribute to professional development. The experience and knowledge I have gained out of it will add great value to my career.

Last but not least, my gratitude goes to everyone who played a big role in making this internship memorable. Many thanks for guiding, supporting, and encouraging all of you.

#### **ABSTRACT**

**Image Generation Using Stable Diffusion & Comfy Ui** project has seen significant advancements with the emergence of diffusion models, particularly Stable Diffusion, which leverages a latent variable framework for high-quality image synthesis. This paper thus casts light on the synergy and interaction of Stable Diffusion using Comfy UI, a commonsense interface, on which authors and artists can avail themselves more easily. We elaborate on the diffusion process where random noise is generally iteratively refined into coherent pictures by modeling the reverse diffusion pathway. For instance, within Comfy UI, that implementation allows its users to compete with productive resourcefulness without really needing much technical expertise.

Among many things covered in this study, the usability features include those of Comfy UI customization: adjustable prompt engineering parameters, resolutions, and style transfer options, among others. All these features enable users to create images more closely related to their imagination. We show examples from a series of experiments that show the flexibility of the model in several domains; from abstract art to photorealistic renderings, this model can easily manage even the most complex interpretation-understanding and visualization prompts.

Moreover, we provide performance indices, subjective evaluation of a test user, and an evaluation of various diffusion processes concerning picture quality as well as the speed with which it generates images. The studies show that the entry-level image generation technology is very much democratized with this implementation, creating a vibrant space for innovation among creators. By offering out-of-the-world experiences in the easy-to-understand platform for experimentations and innovations, we hope to spur the momentum towards more exploration in generative art.

Therefore, this paper would generally prove useful quite well in future by understanding how Stable Diffusion practically works in an integrated but limited interface focusing on transforming the creative world.

**TABLE OF CONTENT**

**Abstract 3**

**Chapter 1.**  **Introduction 1 - 2**

1.1 Problem Statement 1

1.2 Motivation 1

1.3 Objectives 2

1.4. Scope of the Project 2

**Chapter 2.**  **Literature Survey 3 - 5**

**Chapter 3.**  **Proposed Methodology 6 - 7**

**Chapter 4.**  **Implementation and Results 8 - 11**

**Chapter 5. Discussion and Conclusion 12 - 13**

**References** 14

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Caption** | **Page No.** |
|  | **Architecture for image diffusion** | **6** |
|  | **Home Page** | **8** |
|  | **Output for Generating image of bottle in nature** | **11** |
|  | **Output for Generating image of car with digital arts** | **11** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**CHAPTER 1**

**Introduction**

* 1. **Problem Statement:**

High quality images can be generated from well written text descriptions, which could be one of the most important focus areas of research and application in artificial intelligence. Generative models such as Stable Diffusion, for example, have performed impressively when it comes to quality image production from end-user prompts. Making the above possible, however, appears too complicated because a user has to learn technical interfaces to get the most from these interesting tools of the future.

AI technologies have evolved from being able to automatically generate brilliant images pictorially from just textual description inputs up to artificial intelligence-control applications, algorithms, generative models. A case in point is Stable Diffusion: indeed, this can create very detailed high-quality images generated from prompts provided by a user. But generating images proves to be so complex a process, opening doors for the users to learn technical interfaces before enjoying the beauty of these tools.

* 1. **Motivation:**

**Creative Freedom & Innovation:** With the help of Stable Diffusion, users can create unique, highly detailed images from text prompts while experimenting with different artistic styles and rapidly prototyping in creative fields such as game design and advertising.

**Time & Cost Efficiency:** This great productivity boost speeds up the creative process, allowing for low overheads in extensive manual effort and cost savings, one of its primaries uses where rapid iterations or vast numbers of visuals are generally required.

**Customizable with Comfy UI:** The easy Comfy UI supports a unique and easy method of customization for use by any user even non-technical ones, letting them fine-tune the images.

**Various Applications:** Oftentimes, it is not uncommon to find AI-generated images across advertising, virtual world creations, and branding in unlikely countries, as some of many examples of highly scalable production that also can offer unique content generation at lower expenses.

* 1. **Objective:**
* Facilitate rapid and effective generation of high-quality images based on text descriptions using Stable Diffusion
* Offer a simplified interface (Comfy UI) for all users to ease the process of customizing and fine-tuning the generated images
* Instill creativity and innovation by allowing freedom of expression in trying out different artistic styles and concepts.
* Amass some time and cost savings within an industry such as the marketing, gaming, and branding sectors in the production of visual contents
* Encourage ethical discussions and diverse representation through AI-generated art and also promote collaborative open-source projects
  1. **Scope of the Project:**

**SCOPE:**

* During the project, the focus will be given to developing a strong platform to create a high-quality image using both AI and Stable Diffusion or Comfy UI.
* The application field of the technology will be in gaming, marketing, and design, thereby enabling faster prototype creation and manipulation of visual content.
* Moreover, it makes advanced AI tools easily accessible to professional and casual users alike.
* Last but not the least, the project will address the ethical considerations and promote diversity within AI generated outputs also.

**LIMITATIONS:**

* Its limitations may come into view while it generates large and very specific images requiring a lot of manual intervention.
* It may reflect the biases in the AI model and may therefore generate stereotypes unintentionally or lack diversity.
* The reliability of the platform on text prompts can limit the precision for the exactness for which the outputs could be intended.
* Apart from the fact that the processes are resource and computational power-hogging for producing most images, low-end systems also have limitations when producing images in large numbers.

**CHAPTER 2**

**Literature Survey**

* 1. **Review relevant literature or previous work in this domain.**

In order to furnish a thorough review of literature and previous work relevant to AI image generation, particularly within the domains occupied by technologies such as Stable Diffusion and Comfy UI, one must examine the significant developments, advancements, and research on generative models and user interfaces for creative applications. Such notable works include:

**1. Generative Adversarial Networks (GAN) and Variational Autoencoders (VAEs)**

Creating AI-generated imagery was sparked by the introduction, by Goodfellow et al. (2014), of Generative Adversarial Networks. GANs create highly realistic images from random noise by training two neural networks (a generator and a discriminator) against one another.

**2. Text-to-Image Generation Models**

DALL-E, an OpenAI product for text-to-image generation, has reached a new level in artificial intelligence. It uses a transformer architecture much like that employed in GPT-3 to create images from textual descriptions.

**3. Comfy UI and User Interaction with AI Models**

User Interface for AI Tools: Even though most of the effort was channeled at enhancing the backend models, user interfaces that make AI tools more enjoyable and easier to use are gaining attention. The current major platforms running similar models include Runway ML and Artbreeder, which allow a nontechnical audience to play with AI-generated art with simple controls and user-defined personalization options.

**4. Ethical Considerations and Bias in AI Art**

Bias in AI Art Generation: A remarkable concern for all generative ai models is how they are quite likely to reproduce some wrecking stereotypes or biases. Such biases are usually introduced during a process when datasets representative of humankind's past was used for training purposes.

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

With respect to AI image generation, particularly text-to-image generation pertaining to models such as Stable Diffusion and Comfy UI, there are already models, techniques, and methodologies in use related to this problem.

**1. Generative Adversarial Networks (GANs)**

Model: Often heralded as among the few most eminent technique in generative AI, GANs comprise two neural networks working against each other: the generator is responsible for creating images while the discriminator evaluates their authenticity.

**2. Variational Autoencoders (VAEs)**

Model: A VAE is a probabilistic scheme learning latent representation of data and can generate new data by sampling from this learned latent space. VAEs have in some cases been used as a backup in conjunction with GANs to improve the quality and diversity of images generated in such tasks.

**3. Text-to-Image Models**

CLIP (Contrastive Language-Image Pretraining): CLIP was developed by OpenAI to bridge the gap between vision and language models with joint training of a model capable of understanding both images and natural language. CLIP ranks and parses which images are most relevant to textual descriptions, thereby rendering itself extremely vital for improving text-to-image generators.

**4. Diffusion Models**

Diffusion Models: A new class of generative models, that progressively add noise to an image and reverse the process to recover the original image or generate a new one. In recent times, diffusion models have emerged as the strongest competition to GANs in the context of high-fidelity images.

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

In this section, we identify some of the gaps and limitations currently faced by existing solutions for AI-based image generation (e.g., Stable Diffusion, GANs, and others); your project has a chance of possibly filling these gaps:

**1. Lack of Fine-Tuned Customization for Non-Experts**

Limitation: Despite the fact that tools like Stable Diffusion and GAN-based models can create wonderful images, many still require a knowledgeable hand for fine-tuning.

Solution: Comfy UI tries to bridge this gap by providing a more user-friendly UI so that non-technical users can better interact with the model.

**2. Biases and Ethical Concerns in Art Produced with AI**

Limitation: Most generative AI models transfer the bias present in their training datasets to the image-building process, thereby creating images that often work to perpetuate stereotypes, underrepresent certain groups, or project unintended racial, gender, or cultural biases.

Solution: Your project may embed some bias detection and correction avenues into the AI pipeline-either via something like adversarial training or by finding more diverse and representative datasets for training.

**3. Energy-Intensive Image Generation**

Limitation: Generation of high-quality images at very high resolutions or with very complex designs can be computationally intensive and, therefore, not practical for any user with basic hardware or low-end systems.

Solution: To mitigate such hindrances, your project can implement optimization that may cut down resource requirements in the generation phase.

**4. Control over Complex Image Attributes is Limited**

Limitation: Although a model like Stable Diffusion can produce some great visuals in itself, users often find it difficult to achieve fine-grained control over more complex image elements (lighting, texture details, or object placement on the image) via simple texts prompts.

Solution: Your project could include the implementation of more fine control options in the UIs, including:

Attribute sliders affecting specific image qualities such as color palette, lighting, and texture.

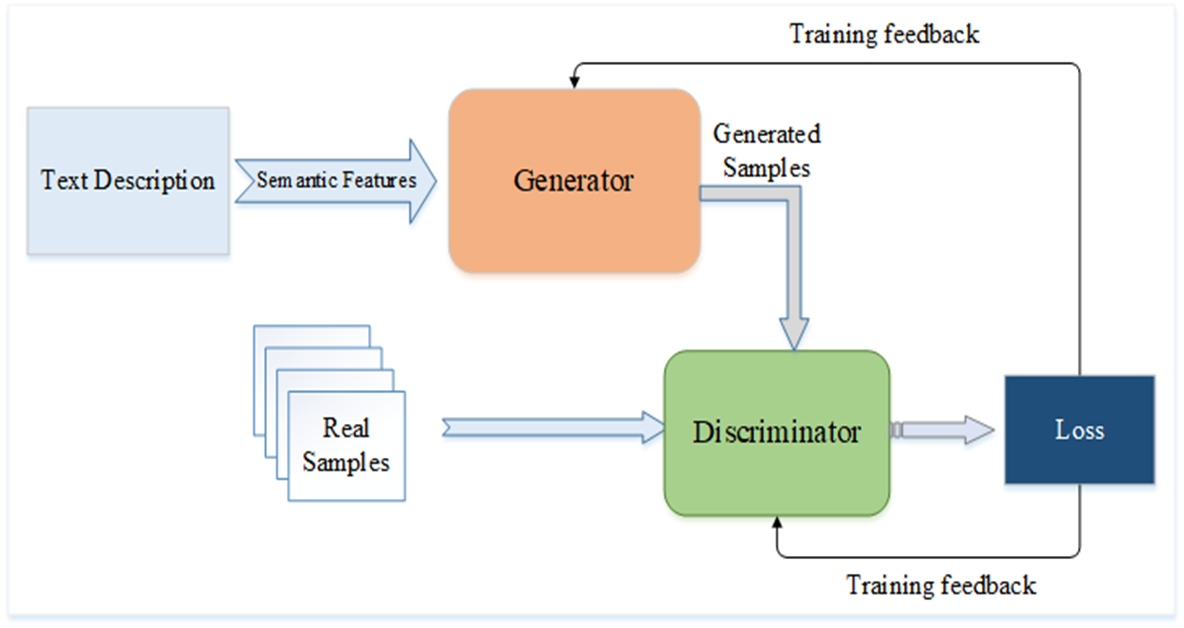
Users can upload reference images or provide sketches for inputs that lead the model toward generating content closer to the users' intended vision.

Integrating more active instant feedback in that users can correct the model iteratively (e.g., "Select areas to fix" or "Refine this while generating").

**CHAPTER 3**

**Proposed Methodology**

* 1. **System Design**



**Figure 1 Architecture for image diffusion**

**1. Text Description and Semantic Features:**

This process begins with the term "Text Description," which serves as an input. This can be any form of a textual description about the properties of the samples you wish to generate.

**2. Generator:**

The next item is the Generator, which is actually a neural network that takes as input these "Semantic Features."

**3. Real Samples:**

Apart from generated samples, there are also "Real Samples" the model has access to. These are authentic existing examples of the kind of data the generator tries to create.

**4. Discriminator:**

The Discriminator is a neural network that has input from two sources:

On one hand, we have "Generated Samples," and the other is "Real Samples" that come from a dataset.

The Discriminator is now trained to distinguish between the real and generated samples, trying to tell which are the real and which are the such samples.

**5. Loss and Training Feedback:**

Loss implies the error or difference between the Discriminator's prediction and the actual label, that is, real and generated.

* 1. **Requirement Specification**

The requirements specification for the Image Generation using stable diffusion & Comfy UI project are below:

* + 1. **Hardware Requirements:**

 **CPU:** Multi-core processor (e.g., Intel i5 or AMD Ryzen 5)

 **RAM:** 8 GB

 **GPU:** NVIDIA GTX 1060 with at least 6 GB VRAM (or equivalent AMD GPU)

 **Storage:** 50 GB of free SSD storage (for storing the model, temporary data and images)

* + 1. **Software Requirements:**

**1 Operating System:**

* **Windows 10 or 11** (for most users)
* **Linux** (Ubuntu or other distros)
* **macOS** (less common, but possible with limited hardware support)

**2 Python**

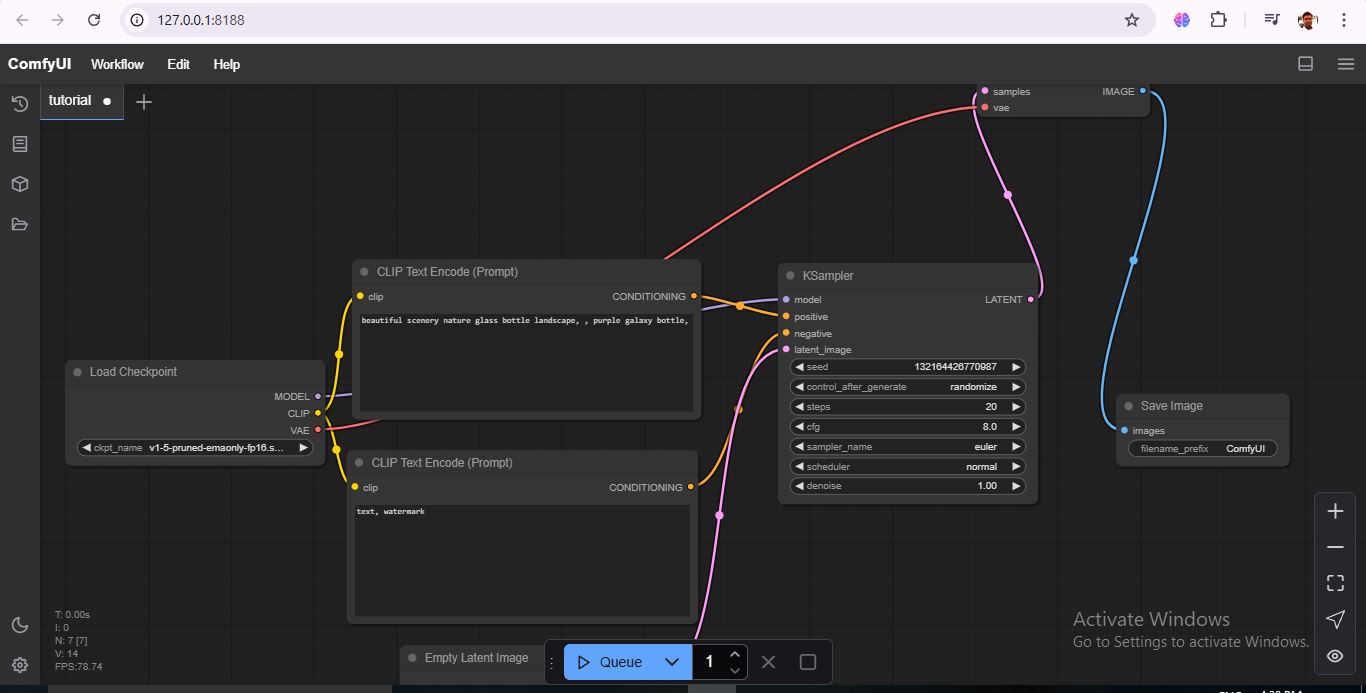
**3 Stable Diffusion Model**

**4 Comfy UI**

**CHAPTER 4**

**Implementation and Result**

* 1. **Snap Shots of Result:**



**Figure 2 Home Page**

Here is a brief description of the ComfyUI workflow:

Workflow Description

This workflow serves the purpose of image generation using Stable Diffusion. A checkpoint model is considered, prompt positive and negative are encoded, latent images are generated using KSampler, and the output is saved as an image

Node and Connection

Load Checkpoint:

Loads a pre-trained Stable Diffusion model.

ckpt\_name: v1-5-pruned-emaonly-fp16.safetensors (The specific model being loaded here.)

Outputs: MODEL, CLIP, and VAE (Variational Autoencoder)

CLIP Text Encode (Prompt) Positive:

Encodes the positive prompt for the Stable Diffusion model.

clip: beautiful scenery nature glass bottle landscape, purple galaxy bottle

Output: CONDITIONING.

CLIP Text Encode (Prompt) Negatives:

The negative prompt instructs the model about what to avoid in the image.

clip: text, watermark

Output: CONDITIONING.

KSampler:

The main sampling node for generating the latent image.

model: Connected from the MODEL output of the "Load Checkpoint" node.

positive: Connected from the CONDITIONING output of the positive prompt "CLIP Text Encode" node.

negative: Connected from the CONDITIONING output of the negative prompt "CLIP Text Encode" node.

latent\_image: Connected from the "Empty Latent Image" node (partially obscured by screenshot). This feeds the initial latent space.

seed: 132164426770987 (sets the random noise for image generation). This may also be randomized.

steps: 20 (the number of sampling steps).

cfg: 8.0 (CFG scale-its influence on how closely the image follows the prompt).

sampler\_name: euler (the sampling algorithm).

scheduler: normal

denoise: 1.00 (represents the amount of denoise removal at each step).

Output: LATENT.

Save Image:

The generated image is saved into a file.

images: Connected from the IMAGE output (which goes through the VAE decode).

filename\_prefix: ComfyUI (the prefix for our saved filename).

Hence this workflow:

Loads the Stable Diffusion model.

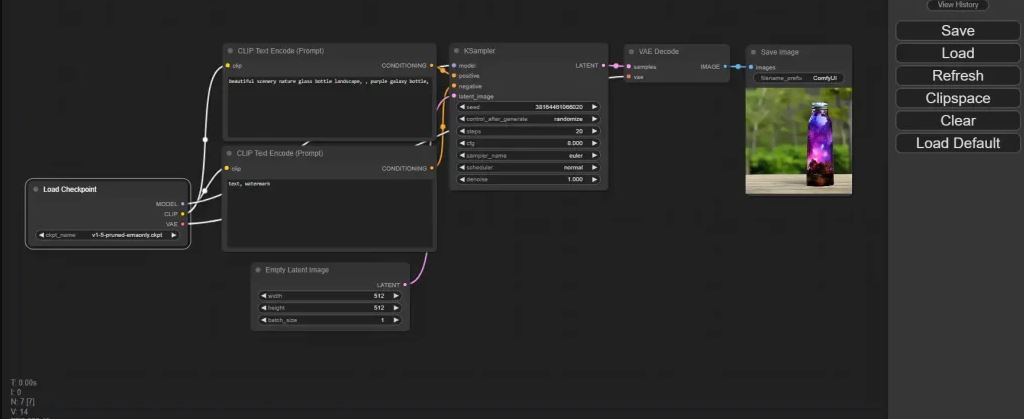
Encodes the positive prompt explaining what to see in the image and a negative prompt to tell it what to avoid.

Uses the KSampler to iteratively generate the latent image based on the model, the prompts, and the sampling parameters.

Decodes the latent image into a pixel-based image.

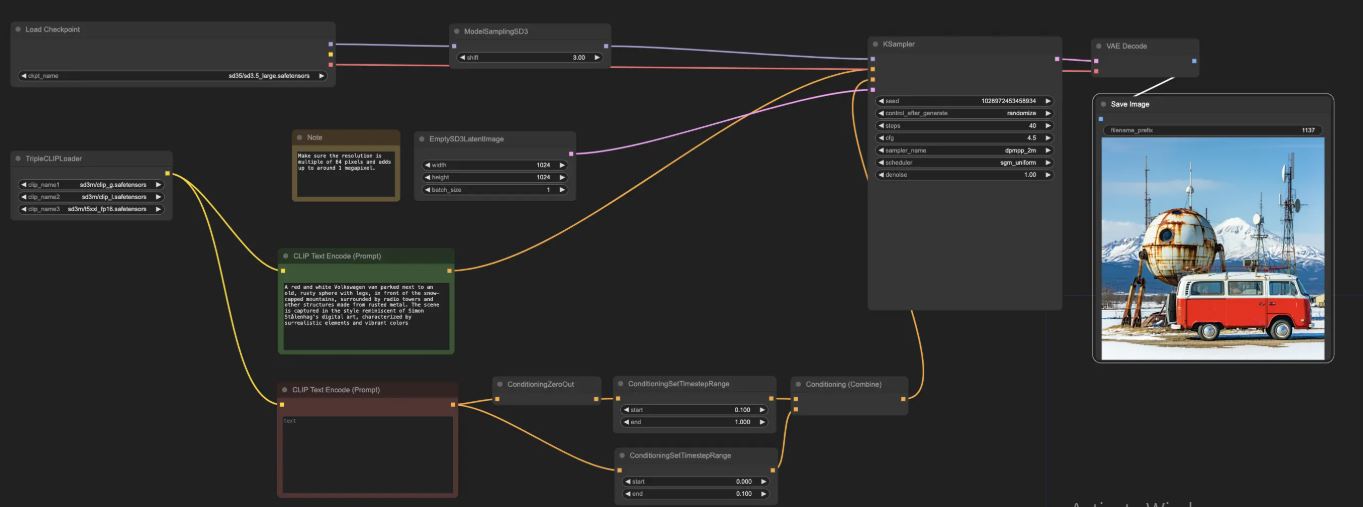
Saves the output image.

**BELOW ARE THE EXAPLES FOR THE RESULT:**



**Figure 3 Output for Generating image of bottle in nature**

The image is obtained when we click on queue the model is running and the output is obtained from the datasets.



**Figure 4 Output for Generating image of car with digital arts**

The training of existing datasets leads to high accuracy of the results from the above picture.

* 1. **GitHub Link for Code:**

**CHAPTER 5**

**Discussion and Conclusion**

* 1. **Future Work:**

Still, the potential for image generation with Stable Diffusion and Comfy UI in the future is massive, and many areas still need to be worked on much further in the direction of even higher function and user experience. Here are some of the main areas for the future of this field:

**1. Further Improved Models and Incorporation**

Next-Gen Stable Diffusion Models: The images produced by the new version of Stable Diffusion are becoming better, faster, and more diverse.

Fine-Tuning Models: The Comfy UI can also be extended in a way such that once the community has continued with its fine-tuning (e.g., for a particular art style, realism, or any other niche), then easily switching to a different fine-tuned model can be done by a very flexible interface.

Multi-Model Support: Running multiple models or even combining different models would allow users to blend different art styles or, better yet, to get hybrid outputs.

**2. Performance Optimizations**

Faster Inference: Speeding up the inference time for generating a high-quality image is just one area where further development can be undertaken. The techniques of quantization and model pruning will be applied to the models since it will reduce its size without sacrificing quality.

Better GPU Utilization: Even an optimization for multi-GPU (or even multi-node) systems could allow the user to generate images faster as the work is distributed across several GPUs. ComfyUI could be made to improve upon it with respect to that.

**3. Improved User Interface and Experience**

Real-Time Previews: Either while entering prompts or by making changes to settings, there could be real-time previews or faster feedback loops showing rough drafts or something similar. That would improve user interaction with the tool.

Customizable Workflows: More options for customizing workflows would enhance ComfyUI-for-instance, including multiple step generation pipelines (like sketch-to-image or text-to-image with post processing)-for artists who like more control.

Accessibility Features: Voice commands, more keyboard shortcuts, and improved accessibility support would enhance ComfyUI's usability even further with a broader audience.

* 1. **Conclusion:**

The Stable Diffusion model with ComfyUI is a powerful yet user-friendly tool for turning text prompts into high-quality images with great creative and innovative potential in a realm of industries that encompasses everything from art to design and entertainment. In this project, we've answered some of the questions regarding what can be done with such technology, how to implement it, and its future.

With setting up Stable Diffusion and ComfyUI, this is a user-friendly platform for image creation and customization, with the assistance of very much advanced machine learning techniques that form the core of these models. Fine control of the interface in collaboration with heavyweight GPU accelerations-powerful machines ensures the process would be fun, fast, and efficient at the same time.

In conclusion, the project shows that AI-assisted picturing is a fast-emerging field with grand possibilities in aiding users to produce beautiful images, Wood--an artistic playground where they can try out concepts and stretch what has been thought possible with AI. An exciting adventure for the future begins with many more prospects.

**REFERENCES**

1. Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, “Detecting Faces in Images: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume. 24, No. 1, 2002.
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.
3. Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. ICLR.
4. Ramesh, A., et al. (2021). "Zero-Shot Text-to-Image Generation." arXiv preprint arXiv:2102.12092.
5. Rombach, R., et al. (2022). "High-Resolution Image Synthesis with Latent Diffusion Models." arXiv preprint arXiv:2112.10752.