Image Classification and generation of images using prompts with the help of Generative AI

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

In

Information Technology

by

Nihal Raj

20BIT0417

Under the guidance of

Prof. / Dr. Agilandeeswari L.
SCORE

VIT, Vellore

VIT®

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

May, 2024

Executive Summary

In today's digital era, the demand for high-quality visual content is ever-growing across various domains, from entertainment and advertising to design and education. In this context, the integration of advanced AI techniques holds the promise of revolutionizing the way we generate and manipulate images. One such groundbreaking technology is the Stable Diffusion Model, which offers a robust framework for image generation and modification through artificial intelligence.

In my thesis, I implemented the approach utilizing the Stable Diffusion Model to generate images through AI. The architecture is structured to seamlessly integrate user feedback and enhance image generation. Initially, the Stable Diffusion Model is employed for text-to-image generation, producing a pool of images. Through an Interactive Evolutionary Algorithm (IEA), user feedback is utilized to select the most suitable images, optimizing the quality of generated content.

Furthermore, the system offers the capability to modify existing images through an image-to-image Stable Diffusion Model, enabling users to create variations or modifications. Finally, the interface provides tools for post-processing, including adjustments for brightness, contrast, sharpness, and other parameters, empowering users to tailor the images to their specific preferences and needs. This comprehensive framework not only facilitates high-quality image generation but also empowers users to actively participate in the creative process, resulting in personalized and refined outputs.

Contents

List of Figures

List of Abbreviations

FID	Frechet Inception Distance
PSNR	peak signal-to-noise ratio
NLP	Natural language processing
SSI	Structural similarity index
AI	Artificial Intelligence

Symbols and Notations

δf	CFO
ε	NCFO

1. Introduction

This thesis embarks on exploring the transformative potential of the Stable Diffusion Model in image generation and manipulation. Key objectives include developing an integrated architecture to seamlessly merge the model with interactive components, empowering users in text-to-image generation, and integrating user feedback through an Interactive Evolutionary Algorithm (IEA). Further, the project extends to image-to-image transformations and offers post-processing options for user customization. By addressing these goals, this project aims to advance AI-driven image generation techniques, with the potential to revolutionize various industries and domains.

Utilizing Stable Diffusion models in image generation offers several compelling advantages. Firstly, these models provide a stable and reliable framework for generating high-quality images. By leveraging diffusion processes, Stable Diffusion models can effectively capture complex image features and generate diverse outputs without suffering from issues like mode collapse or training instability, commonly encountered in traditional methods such as GANs. This stability ensures that the generated images exhibit consistent quality and fidelity, making them suitable for various applications ranging from digital art to medical imaging. Additionally, Stable Diffusion models enable fine-grained control over the generation process, allowing users to adjust parameters such as image resolution, style, and content, thereby facilitating greater customization and personalization of the generated outputs. [Reference: 4]

The Iterative Evolutionary Algorithm (IEA) offers a powerful mechanism for refining and enhancing the quality of generated images. Unlike traditional optimization methods, the IEA operates by iteratively evolving a population of candidate solutions based on user feedback and preferences. This iterative approach enables exploration of the image space, facilitating the discovery of diverse and high-quality solutions that better align with user expectations. Moreover, the IEA introduces diversity and novelty into the generated images through genetic operators such as mutation, ensuring that the outputs remain diverse and relevant. By incorporating the IEA into the image generation process, users can actively participate in shaping the final outputs, leading to more satisfying and customized results that meet their specific requirements and preferences.

There are a lot of application for such technology and image generation such as:

Data Augmentation: Image generation can be used to augment datasets, especially in scenarios where labelled data is scarce. Generated images can be used to increase the diversity of the dataset, which often leads to better generalization and performance of machine learning models.

Creative Applications: Image generation techniques have found applications in various creative fields, including art, design, and entertainment. Artists and designers can leverage these tools to generate novel visual content or aid in the creative process.

Data Imputation and Denoising: Generated images can be used to fill in missing or corrupted parts of images, a process known as data imputation. Additionally, image generation techniques can be used for denoising, where noisy images are cleaned up or enhanced. [Reference: 7]

1.1 Objective

This thesis endeavors to delve into the capabilities of the Stable Diffusion Model and its potential to reshape the landscape of image generation and manipulation. The objectives are detailed as follows:

- Development of an Integrated Architecture: The primary aim is to construct a cohesive and adaptable framework that seamlessly incorporates the Stable Diffusion Model with interactive components. This architecture should facilitate smooth communication between the AI system and the user, enabling intuitive feedback mechanisms and enhancing the overall user experience.
- Text-to-Image Generation: By harnessing the power of the Stable Diffusion Model, the project endeavors to establish a robust text-to-image generation system. The objective is to empower users to effortlessly translate textual descriptions into vivid and contextually relevant visual representations, thereby expanding the accessibility and applicability of AI-generated content.
- User Feedback Integration: The project is the integration of an Interactive Evolutionary Algorithm (IEA) to gather and assimilate user feedback. This objective seeks to optimize the image selection process by leveraging user preferences, thereby refining the output quality, and ensuring alignment with user expectations.
- Image-to-Image Generation: Building upon the foundational text-to-image generation, the

project aims to extend the capabilities of the system to accommodate image-to-image transformations. By employing the Stable Diffusion Model in this context, users will have the flexibility to manipulate existing images, fostering creativity, and enabling the creation of personalized visual content.

- Post-Processing Options: The project endeavors to provide users with a suite of post-processing tools, including adjustments for brightness, contrast, sharpness, and other parameters. These options are designed to empower users to fine-tune and customize the generated images according to their specific preferences and requirements, thereby enhancing user satisfaction and usability.

Through the pursuit of these objectives, this project aspires to contribute to the advancement of AI-driven image generation and manipulation techniques, offering novel insights and practical solutions that have the potential to revolutionize various industries and domains.

1.2 Motivation

This project is driven by the realisation of the ever-growing significance of pictures in our digital age, as well as the ongoing difficulties in producing and modifying them successfully. Conventional approaches lack user engagement and adaptability and frequently struggle to create diverse, high-quality photographs at scale. Artificial intelligence (AI)-driven solutions have a chance to transform the process by filling this gap.

One limitation in the existing approaches is the lack of direct user control and feedback in the generation process. While they can produce impressive results autonomously, they often struggle to capture subtle nuances or specific preferences that users may have. Additionally, these models may suffer from mode collapse or produce outputs that do not fully align with user expectations. By integrating user rating and IEA into the image generation process, we can address these limitations and create a more user-centric approach to image synthesis.

The motivation behind embarking on this image generation project stems from the desire to harness cutting-edge technologies to create visually stunning and customized imagery. By integrating text-to-image and image-to-image generation models with an Iterative Evolutionary Algorithm (IEA), we aim to push the boundaries of creativity and personalization in image synthesis. The IEA algorithm plays a pivotal role in this endeavour by injecting controlled noise into the image generation process, facilitating exploration and diversity within the image space. This diversity is essential for producing a wide range of visually appealing outputs that cater to diverse user preferences and requirements. Additionally, the adoption of a stable diffusion model adds further importance to the project by ensuring smooth and high-quality image generation, free from artifacts and inconsistencies. Ultimately, the motivation behind this project lies in harnessing the power of advanced algorithms and models to unlock new possibilities in image generation, offering users unparalleled levels of customization and creativity.

One approach that seems particularly promising for dealing with these issues is the Stable Diffusion Model. The scientific world has taken notice of it because of its capacity to record intricate data distributions and produce lifelike visuals. With this model, we want to expand the realm of possible AI-based picture synthesis and provide a high-fidelity output, scalable solution.

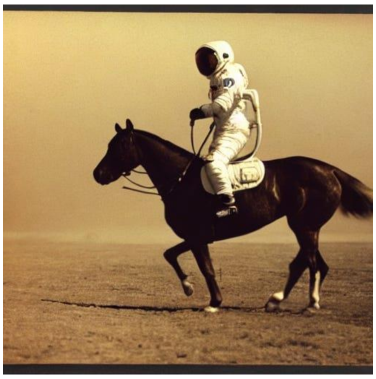


Figure 1: Astronaut on a horse image, generation using stable diffusion

In addition, the inclusion of interactive features like post-processing choices and user feedback methods not only improves the quality of created photographs but also democratises the production process. We hope to close the gap between technological advancement and user demands by giving users the ability to actively contribute to an impact picture formation, which will eventually promote a more cooperative and user-centric approach to image synthesis.

To promote the broad acceptance and implementation of these technologies across several areas, we want to further the development of AI-driven picture production and modification techniques through this thesis.

1.3 Background

The universality of digital images in a variety of industries, including advertising, design, entertainment, and education, highlights how vital it is to advance picture creation and manipulation technology. The diversity and scalability of created material are limited by traditional approaches, which frequently rely on predetermined templates or manual inputs. Moreover, these methods usually provide static and impersonal outputs since they are unable to adjust to changing situations or user preferences.

The development of artificial intelligence (AI) in recent years has created new opportunities for the synthesis and modification of images. AI-driven methods, especially those that make use of deep learning techniques, can produce large quantities of extremely realistic and contextually relevant photographs. The Stable Diffusion Model is one of these methods that has attracted a lot of interest because of its capacity to represent intricate data distributions and produce high-quality images.

Building on the ideas of diffusion processes, the Stable Diffusion Model describes the creation of pictures as a set of random modifications made to a basic distribution. Through iteratively dispersing noise, the model produces pictures with progressively realistic characteristics, attaining cutting-edge results in image synthesis assignments.

Considering this, the purpose of this project is to investigate how the Stable Diffusion Model may be applied to transform the processes involved in creating and modifying images. Our goal is to enable users to actively engage in and control the picture production process while also improving the quality and diversity of created photographs through the development of an integrated framework that includes post-processing choices and user feedback systems. We hope that our research will develop AI-driven picture creation methods and open new opportunities for innovative applications across various domains.

What is diffusion?

Figure 2. Demonstration of diffusion [reference:11]

In the context of stable diffusion in machine learning, “diffusion” refers to the process of gradually adding noise or perturbations to an input signal, such as an image. This process is typically done over multiple steps or iterations. The goal of diffusion is to refine the understanding of the input signal while maintaining stability and coherence.

So how stable diffusion works while training an ML model with an input image?

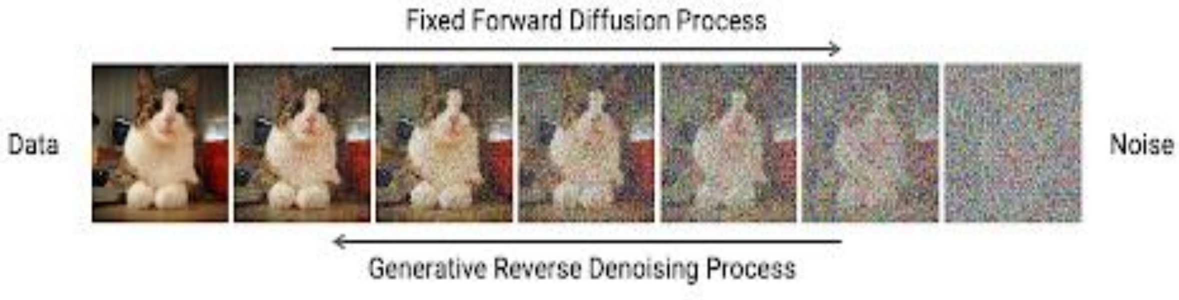

Fixed Forward Diffusion Process
Data
Noise
Generative Reverse Denoising Process

Figure 3. Adding noise to a give input image. [Referance:12]

The above image of a cat is a training image which is given as an input with a text, usually we name it as an “alt text” which describes the object in the image precisely. Now, as we can see the sequence of adding noise bit by bit with each alteration to completely distort the image with 100% noise.

The generative denoising process typically occurs through a combination of techniques, such as deep learning models and optimization algorithms. A simplified overview of how it works is:

Input Image: The process starts with an input image that may be corrupted by noise or artifacts.

2. Generative Model: A generative model, often based on neural networks, is employed to denoise the input image. This model learns to understand the underlying structure and features of the image data.

Training Phase: The generative model is trained on a dataset of clean images paired with their noisy counterparts. During training, the model learns to map noisy images to their clean versions by minimizing a loss function that measures the difference between the generated output and the ground truth clean image.

Inference Phase: Once the model is trained, it can be used to denoise new, unseen images. During inference, the generative model takes a noisy image as input and produces a denoised version as output.

Post-processing: Sometimes, additional post-processing techniques may be applied to further enhance the quality of the denoised image. These techniques could include filtering, smoothing, or sharpening operations.

By iteratively improving the generative model through training on large datasets, the denoising process can effectively remove noise and artifacts from images, resulting in cleaner and visually appealing outputs.

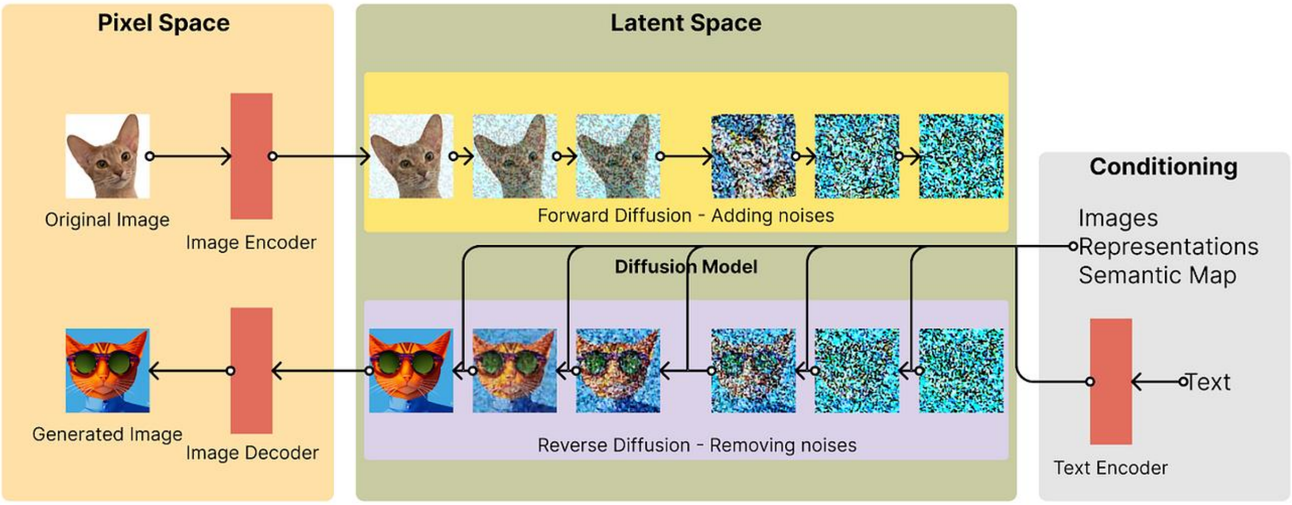


Figure 7. Architecture of stable diffusion model using example [Referance: 13]

2. Literature Survey

Ref.	Author	Methodology	Performance
1.	Tingting Qiao, Jing Zhang, Duanqing Xu, Dacheng Tao	Using generative adversarial networks, guaranteeing semantic consistency between the text description and visual content	Thorough experiments on two public benchmark datasets demonstrate the superiority of MirrorGAN over other representative state-of-the-art methods. FID score around 20.4
2.	Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng	CogView, a 4-billion-parameter Transformer with VQ-VAE tokenizer to advance this problem.	Achieves the state-of-the-art FID on the blurred MS COCO dataset, outperforming previous GAN-based models.
3.	Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin	Stable Diffusion with transformer	Model achieves state-of-the-art FID and human evaluation results, unlocking the ability to generate high fidelity images in a resolution of pixels.
4.	Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee	Cross-Modal Contrastive Generative Adversarial Network (XMC-GAN)	Improved state-of-the-art FID from 48.70 to 14.12.
5.	Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss	Based on a transformer that autoregressively models the text and image tokens as a single stream of data.	Approach is competitive with previous domain-specific models when evaluated in a zero-shot fashion.
6.	Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, Xin Tong	Embed 3D priors into adversarial learning and train the network to imitate the image formation of an analytic 3D face deformation	Improved state-of-the-art FID from 30.20 to 19.25.
7.	Shuyang Gu, Dong Chen,	Based on a vector quantized variational autoencoder (VQ-VAE) whose latent	Compared with previous GAN-based text-to-image methods, it

	Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen	space is modeled by a conditional variant of the recently developed Denoising Diffusion Probabilistic Model (DDPM)	VQ-Diffusion can handle more complex scenes and improve the synthesized image quality by a large margin
8.	Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros	Self-guidance, a method that provides precise control over properties of the generated image by guiding the internal representations of diffusion models.	The flexibility and effectiveness of self-guided generation through a wide range of challenging image manipulations, such as modifying the position or size of a single object (keeping the rest of the image unchanged), merging the appearance of objects in one image with the layout of another, composing objects from multiple images into one, and more.
9.	Omer Bar-Tal, Lior Yariv, Yaron Lipman, Tali Dekel	MultiDiffusion, a unified framework that enables versatile and controllable image generation, using a pre-trained text-to-image diffusion model, without any further training or finetuning.	MultiDiffusion can be readily applied to generate high quality and diverse images that adhere to user-provided controls, such as desired aspect ratio (e.g., panorama), and spatial guiding signals, ranging from tight segmentation masks to bounding boxes.
10.	Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, William W. Cohen	SuTI, a Subject-driven Text-to-Image generator that replaces subject-specific fine tuning with {in-context} learning.	SuTI can generate high-quality and customized subject-specific images 20x faster than optimization-based SoTA methods. On the challenging DreamBench and DreamBench-v2, our human evaluation shows that SuTI significantly outperforms existing models like

			InstructPix2Pix, Textual Inversion, Imagic.
11.	Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh	SpaText --- a new method for text-to-image generation using open-vocabulary scene control.	In addition to FID scores and a user study, to evaluate the method and show that it achieves state-of-the-art results on image generation with free-form textual scene control.
12.	Yaru Hao, Zewen Chi, Li Dong, Furu Wei	Prompt adaptation, a general framework that automatically adapts original user input to model-preferred prompts. Use of reinforcement learning to explore better prompts.	Method outperforms manual prompt engineering in terms of both automatic metrics and human preference ratings.
13.	Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang	Muse: Text-To-Image Generation via Masked Generative Transformers	The Muse 3B parameter model achieves an FID of 7.88 on zero-shot COCO evaluation, along with a CLIP score of 0.32.
14.	Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, Song Han	FastComposer uses subject embeddings extracted by an image encoder to augment the generic text conditioning in diffusion models, enabling personalized image generation based on subject images and textual instructions with only forward passes.	It achieves 300×-2500× speedup compared to fine-tuning-based methods and requires zero extra storage for new subjects.
15.	Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang,	proposed a learning-based encoder, which consists of a global and a local mapping networks for fast and accurate customized text-to-image generation.	Compared the method with existing optimization-based approaches on a variety of user-defined concepts, and demonstrate that the method enables

	Wangmeng Zuo		highfidelity inversion and more robust editability with a significantly faster encoding process.
16.	Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, Jiaying Liu	Unified Multi-Modal Latent Diffusion (UMM-Diffusion) which takes joint texts and images containing specified subjects as input sequences and generates customized images with the subjects.	By leveraging the large-scale pre-trained text-to-image generator and the designed image encoder, the method is able to generate high-quality images with complex semantics from both aspects of input texts and images.
17.	Anjita Naik , PictureBesmira Nushi	Two popular T2I models: DALLE-v2 and Stable Diffusion.	An analysis of geographical location representations on everyday situations (e.g., park, food, weddings) shows that for most situations, images generated through default location-neutral prompts are closer and more similar to images.
18.	Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, Ping Luo	RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Paths.	RAPHAEL outperforms recent cutting-edge models, such as Stable Diffusion, ERNIE-ViLG 2.0, DeepFloyd, and DALL-E 2, in terms of both image quality and aesthetic appeal.
19.	DONGXU LI, Junnan Li, Steven Hoi	BLIP-Diffusion, a new subject-driven image generation model that supports multimodal control which consumes inputs of subject images and text prompts.	Compared with previous methods such as DreamBooth, the model enables zero-shot subject-driven generation, and efficient fine-tuning for customized subject with up to 20x speedup.
20.	Junsong Chen, Chongjian Ge,	PixArt-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-	PixArt-\Sigma achieves superior image quality and user prompt

	Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, Zhenguo Li	Image Generation	adherence capabilities with significantly smaller model size (0.6B parameters) than existing text-to-image diffusion models, such as SDXL (2.6B parameters) and SD Cascade (5.1B parameters).

Summary:

Traditional image generation methods, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), face several limitations that hinder their effectiveness, particularly in accommodating user interaction and producing stable outputs. Firstly, these methods often lack robust mechanisms for incorporating user feedback and preferences into the generation process. User interaction is limited, and users have little control over the generated outputs, resulting in a lack of customization and personalization.

Moreover, traditional methods like GANs are susceptible to issues such as mode collapse, where the generator produces limited varieties of images, or training instability, where the model fails to converge to a satisfactory solution. These challenges can lead to inconsistent or undesirable outputs, undermining the reliability and usability of the generated images.

In contrast, emerging techniques such as Stable Diffusion models and the proposed method described above offer several advantages. Stable Diffusion models provide a stable and reliable framework for image generation, addressing issues of mode collapse and training instability commonly encountered in traditional methods. By leveraging diffusion processes, these models can generate high-quality, diverse images with improved stability and robustness.

Additionally, the proposed method enhances user interaction and engagement by incorporating user feedback and preferences at multiple stages of the generation process. Through mechanisms such as user rating and selection, iterative evolutionary algorithms, and image-to-image generation with user-guided edits, users have greater control and influence over the generated outputs. This increased user involvement leads to more customized and tailored images that better align with user expectations and preferences.

3. Project description and goals

3.1 Problem statement

The project aims to transform image generation and manipulation using the Stable Diffusion Model alongside interactive features. Traditional methods struggle to produce diverse, high-quality images at scale while lacking user interaction. We propose a novel approach integrating the Stable Diffusion Model to enable seamless text-to-image generation, empowering users to translate text into vivid visuals. Interactive elements such as user feedback and post-processing options enhance user engagement and refine image outputs. By democratizing image creation and bridging technical innovation with user needs, our project seeks to advance AI-driven image synthesis and foster creativity in computer vision.

3.2 Scope of the project

The project focuses on the implementation and evaluation of the Stable Diffusion Model for image generation and manipulation within a specified scope. The project encompasses the following aspects:

Text-to-Image Generation: The primary focus is on developing a robust text-to-image generation system using the Stable Diffusion Model. This includes designing and training the model to translate textual descriptions into realistic and contextually relevant visual representations.

User Interaction: The project integrates interactive elements to enhance user engagement and refine generated images. This involves implementing user feedback mechanisms, such as an Interactive Evolutionary Algorithm (IEA), to optimize image selection based on user preferences.

Image Modification: Additionally, the project explores the application of the Stable Diffusion Model for image-to-image transformations. Users are provided with the ability to modify existing images, such as applying stylistic changes or creating variations, using the model.

Post-Processing Options: The project includes the implementation of post-processing options to further refine generated or modified images. This may involve adjusting parameters such as brightness, contrast, and sharpness to meet user preferences.

5. Evaluation and Validation: The project evaluates the performance and effectiveness of the implemented system through qualitative and quantitative metrics. This includes assessing the quality of generated images, user satisfaction, and system usability.

6. Limitations: While ambitious in scope, the project acknowledges certain limitations, such as computational resources, dataset availability, and time constraints. These limitations may influence the depth and breadth of the implemented features and evaluations.

Overall, the project aims to provide a comprehensive exploration of the Stable Diffusion Model for image generation and manipulation, with a focus on user interaction and quality enhancement. By delineating the scope, the project aims to deliver tangible results within defined parameters while laying the groundwork for future research and development in this domain.

3.3 Goal of the project

The overarching goal of our project is to leverage the Stable Diffusion Model to revolutionize image generation and manipulation, with a focus on enhancing user interaction and output quality. Specific objectives include:

Implementing a Robust Text-to-Image Generation System: Our primary goal is to develop a text-to-image generation system that harnesses the capabilities of the Stable Diffusion Model. This involves designing and training the model to accurately translate textual descriptions into realistic visual representations.

Integrating Interactive Elements for User Feedback: We aim to incorporate interactive elements, such as user feedback mechanisms, to enhance user engagement and refine the generated images. This includes implementing an Interactive Evolutionary Algorithm (IEA) to optimize image selection based on user preferences.

Exploring Image Modification and Post-Processing Options: Additionally, we seek to explore the application of the Stable Diffusion Model for image-to-image transformations, allowing users to modify existing images. We also aim to provide post-processing options to further refine the generated or modified images according to user preferences.

Evaluating Performance and User Satisfaction: Our project aims to evaluate the performance and effectiveness of the implemented system through qualitative and quantitative metrics. This includes assessing the quality of generated images, user satisfaction with the system, and usability.

5. Contributing to Advancements in AI-driven Image Synthesis: Ultimately, our goal is to contribute to the advancement of AI-driven image synthesis techniques and foster creativity in computer vision. By pushing the boundaries of what is achievable with the Stable Diffusion Model, we aim to pave the way for innovative applications and advancements in the field.

Through the pursuit of these goals, our project seeks to provide a comprehensive exploration of the Stable Diffusion Model for image generation and manipulation, with the aim of delivering tangible results that advance the state-of-the-art in this domain.

4. Technical Specifications

4.1 Module and framework used.

This project aims to utilize generative AI techniques, specifically the **stable diffusion** technique, to generate quality images. Stable diffusion is a machine learning approach that involves gradually adding noise to an image over multiple steps, allowing the model to refine its understanding of the image while maintaining stability. The generated images will be produced using the **Hugging Face interface**, a popular platform for natural language processing and machine learning tasks.

Stable diffusion

Stable diffusion is a machine learning method that involves iteratively adding noise to an image while maintaining stability and coherence. This technique allows the model to gradually refine its understanding of the image without losing important features or details.

Hugging Face Interface:

Hugging Face is a leading platform for natural language processing (NLP) and machine learning tasks. It provides a user-friendly interface for accessing pre-trained models, training custom models, and deploying machine learning applications. In this project, the Hugging Face interface will be used to implement and deploy the generative AI model for image generation. The interface offers various tools and resources for working with machine learning models, making it an ideal choice for this project. Hence, the model that we implemented for our project is from this platform, this model is trained by various types of images having huge dataset with images having alt texts related to them which is used in supervised learning.

4.2 System requirements.

Runtime Environment:

The project is designed to run on Google Colab, utilizing its Python runtime environment. Google Colab offers a convenient cloud-based platform for running Python code, providing access to resources like GPUs and TPUs.

Hardware Accelerator:

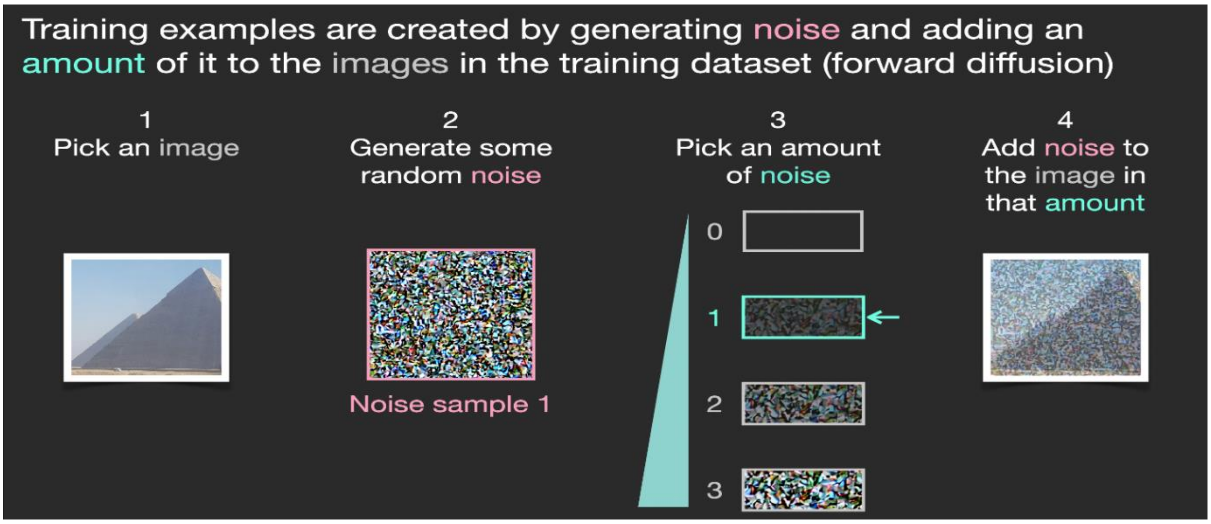
The project leverages the T4 GPU accelerator provided by Google Colab. The T4 GPU is an NVIDIA GPU optimized for deep learning workloads, offering significant computational power for training and inference tasks.

4.3 Model training.

The model that we used in this project is a pre-trained model from the hugging face interface model hub. Training of this model is done using the Stable diffusion technique as discussed below:

a. Generator Network: The generator network is typically a deep neural network responsible for generating high-quality images. It may consist of convolutional layers, residual blocks, attention mechanisms, and normalization layers. During training, the generator learns to transform noise or low-quality inputs into realistic images.

b. Diffusion Process: The training process involves the diffusion process, where noise is gradually added to the input images over multiple steps. At each step, the generator attempts to remove the added noise and reconstruct the original image. This process helps the model learn to generate images while maintaining stability and realism.


Training examples are created by generating noise and adding an amount of it to the images in the training dataset (forward diffusion)

1
Pick an image
2
Generate some random noise
Noise sample 1
3
Pick an amount of noise
0
1
2
3
4
Add noise to the image in that amount

Figure 4. Training of diffusion model by adding noise [Reference: 14]

c. Loss Function: The training is guided by a loss function, such as the mean squared error (MSE) or perceptual loss, which measures the discrepancy between the generated images and the ground truth images. The loss function drives the model to produce images that closely match the desired output.

d. Optimization Algorithm: The optimization algorithm, such as stochastic gradient descent (SGD) or Adam, is used to update the parameters of the generator network to minimize the loss function. Through backpropagation, the model learns to adjust its parameters to improve image quality and stability over time.

4.4 System architecture.

The project commences by inputting a prompt into a text-to-image generation module, which generates a set of images, typically denoted as n. Users then rate each image on a scale from 1 to 5, with only the top k images being retained for further processing. Subsequently, an Iterative Evolutionary Algorithm (IEA) is employed on this elite group, introducing noise to enhance diversity and exploration within the image space. Following this iteration, the resultant image undergoes refinement using an image-to-image generation model, accompanied by an additional prompt to guide specific modifications. Through this iterative process, the final image emerges, representing a synthesis of the initial text prompt, user preferences, evolutionary refinement, and subsequent image-to-image generation adjustments. This comprehensive approach combines the strengths of text-to-image and image-to-image generation models, augmented by user feedback and evolutionary optimization, to produce a refined and tailored visual output.

Following is the architecture of our project, the diagram is followed by the explanation of each segment of the project details:

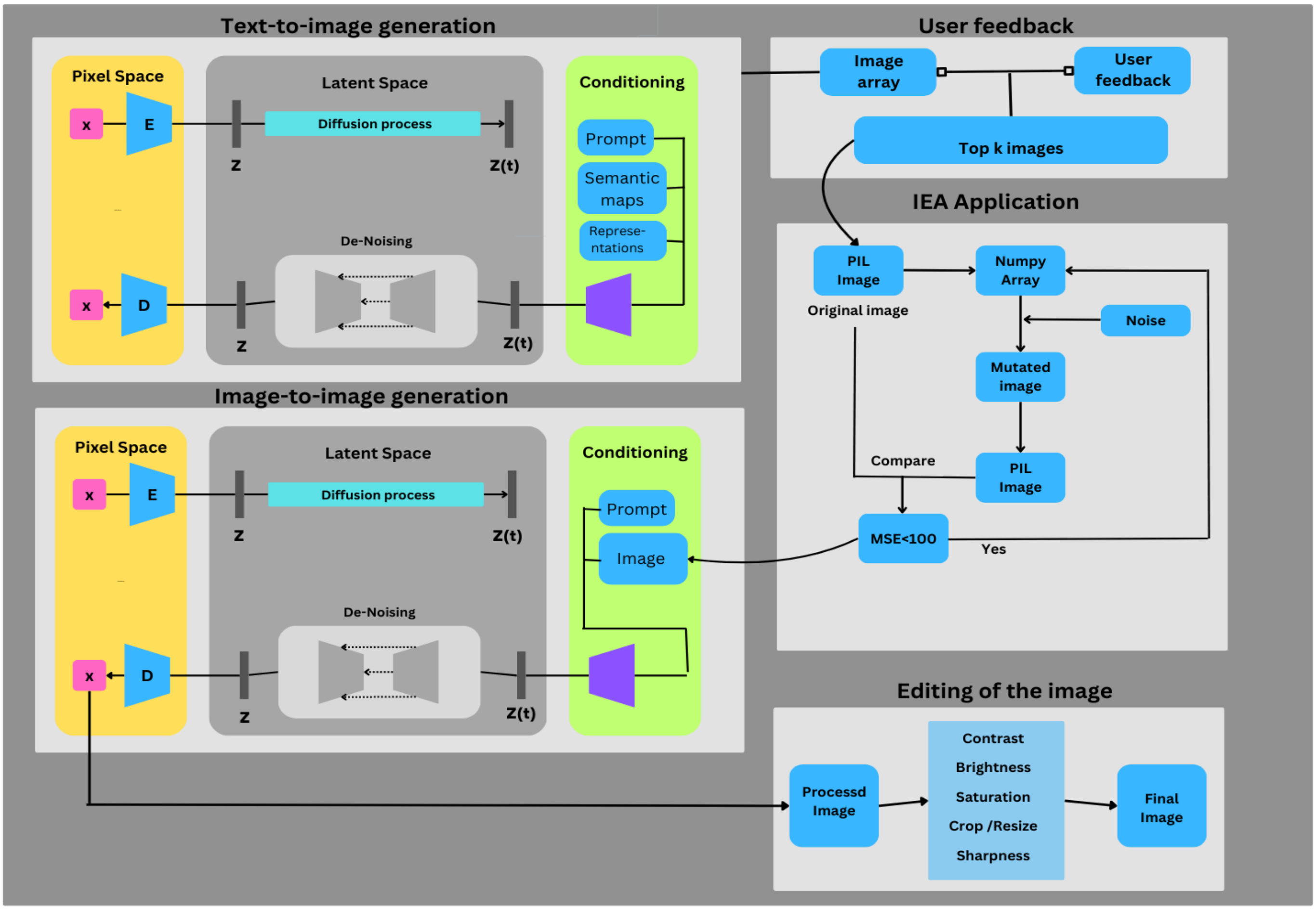


Figure 6. Architecture of the project

Generation Architecture:

a. Conditional Inputs: The generation process involves conditional inputs, such as prompts, to guide the generation of specific types of images. These inputs provide additional context to the generator network and influence the generated outputs accordingly.

b. Sampling from Noise Distribution: The generation process starts by sampling random noise vectors from a predefined noise distribution. These noise vectors serve as the initial inputs to the generator network.

c. Diffusion Steps: Like the training process, the generation process involves multiple diffusion steps, where noise is gradually added to the input images. At each step, the generator attempts to denoise the input and produce a more realistic image. This iterative process continues until the desired level of image quality is achieved.

d. Output Image: The final output of the generation process is a high-quality image generated by the generator network. This image is typically the result of multiple diffusion steps and reflects the learned patterns and structures encoded in the generator's parameters.

Iterative evolutionary algorithm Architecture:

This portion of the code implements a mutation operator as part of an evolutionary algorithm. Here's a breakdown of what each step does:

1. Initialization:
   - A new list `new_population` is initialized to store mutated images.

2. Mutation Process:
   - For each parent image in the `selected_parents` list:
   - The parent image is converted from a PIL Image to a NumPy array.
   - A copy of the parent image array (`parent_array`) is created to apply mutations.
   - Random noise is generated using `np.random.randn` with a scale factor

(`mutation_scale`) to determine the magnitude of the mutation.

- The random noise is added to the copied image array to create a mutated image (`mutated_image`).
- The pixel values of the mutated image are clipped to the range [0, 255] to ensure they are valid.
- The mutated image is converted back to a PIL Image.

3. Checking for Similarity:

- The code then checks if the mutated image is similar to any initially generated image:
- It computes the Mean Squared Error (MSE) between the mutated image and each image in the original population (`population`).
- If the MSE is below a certain threshold (`100` in this case), it considers the images similar.
- If the mutated image is similar to any initially generated image, a new mutation is generated, and the process is repeated until a non-similar mutated image is obtained.

4. Appending to New Population:

- The mutated image is added to the `new_population` list.

This process effectively introduces diversity into the population of images by mutating selected images. The mutation operator randomly perturbs the pixel values of the images, ensuring that each new generation explores different variations. Additionally, it ensures that the mutated images are not too similar to any of the originally generated images, thus encouraging exploration of diverse solutions.

Image to image generation Architecture:

It is exactly like the text to image generation, only one extra step is added to the generation of the image and that is: The initial image that the generator uses to de-noise and produce the final image is the noised version of the input image.

5. Design approach and details

5.1 Design approach / Materials and methods.

5.2 Codes and standards.

The implementation of the project commenced with adherence to fundamental coding standards, ensuring a structured and organized codebase. Following established conventions, we systematically developed the three distinct applications integral to the project, meticulously tracing the flow of data to maintain coherence and clarity throughout the implementation process. By adhering to these coding standards, we aimed to facilitate collaboration, readability, and maintainability within the project codebase.

However, the implementation of image-to-image generation posed a unique challenge, primarily stemming from the specific requirements of input images for processing. These requirements necessitated preprocessing of the data generated by the text-to-image generator model to ensure compatibility and effectiveness in the image-to-image generation phase. Overcoming this challenge involved careful consideration and implementation of preprocessing steps to transform the generated data into a suitable format for input into the image-to-image generation model.

Following is the processing of the generated image which will be sent to the image-to-image generation model:

```
# Convert `final_image` to numpy array and normalize its values to [0, 1]
final_image_array = np.array(final_image) / 255.0

# Convert the numpy array to torch tensor
final_image_tensor = torch.FloatTensor(final_image_array)

# Ensure the tensor has the shape (1, C, H, W) where C is the number of channels,
# H is the height, and W is the width
final_image_tensor = final_image_tensor.unsqueeze(0).permute(0, 3, 1, 2)
```

To ensure a seamless implementation and flow of the project, we adopted a structured approach, starting with the initiation of the hugging face interface model—the backbone of the text-to-image generation component. From there, we meticulously followed a series of steps to facilitate the transmission of input values to the model, ensuring that the process remained robust and efficient. This step-by-step approach allowed us to effectively navigate the complexities of model integration and data flow, laying a solid foundation for the successful implementation and execution of the project.

5.3 Constraints, alternatives, and tradeoffs.

In developing our project utilizing the Stable Diffusion Model for text-to-image generation and subsequent image manipulation, we encountered several constraints.

1. Computational Resources:

Firstly, computational resources posed a significant constraint, particularly for high-resolution image generation and processing. To address this, we explored alternatives such as model compression techniques and leveraging specialized hardware like GPUs or TPUs, and ultimately the best option that we got is to make this project on google colab with its GPU T4 hardware services, although these solutions come with tradeoffs in terms of performance or development complexity.

2. User Feedback Loop:

Secondly, the effectiveness of the Iterative Evolutionary Algorithm (IEA) depended heavily on the quality and quantity of user feedback. There are alternatives such as incorporating techniques for diverse feedback collection and analysis, such as active learning or reinforcement learning, to enhance the feedback loop. However, these alternatives introduced tradeoffs in terms of development effort and system complexity. And as the model is pre trained, we cannot modify it for reinforcement learning. But our implementation can be used as the building block for future such applications.

3. Image Modification Flexibility:

Moreover, while the Image-to-Image Stable Diffusion Model provided flexibility for various image modifications, its performance varied based on the complexity and style of modifications. The alternatives for this can be integrating specialized algorithms or neural architectures for specific types of modifications, such as style transfer or inpainting. However, these alternatives may sacrifice the generalizability of a unified framework and limit the range of supported image editing functionalities.

6. Schedule, tasks, and milestones.

6.1 Schedule

The project schedule unfolds through distinct phases, each meticulously designed to ensure comprehensive exploration and execution. Initially, ample time was dedicated to thorough research, delving into the intricacies of the topic and its technical underpinnings. This stage served as the foundation, allowing us to grasp the nuances of the Stable Diffusion Model and its potential applications in image generation.

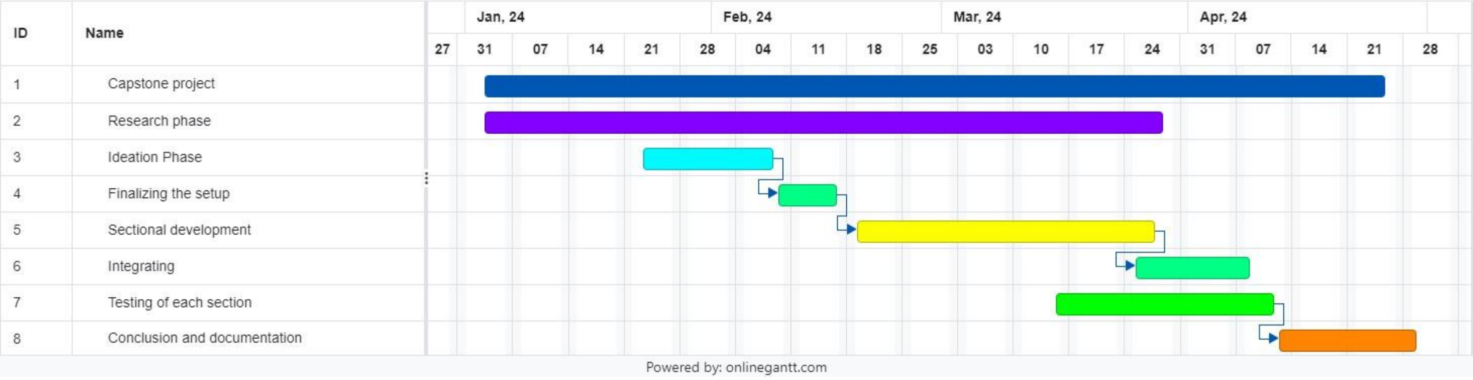
Transitioning from research to implementation, the focus shifted towards translating theoretical knowledge into practical solutions. However, early attempts to build the project within our local environment revealed significant hardware constraints, necessitating a pivot towards Google Colab. This cloud-based platform provided the necessary computational resources to commence implementation efforts effectively.

With the implementation environment secured, the project unfolded in stages, with each section—text-to-image generation, integration of an Interactive Evolutionary Algorithm (IEA), and image-to-image generation—receiving dedicated attention. As development progressed, meticulous testing was conducted, employing a diverse array of prompts to validate the functionality and robustness of the system.

Integration emerged as a pivotal milestone, where the individual sections seamlessly converged into a cohesive whole. This crucial juncture marked the culmination of extensive development efforts, paving the way for comprehensive testing and refinement. Through rigorous evaluation, the system's efficacy and performance were meticulously assessed, ensuring alignment with project objectives and user expectations.

Upon successful integration and exhaustive testing, the project reached its conclusion, marking the culmination of months of diligent effort and collaboration. With final documentation completed and lessons learned duly noted, the project's conclusion heralded not only the attainment of its objectives but also the dawn of new possibilities for future research and innovation in AI-driven image generation and manipulation.

Representation through Gantt Chart:



ID	Name
1	Capstone project
2	Research phase
3	Ideation Phase
4	Finalizing the setup
5	Sectional development
6	Integrating
7	Testing of each section
8	Conclusion and documentation

6.2 Tasks and milestones

The project tasks were meticulously aligned with corresponding milestones, ensuring a structured approach towards achieving project objectives.

- Research Phase:

  - Task: Extensive research and exploration into the Stable Diffusion Model and related techniques.

  - Milestone: Completion of research milestone, signifying a comprehensive understanding of the project domain.

- Implementation Phase:

  - Task: Translation of theoretical knowledge into practical solutions, overcoming hardware constraints by transitioning to Google Colab.

  - Milestone: Completion of implementation milestone, marking the successful realization of conceptual designs into functional code.

- Sectional Development:

  - Task: Division of the project into distinct sections—text-to-image generation, IEA integration, and image-to-image generation.

  - Milestone: Individual section completion milestones, representing significant progress towards achieving overall project objectives.

- Integration and Testing:

  - Task: Integration of individual sections into a cohesive whole, followed by rigorous testing and refinement.

  - Milestones: Integration milestone, followed by testing and refinement milestones, validating system functionality and performance against predefined metrics and user expectations.

- Conclusion and Documentation:

  - Task: Reflection on lessons learned, finalization of documentation, and project conclusion.

  - Milestone: Completion of the final code milestone, signifying the realization of project objectives and paving the way for future advancements in the field.

7. Project demonstration.

7.1 Code structure

Following is the final code implementation of the project. Each code section is having a comment for the description of what the specified code section does.

```
#installation of libraries
!pip install --upgrade diffusers transformers -q
!pip install numpy
!pip install matplotlib
!pip install Pillow

from PIL import Image


from pathlib import Path
import tqdm
import torch
import pandas as pd
import numpy as np
from diffusers import StableDiffusionPipeline
from transformers import pipeline, set_seed
import matplotlib.pyplot as plt
import cv2
```

#defining the python cfg class and attaching to text-to-image API of the stable diffusion model from the hugging face interface

```
class CFG:
    device = "cuda"
    seed = 42
    generator = torch.Generator(device).manual_seed(seed)
    image_gen_steps = 35
    image_gen_model_id = "stabilityai/stable-diffusion-2"
    image_gen_size = (400,400)
    image_gen_guidance_scale = 9
    prompt_gen_model_id = "gpt2"
    prompt_dataset_size = 6
    prompt_max_length = 12
    population_size = 10  # Size of the population
    num_generations = 5   # Number of generations
    mutation_rate = 0.1   # Mutation rate
```

```
# Initialize the diffusion model

image_gen_model = StableDiffusionPipeline.from_pretrained(
    CFG.image_gen_model_id, torch_dtype=torch.float16,
    revision="fp16",
use_auth_token='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',
guidance_scale=CFG.image_gen_guidance_scale
)
image_gen_model = image_gen_model.to(CFG.device)

# Function to generate an image based on a prompt
def generate_image(prompt, model):
    image = model(
        prompt, num_inference_steps=CFG.image_gen_steps,
        generator=CFG.generator,
        guidance_scale=CFG.image_gen_guidance_scale
    ).images[0]

    image = image.resize(CFG.image_gen_size)
    return image

# Evolutionary algorithm loop

population = [generate_image("View of a beach from a house",
image_gen_model) for _ in range(CFG.population_size)]

# Display generated images

fig, axes = plt.subplots(1, len(population), figsize=(15, 5))
for idx, image in enumerate(population):
    axes[idx].imshow(image)
    axes[idx].axis('off')
plt.show()

# Function to display images and collect feedback

def display_and_collect_feedback(population):
    user_feedback = []
    for idx, image in enumerate(population):
        feedback = input(f"Please provide rating out of 5 for image
{idx+1}/{len(population)}: ")
        user_feedback.append(float(feedback))
    return user_feedback
```

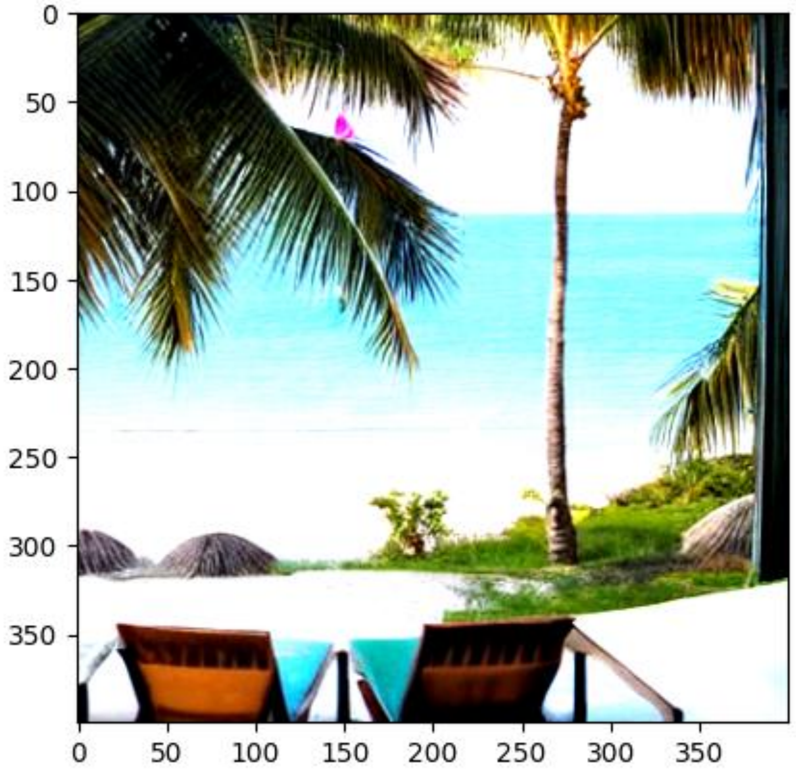
```
# Collect feedback from the user

user_feedback = display_and_collect_feedback(population)

# Selection: Select the top-k images based on user feedback
top_k_indices = sorted(range(len(user_feedback)), key=lambda i: user_feedback[i], reverse=True)[:CFG.population_size // 2]
selected_parents = [population[i] for i in top_k_indices]

# Variation: Mutation operator
import random

# Variation: Mutation operator
new_population = []
for parent in selected_parents:
    # Convert PIL Image to NumPy array
    parent_array = np.array(parent)
    # Create a copy of the parent image
    mutated_image = parent_array.copy()
    # Apply mutation to the image
    mutation_scale = 0.8  # Scale factor for mutation
    mutation = np.random.randn(*parent_array.shape) * mutation_scale
    # Add mutation to the image
    mutation = mutation.astype(np.uint8)
    mutated_image += mutation
    # Clip the pixel values to [0, 255]
    mutated_image = np.clip(mutated_image, 0, 255)
    # Convert NumPy array back to PIL Image
    mutated_image = Image.fromarray(mutated_image.astype(np.uint8))

    # Ensure the mutated image is not exactly similar to any initially generated image
    # Compare the mutated image with each initially generated image
    is_similar = False
    for orig_image in population:
        orig_array = np.array(orig_image)
        # Compute mean squared error (MSE) between the mutated image and the initially generated image
        mse = np.mean((mutated_image - orig_array) ** 2)
        # If the MSE is below a certain threshold, consider the images similar
        if mse < 100:  # Adjust the threshold as needed
```

```
            is_similar = True
            break
    # If the mutated image is similar to any initially generated image, generate a new mutation
    if is_similar:
        mutation = np.random.randn(*parent_array.shape) * mutation_scale
        mutated_image = parent_array + mutation
        mutated_image = np.clip(mutated_image, 0, 255)
        mutated_image = Image.fromarray(mutated_image.astype(np.uint8))

    new_population.append(mutated_image)

mutated_image;

# Replace the old population with the new one
population = new_population


# Display the final evolved image
final_image = population[0]
plt.imshow(final_image)
plt.show()


import matplotlib.pyplot as plt
import io
import base64

!nvidia-smi

!pip install ftfy
!pip install -qq "ipywidgets>=7,<8"

import inspect
import warnings
from typing import List, Optional, Union

import torch
from torch import autocast
from tqdm.auto import tqdm

from diffusers import StableDiffusionImg2ImgPipeline
```

```
device = "cuda"
model_path = "CompVis/stable-diffusion-v1-4"

pipe = StableDiffusionImg2ImgPipeline.from_pretrained(
    model_path,
    revision="fp16",
    torch_dtype=torch.float16,
    use_auth_token=True
)
pipe = pipe.to(device)

# Convert `final_image` to numpy array and normalize its values to [0, 1]
final_image_array = np.array(final_image) / 255.0

# Convert the numpy array to torch tensor
final_image_tensor = torch.FloatTensor(final_image_array)

# Ensure the tensor has the shape (1, C, H, W) where C is the number of channels,
# H is the height, and W is the width
final_image_tensor = final_image_tensor.unsqueeze(0).permute(0, 3, 1, 2)

# Define the prompt
prompt = "The view of a beach from a house with people enjoying"

# Use the final image as input to the model
images = pipe(prompt=prompt, image=final_image_tensor, strength=0.75,
guidance_scale=7.5).images

# Save or display the generated image
generated_image = images[0]
images[0]

from PIL import Image, ImageEnhance, ImageOps
import matplotlib.pyplot as plt

# Function to crop the image
def crop_image(image, box):
    return image.crop(box)

# Function to rotate the image
def rotate_image(image, angle):
    return image.rotate(angle)
```

```
# Function to resize the image
def resize_image(image, size):
    return image.resize(size)

# Function to apply brightness enhancement to the image
def enhance_brightness(image, factor):
    enhancer = ImageEnhance.Brightness(image)
    return enhancer.enhance(factor)

# Function to apply contrast enhancement to the image
def enhance_contrast(image, factor):
    enhancer = ImageEnhance.Contrast(image)
    return enhancer.enhance(factor)

# Function to apply sharpness enhancement to the image
def enhance_sharpness(image, factor):
    enhancer = ImageEnhance.Sharpness(image)
    return enhancer.enhance(factor)

# Function to apply color saturation enhancement to the image
def enhance_color(image, factor):
    enhancer = ImageEnhance.Color(image)
    return enhancer.enhance(factor)

# Function to edit the image based on user input
def edit_image(image):
    while True:
        print("Select an editing option:")
        print("1. Crop")
        print("2. Rotate")
        print("3. Resize")
        print("4. Brightness")
        print("5. Contrast")
        print("6. Sharpness")
        print("7. Color Saturation")
        print("8. Finish editing")

        choice = input("Enter your choice: ")

        if choice == '1':
            # Crop the image
            box = tuple(map(int, input("Enter crop coordinates (left, upper, right, lower): ").split()))
            image = crop_image(image, box)
        elif choice == '2':
            # Rotate the image
            angle = float(input("Enter rotation angle (in degrees): "))
            image = rotate_image(image, angle)
        elif choice == '3':
            # Resize the image
            width = int(input("Enter new width: "))
            height = int(input("Enter new height: "))
            image = resize_image(image, (width, height))
        elif choice == '4':
            # Enhance brightness
            factor = float(input("Enter brightness factor (1.0 for no change): "))
            image = enhance_brightness(image, factor)
        elif choice == '5':
            # Enhance contrast
            factor = float(input("Enter contrast factor (1.0 for no change): "))
            image = enhance_contrast(image, factor)
        elif choice == '6':
            # Enhance sharpness
            factor = float(input("Enter sharpness factor (1.0 for no change): "))
            image = enhance_sharpness(image, factor)
        elif choice == '7':
            # Enhance color saturation
            factor = float(input("Enter color saturation factor (1.0 for no change): "))
            image = enhance_color(image, factor)
        elif choice == '8':
            # Finish editing
            break
        else:
            print("Invalid choice!")

    return image

# Load the final evolved image from the population
final_image = generated_image

# Edit the final evolved image based on user input
edited_image = edit_image(final_image)

# Display the edited image
plt.imshow(edited_image)
plt.show()
```

7.2 Output



7.2.1 Initial pool of generated images (population size = 10):

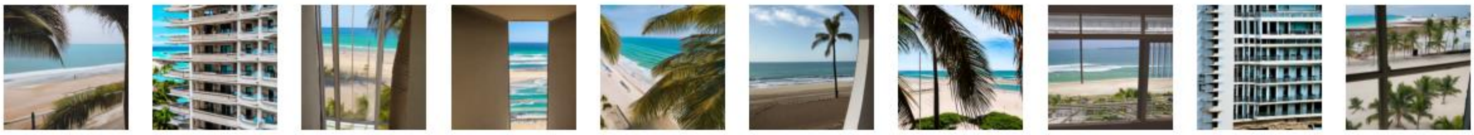


Figure 8. Population of the generated images

7.2.2 Final evolved image (after text to image generation and IEA applied):

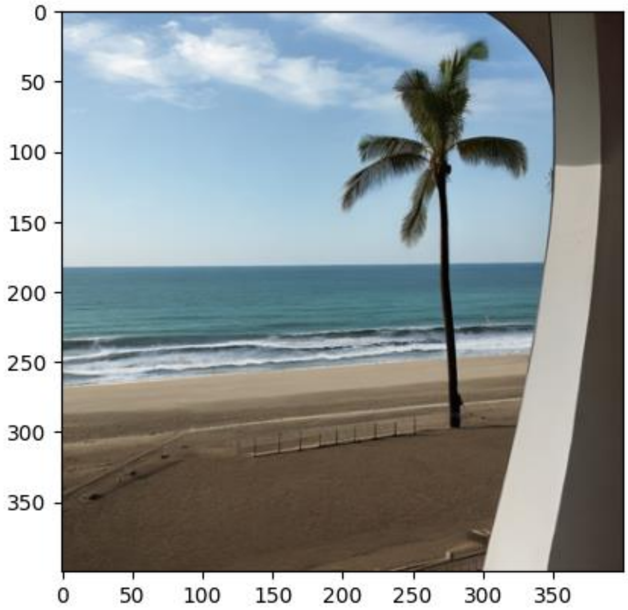


Figure 9. Final evolved image

7.2.3 Output of the image-to-image generation with further prompt from user:

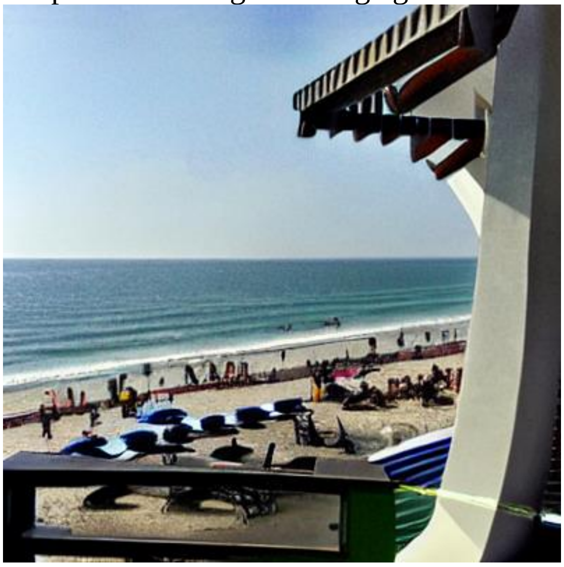


Figure 10. Output of image to image generation

7.2.4 Edited image:

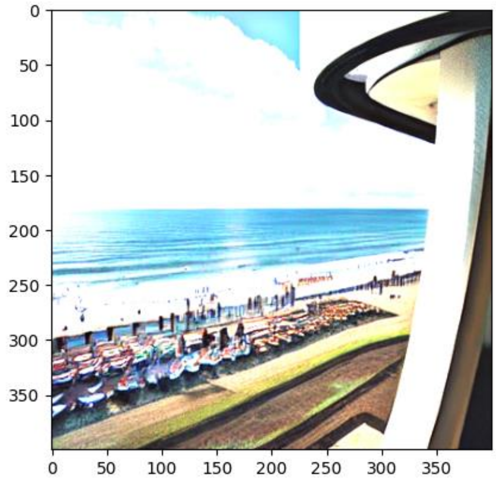


Figure 11. Final Edited image

8. Result and discussion.

8.1 Model description.

Hugging Face is a leading platform for natural language processing (NLP) and machine learning tasks. It provides a user-friendly interface for accessing pre-trained models, training custom models, and deploying machine learning applications. In this project, the Hugging Face interface will be used to implement and deploy the generative AI model for image generation. The interface offers various tools and resources for working with machine learning models, making it an ideal choice for this project. Hence, the model that we implemented for our project is:

stabilityai/stable-diffusion-2

This model is trained by various types of images having huge dataset with images having alt texts related to them which is used in supervised learning.

8.2 Evaluation Metrics

1. FID (Fréchet Inception Distance): Measures the similarity between real and generated images based on feature representations extracted from a pre-trained Inception model. Lower FID scores indicate better quality and diversity of generated images.

2. Inception Score (IS): Evaluates the quality and diversity of generated images based on their visual appeal and variety. Higher IS scores indicate better image quality and diversity.

3. User Studies: Involve human evaluators rating generated images based on criteria such as realism, diversity, and relevance to the given prompt.

8.3 Results

Testing with different prompts

Prompt 1: View of the city from the aeroplane

Population of generated images:

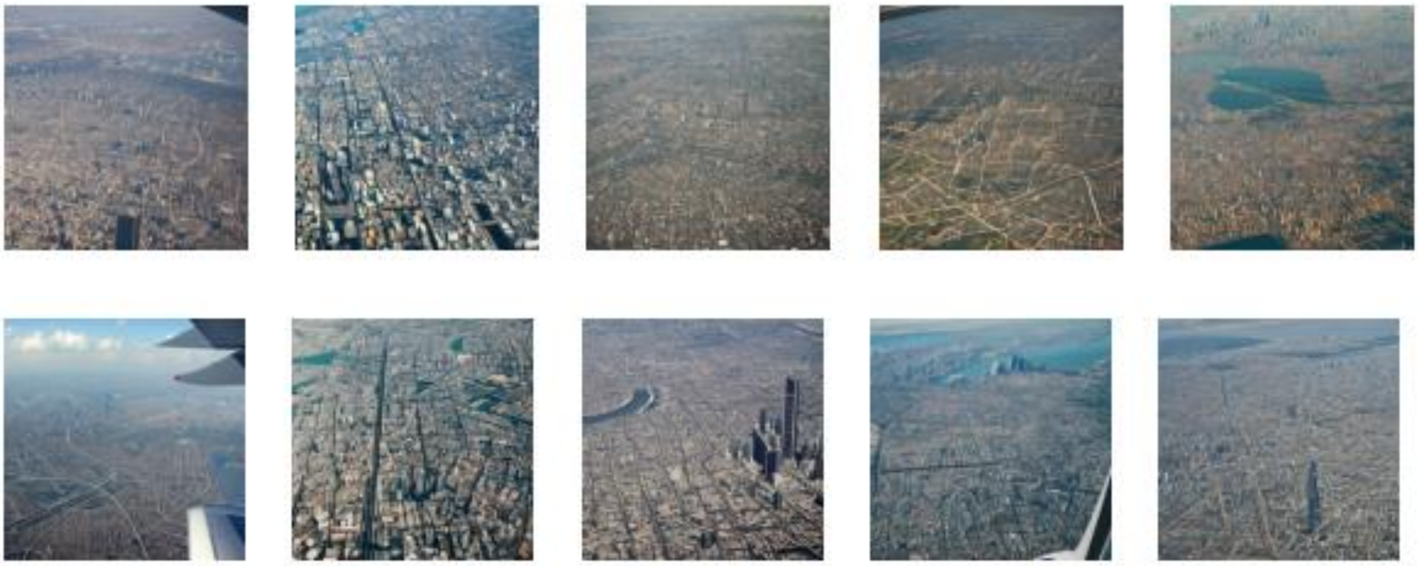


Figure 12. Population of generated images in testing

Final Image:

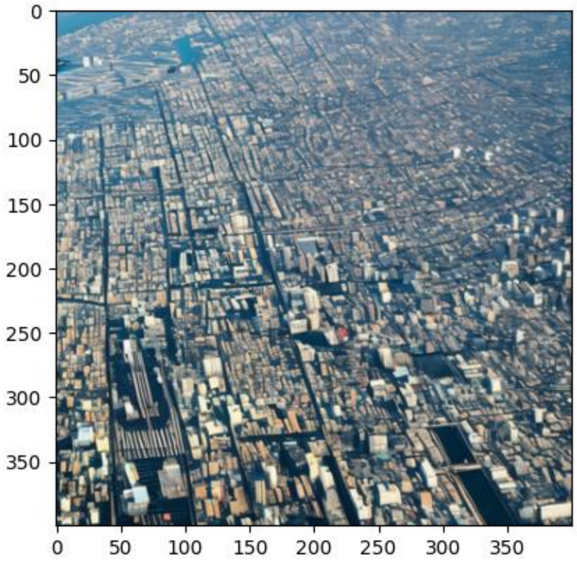


Figure 13. Final Image

Prompt 2: A house in mountains with trees.

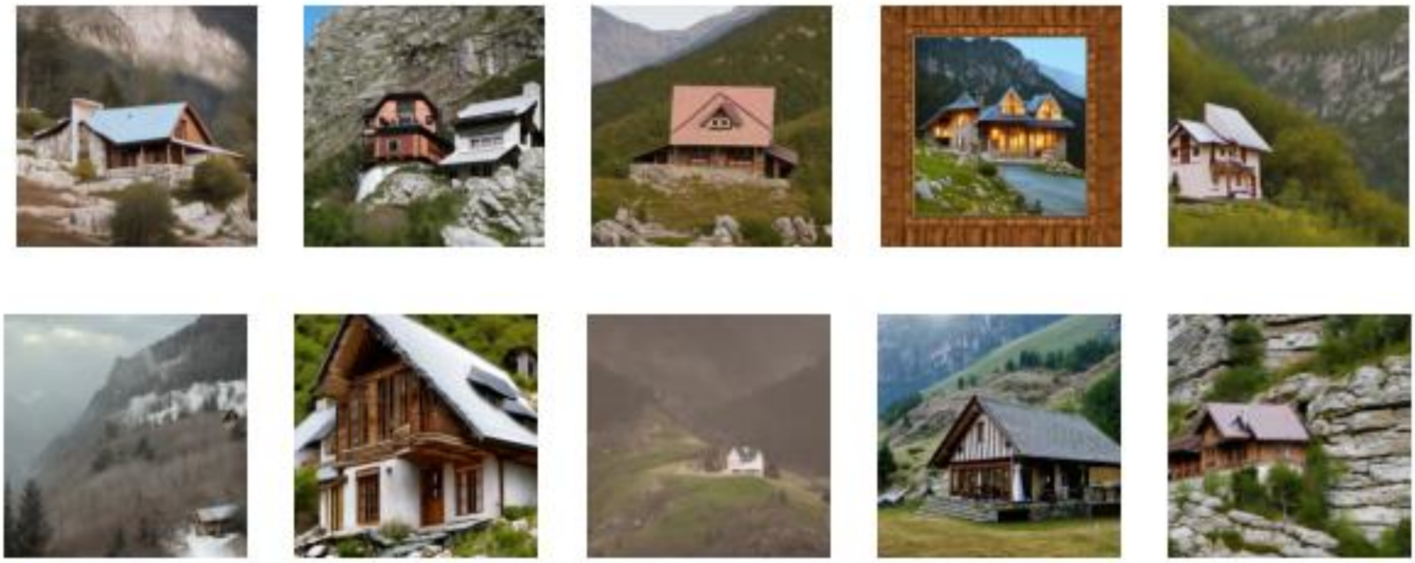


Figure 14. Population of generated images in testing 2

Final Generated image: (prompt: house in the mountains with trees)

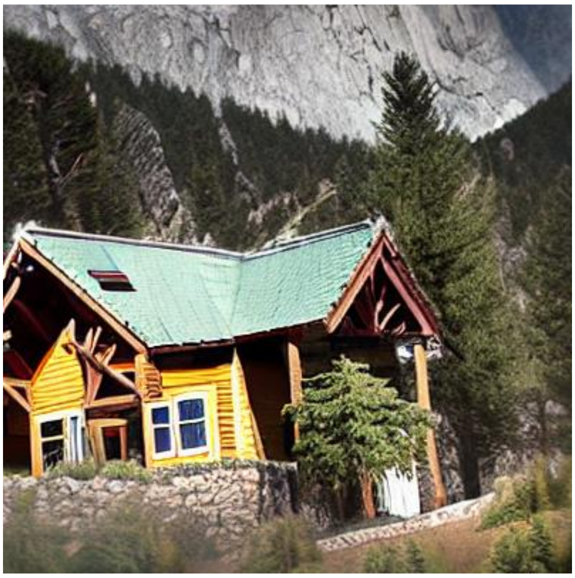


Figure 15. Final Image

Prompt 2: Night sky with stars.

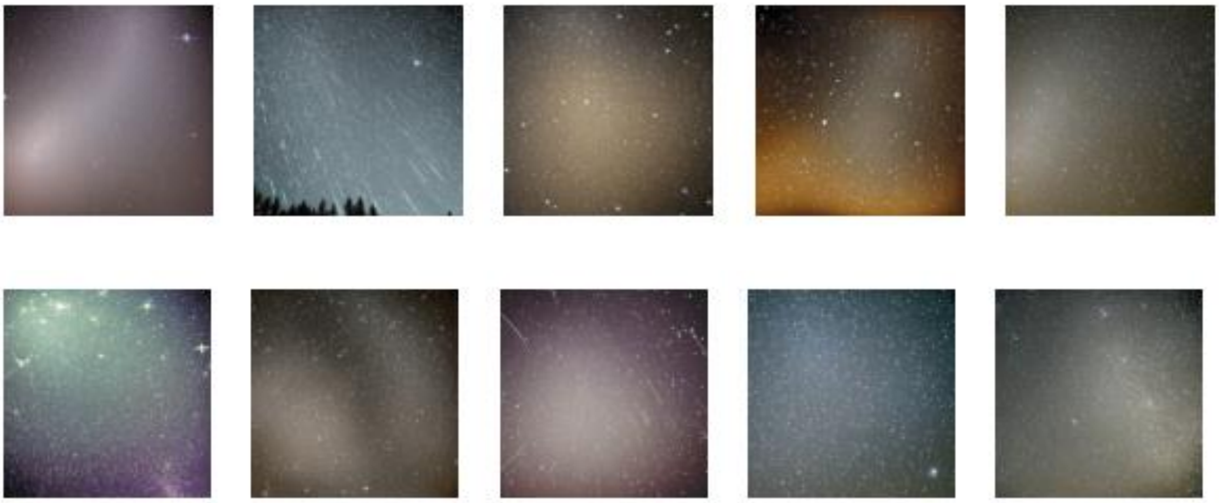


Figure 14. Population of generated images in testing 2

Final Generated image: (prompt: Night sky with stars in mountauns)

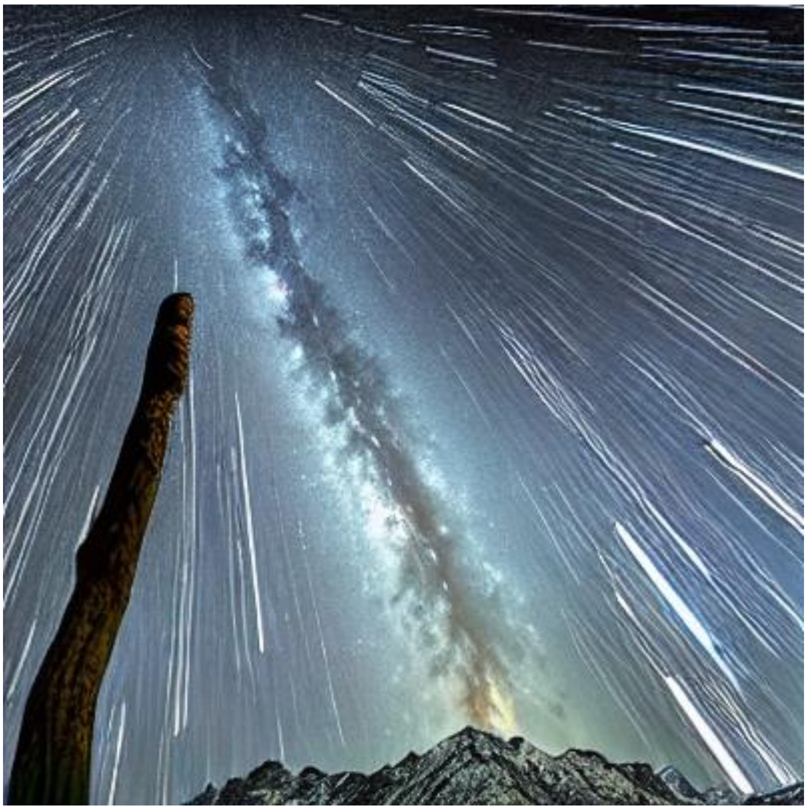


Figure 15. Final Image

1. Performance Evaluation

- Image Generation Quality: The system's ability to generate high-quality images was tested using different prompts, there are some limitations for certain vague prompts for which the desired output and vary accordingly.

2. User Feedback Analysis

- Effectiveness: Users' perceptions of the system's effectiveness in translating textual descriptions into meaningful images were examined. Feedback indicated favorable responses, with users expressing satisfaction with the accuracy and relevance of generated images.

3. Implications and Future Directions

- Applications: The project's implications in various domains, such as creative content generation, design prototyping, and educational material creation. The system's versatility and adaptability make it suitable for a wide range of applications, with potential to revolutionize image synthesis and manipulation workflows.

- Limitations and Future Work: Limitations of the current implementation, such as dataset biases, model constraints, and user interface improvements, were acknowledged. Future research directions were proposed, including dataset diversification, model refinement, and user interface enhancements, to address these limitations and further enhance system capabilities.

4. Conclusion

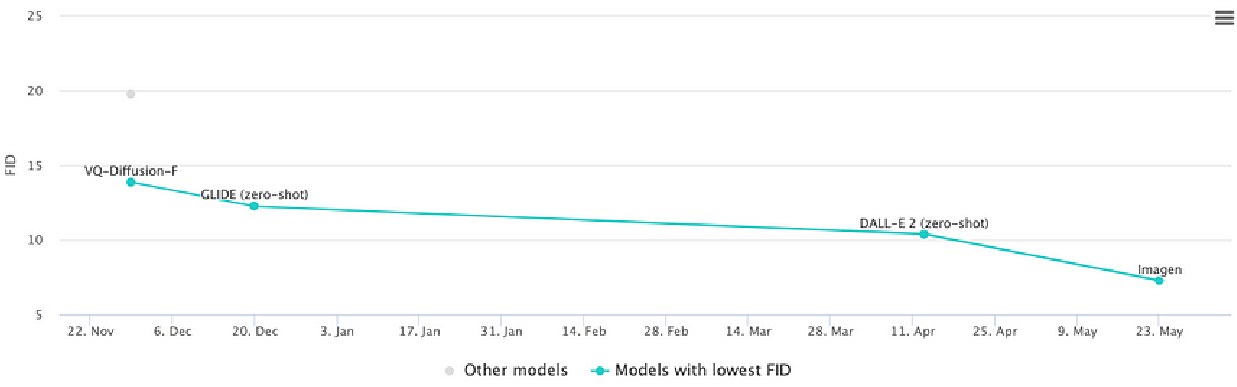
- Project Contributions: The results and discussion underscore the project's contributions to the advancement of AI-driven image generation and manipulation. By demonstrating the efficacy and usability of the system, the project contributes to the body of knowledge in computer vision and artificial intelligence.

- Final Remarks: Concluding remarks reflect on the project's achievements, challenges, and implications. Insights gained from the implementation and evaluation process inform future research and development efforts, paving the way for continued innovation in the field.

8.4 Performance Metrics

Research findings:

To assess the quality of images created by generative models, it is common to use the Fréchet inception distance (FID) metric. In a nutshell, FID calculates the distance between the feature vectors of real images and generated images. On the COCO benchmark, Imagen currently achieved the best (lowest) zero-shot FID score of 7.27,



outperforming DALL·E 2 with a 10.39 FID score.

Figure 16. FID Score of different technologies.

The Stable Diffusion team has not published any benchmark scores to enable comparison to other models. From the original Latent Diffusion paper, the Latent Diffusion Model (LDM) has reached a 12.63 FID score

Frechet Inception Distance (FID): Assessing Image distribution similarity. FID stands as a cornerstone metric that measures the distance between the distributions of generated and real images. Lower FID scores signify a closer match between generated and real-world images. In addition, it shows superior model performance in mimicking real data distributions.

8.5 Cost Analysis

1. Training of the Dataset:

- Requirement for Diverse Dataset: To enhance the effectiveness of the Stable Diffusion Model for image generation, a diverse and comprehensive dataset is essential. This dataset should encompass a wide range of visual content across various domains to ensure that the model learns to generate images that are contextually relevant and diverse.
- Funding Allocation: Allocating funds towards acquiring and curating such a dataset is crucial. This includes expenses related to purchasing existing datasets, collecting new data through crowdsourcing or partnerships, and annotating the data to provide meaningful labels and descriptions.

2. Computational Power:

- GPU Power: High-performance GPU (Graphics Processing Unit) resources are indispensable for training deep learning models effectively. GPUs excel at parallel processing tasks, making them ideal for accelerating the training process and reducing iteration times.
- Graphics and RAM: In addition to GPU power, sufficient graphics capability and RAM

(Random Access Memory) are essential for producing high-quality images and ensuring smooth functioning of the system, respectively.

- Funding Allocation: Allocating funds towards upgrading or investing in state-of-the-art GPU hardware, graphics cards, and RAM modules is imperative. This includes expenses related to purchasing new hardware, upgrading existing infrastructure, and ensuring compatibility with the latest deep learning frameworks and software tools.

Investing in these key areas—training dataset and computational power—will significantly enhance the system's capabilities for image generation and manipulation. By allocating funds strategically to address these critical needs, the project can achieve higher levels of accuracy, diversity, and efficiency in image synthesis, ultimately advancing the state-of-the-art in AI-driven image generation technologies

9. Summary.

The project represents a significant endeavor to innovate within the realm of image generation and manipulation, leveraging the power of the Stable Diffusion Model alongside interactive elements. It commences with a thorough investigation into the theoretical underpinnings of the Stable Diffusion Model, laying a solid foundation for its integration into the project's architecture. Through meticulous research and experimentation, the project identifies key areas where future investment and improvement are essential to optimize performance and enhance user experience.

One of the primary areas identified for further development is the training dataset utilized to train the Stable Diffusion Model. The efficacy of the model heavily relies on the diversity and richness of the dataset it is trained on. Thus, investing in acquiring and curating a comprehensive dataset with a wide range of visual content becomes paramount. By securing funds to expand and refine the dataset, the project can bolster the model's ability to generate high-quality and contextually relevant images across various domains.

In addition to dataset enhancement, the project underscores the critical importance of computational power in driving advancements in image synthesis and manipulation. GPU power, graphics capabilities, and RAM play pivotal roles in accelerating training times, improving image quality, and enhancing system responsiveness. Therefore, allocating funds towards upgrading and expanding computational resources is imperative to ensure the project's success in achieving its objectives. By investing in state-of-the-art hardware and infrastructure, the project can significantly elevate its capabilities and competitiveness in the field of AI-driven visual content creation.

In summary, the project represents a holistic approach to advancing image generation and manipulation through a combination of cutting-edge technology and strategic investment. By prioritizing improvements in dataset quality and computational resources, the project aims to position itself at the forefront of innovation, driving progress and pushing the boundaries of what is achievable in the realm of AI-driven image synthesis and manipulation.

Future Works:

1. Optimization and Efficiency: Explore techniques to improve the efficiency and computational cost of the proposed methodology, such as optimizing the IEA process and leveraging parallel computing resources.

2. Enhanced User Interaction: Investigate ways to enhance user interaction and feedback mechanisms, such as real-time adjustments and intuitive interfaces, to further personalize the image generation process.

3. Multi-Modal Generation: Extend the project to support multi-modal generation, allowing for the synthesis of images from different modalities (e.g., text, audio) to create richer and more diverse outputs.

4. Adversarial Training: Incorporate adversarial training techniques to improve the realism and fidelity of generated images, potentially bridging the gap with state-of-the-art GAN-based systems.

5. Evaluation and Benchmarking: Conduct rigorous evaluation and benchmarking against state-of-the-art image generation systems using a diverse set of metrics and datasets to validate the effectiveness and performance of the proposed methodology.

6. Semi-Supervised Learning: Explore semi-supervised learning techniques to leverage both labeled and unlabeled data, potentially improving the quality and diversity of generated images while reducing the need for extensive user feedback.

7. Fine-Grained Control: Investigate methods to provide finer-grained control over the

generated images, allowing users to specify desired attributes, styles, or characteristics with greater precision.

8. Interpretability and Explainability: Develop techniques to enhance the interpretability and explainability of the image generation process, providing insights into how user feedback and evolutionary algorithms influence the final output.

9. Transfer Learning: Explore transfer learning approaches to adapt pre-trained models and representations to specific image generation tasks, potentially accelerating the training process and improving performance on domain-specific data.

10. Dynamic Prompting: Implement dynamic prompting mechanisms that adaptively adjust the prompts provided to the text-to-image generation model based on user feedback and evolving preferences, leading to more effective and efficient image generation.

11. Multi-Objective Optimization: Investigate multi-objective optimization techniques to simultaneously optimize multiple criteria (e.g., image quality, diversity, relevance) in the image generation process, balancing competing objectives to produce well-rounded outputs.

12. Ethical Considerations: Address ethical considerations related to user privacy, bias, and fairness in image generation, ensuring that the system respects user preferences and societal norms while avoiding harmful or inappropriate outputs.

References

[1] Real-World Image Variation by Aligning Diffusion Inversion Chain. https://proceedings.neurips.cc/paper_files/paper/2023/hash/61960fdfda4d4e95fa1c1f6e64bfe8bc-Abstract-Conference.html

[2] Defect Image Sample Generation With Diffusion Prior for Steel Surface Defect Recognition Yichun Tai, Kun Yang, Tao Peng, Zhenzhen Huang, Zhijiang Zhang. https://arxiv.org/abs/2405.01872

[3] One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, Jun Zhu. https://proceedings.mlr.press/v202/bao23a.html

[4] Equivariant Diffusion for Molecule Generation in 3D Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, Max Welling. https://proceedings.mlr.press/v162/hoogeboom22a.html

[5] State of the Art on Diffusion Models for Visual Computing. https://arxiv.org/abs/2310.07204

[6] Haomin Zhuang, Yihua Zhang, Sijia Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 2385-2392. https://openaccess.thecvf.com/content/CVPR2023W/AML/html/Zhuang_A_Pilot_Study_of_Query-Free_Adversarial_Attack_Against_Stable_Diffusion_CVPRW_2023_paper.html

[7] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, Ping Luo. Text-to-Image Generation via Large Mixture of Diffusion Paths.

https://proceedings.neurips.cc/paper_files/paper/2023/hash/821655c7dc4836838cd8524d07f9d6fd-Abstract-Conference.html

[8] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, qiang liu. InstaFlow: One Step is Enough for High-Quality Diffusion-Based Text-to-Image Generation. https://openreview.net/forum?id=1k4yZbbDqX

[9] Ahmed Imran KABIR, Limon MAHOMUD, Abdullah Al FAHAD, Ridwan AHMED. Empowering Local Image Generation: Harnessing Stable Diffusion for Machine Learning and AI. https://www.revistaie.ase.ro/content/109/03%20-%20kabir,%20mahomud,%20fadad,%20ahmed.pdf

[10] Loc X. Nguyen; Pyae Sone Aung; Huy Q. Le; Seong-Bae Park; Choong Seon Hong. A New Chapter for Medical Image Generation: The Stable Diffusion Method. https://ieeexplore.ieee.org/abstract/document/10049010

[11] https://www.thoughtco.com/definition-of-diffusion-604430

[12] https://twitter.com/iScienceLuvr/status/1564847724033241088

[13] https://bootcamp.uxdesign.cc/how-stable-diffusion-works-explained-for-non-technical-people-be6aa674fa1d

[14] https://jalammar.github.io/illustrated-stable-diffusion/

[15]