DAYANANDA SAGAR UNIVERSITY

SCHOOL OF ENGINEERING

Bachelor of Technology

In
Computer Science and Engineering

(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

A Project Report On

Harnessing Stable Diffusion Model for High- Resolution Text-to-Image Synthesis

Submitted By

Ashish Patil - ENG21AM0013

Atharva T - ENG21AM0014

Diksha Sinha - ENG21AM0033

Tenzin Ludup - ENG22AM3011

Under the guidance of

Dr. Vegi Fernando A

Assistant Professor, CSE(AIML), DSU

2024 - 2025

Department of Computer Science and Engineering (AI & ML)

DAYANAND SAGAR UNIVERSITY

Bengaluru - 560068

SCHOOL OF
ENGINEERING

Dayananda Sagar University

Devarakaggalahalli, Harohalli Kanakapura Road, Dt, Ramanagara, Karnataka 562112, India

Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning)

CERTIFICATE

This is to certify that the project entitled **Harnessing Stable Diffusion Model for High-Resolution Text-to-Image Synthesis** is a Bonafide work carried out by **Ashish Patil (ENG21AM0013)**, **Atharva T (ENG21AM0014), Diksha Sinha (ENG21AM0033)** and **Tenzin Ludup (ENG22AM3011)** in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning), during the year 2024-2025.

Dr. Vegi Fernando A
Assistant Professor
Dept. of CSE (AIML)
School of Engineering
Dayananda Sagar University

Signature ....

Dr. Vinutha N
Project Co-Ordinator Dept. of CSE (AIML) School of Engineering
Dayananda Sagar University

Signature ....

Dr. Jayavrinda Vrindavanam

Professor & Chairperson
Dept. of CSE (AIML)
School of Engineering
Dayananda Sagar University

Signature ....

Name of the Examiners:

1 ………………………
2 ………………………
3 ………………………

Signature with date:

…………………………
…………………………
…………………………

Acknowledgement

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work. First, we take this opportunity to express our sincere gratitude to **School of Engineering and Technology, Dayananda Sagar University** for providing us with a great opportunity to pursue our Bachelor's degree in this institution. We would like to thank **Dr. Udaya Kumar Reddy K R**, Dean, School of Engineering and Technology, Dayananda Sagar University for his constant encouragement and expert advice. It is a matter of immense pleasure to express our sincere thanks to **Dr. Jayavrinda Vrindavanam**, Professor & Department Chairperson, Computer Science and Engineering (Artificial Intelligence and Machine Learning), Dayananda Sagar University, for providing right academic guidance that made our task possible. We would like to thank our guide **Prof. Dr. Vegi Fernando A**, Assistant Professor, Dept. of Computer Science and Engineering, for sparing his valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project. We would like to thank our Project Coordinator **Dr. Vinutha N** as well as all the staff members of Computer Science and Engineering (AIML) for their support. We are also grateful to our family and friends who provided us with every requirement throughout the course. We would like to thank one and all who directly or indirectly helped us in the Project work.

ASHISH PATIL - ENG21AM0013

ATHARVA T - ENG21AM0014

DIKSHA SINHA -ENG21AM0033

TENZIN LUDUP -ENG22AM3011

Harnessing Stable Diffusion Model for High-Resolution Text-to-Image Synthesis

Ashish Patil, Atharva T, Diksha Sinha, Tenzin Ludup

Abstract

This project's main goal is to decode verbal cues into equivalent high-quality visuals by putting an advanced AI diffusion model into practice. The diffusion model is a class of generative AI techniques based on the iterative denoising process. It begins with a disorganized image and carefully refines it over a number of cycles to get a coherent and realistic final. In addition to being visually appealing, this systematic process allows the production of visuals that closely correspond to the spoken description supplied by the user. The project makes use of pre-trained models, such as Stable Diffusion, which are well-known for their ability to generate images from text. Fine-tuning is done on particular datasets that are suited to various use cases in order to improve the model's accuracy and adaptability. The system can process a wider variety of textual inputs thanks to this customisation, ranging from straightforward descriptive statements to intricate and subtle prompts. The application of cutting-edge methods like latent space

processing, which effectively reduces computing overhead and compresses and manipulates data, guarantees that the system retains high-quality output. A U-Net design, a kind of neural network that excels at producing images with minute details and global structures, is used in the study. The model effectively focuses on relevant sections of the input text when paired with attention mechanisms, ensuring accurate and intricate visual outputs.

A wide range of creative industries, including media, art, and design, benefit from this technology's enhanced content generation and visualization capabilities. E-commerce can be utilized to produce product images, healthcare to enhance medical imaging, and education to help Students visualize complex concepts. Through an examination of diffusion model architectures and principles, the study establishes a connection between literary imagination and digital visualization. It establishes the foundation for further advancements by demonstrating AI's potential in content creation and human-computer interaction.

Contents

Contents

Chapter 1

Introduction

The way we produce, Artificial intelligence (AI) has transformed the way we create, consume, and interact with digital information, sparking innovative solutions in a variety of sectors. With tools and methods for creating sophisticated, realistic, and variable outputs that serve a wide range of applications, generative design is a leading field among these breakthroughs. Diffusion models have revolutionized the field of generative AI. These models provide high-quality, coherent, and realistic results by iteratively filtering noisy data. By employing this new method, diffusion models have pushed the boundaries of AI's ability to generate intricate and detailed outputs.

The primary goal of this project is to develop and implement an artificial intelligence diffusion model capable of producing photorealistic pictures from textual descriptions. Fundamentally, the program seeks to develop a link between human creativity and machine generated pictures. Using complex algorithms such as probabilistic de-noising methods and sophisticated attention processes, the model can precisely grasp textual prompts and generate accompanying visuals with exceptional fidelity and detail. Even abstract or complex notions can be properly portrayed thanks to this capability, which allows the model to offer pictures that are both realistic and well-matched to the intricacies of the input descriptions.

This work has a wide range of applications. This technology can be used in e-commerce to produce dynamic product graphics from textual descriptions, which improves consumer experiences and speeds up the design process. AI-generated photographs can provide personalized visual material in advertising, removing the need for extensive manual effort. It can dramatically save production time in the entertainment industry by assisting with the creation of visually appealing materials for films, video games, and virtual reality experiences. This technology has the ability to make learning more dynamic and engaging by assisting with the visualization of challenging concepts.

Additionally, this technology gives artists and developers a powerful tool for swiftly and easily realizing their ideas in creative fields like as media, design, and virtual reality. Beyond these helpful uses, the program contributes to AI research by proving diffusion models' potential and examining their underlying mechanics. This work lays the groundwork for greater alignment between human imagination and computer-generated outputs in future breakthroughs in generative artificial intelligence and applications.

However, the high computational requirements of such models often restrict their deployment on low-power devices. To address this challenge, our project integrates several lightweight enhancements into the Stable Diffusion framework, making it more efficient and accessible without compromising performance.

Key innovations include Low-Rank Adaptation (LoRA), which allows fine-tuning with significantly fewer parameters; HyperNetwork-based prompt embeddings, which dynamically generate more context-aware features for better alignment between text and image; and a Mixture of Denoisers, which accelerates the diffusion process while preserving output quality. Additionally, we incorporate a Multi-Modal Cross-Attention Transformer module to improve the fusion of textual and visual information, resulting in higher semantic relevance in generated outputs.

These enhancements collectively reduce inference time, improve prompt understanding, and allow Stable Diffusion to run effectively on edge devices like Jetson Nano and Raspberry Pi, making powerful AI-driven creativity tools more accessible for real-world applications in areas such as AR/VR, education, mobile apps, and digital art.

1.1 Scope

The goal of this research is to create, design, and implement an AI diffusion model capable of producing realistic visuals from textual descriptions. The project's primary purpose is to bridge the gap between textual imagination and visual representation by employing diffusion models to transform abstract notions into physical visuals. The model attempts to grasp textual cues with remarkable coherence and detail by employing cutting-edge probabilistic algorithms and incorporating tactics such as latent space optimization and attention processes, resulting in outputs that closely match user intent.

This study looks into the various applications of this technology in several fields. It has the ability to alter media and design by automating high-quality graphic production, as well as improving immersive experiences in virtual reality and gaming by creating dynamic and responsive environments. The capacity to produce visual content directly from product descriptions could considerably benefit practical applications such as e-commerce and the creative industries by reducing workflows and increasing customer engagement. Furthermore, the technology has potential uses in accessibility, medical imaging, and educational tools, where the ability to envision complex concepts or circumstances may improve comprehension and utility. This study employs rigorous evaluation methodologies using industry-standard measurements such as the Fr'echet Inception Distance (FID) [1] , The Structural Similarity Index Measure (SSIM) [2] and BLEU [3] Scores to ensure the model's efficacy. These metrics provide quantifiable data regarding the created images' accuracy, quality, and realism. This paper enhances the field of AI-driven generative systems by improving our theoretical understanding of diffusion models and providing a scalable, usable strategy for producing high-quality results. By demonstrating AI's revolutionary potential for bridging the gap between computational visualization and human imagination, it hopes to pave the way for creative and economic developments.

Chapter 2

Problem Definition

Currently current generative models, such as GANs and VAEs, struggle to convert complex oral descriptions into extremely detailed and cohesive visual representations. Frequently, they mistake little cues, resulting in inconsistent or unrealistic outcomes. A description like "A misty forest at dawn with a river flowing under a wooden bridge" will not convey the desired tone or subtleties. These constraints make it difficult to use these models in settings requiring realism and precision.

To address these limitations, our study employs diffusion models, a cutting-edge tool in generative AI. Diffusion models iteratively enhance noisy data through a de-noising process, beginning with a random noise image and progressing to a high-quality, realistic representation over time. This rigorous technique allows for greater control over the generating process, ensuring that the outputs closely match the input descriptions in terms of both detail and tone. Furthermore, diffusion models excel at processing complex stimuli, capturing diverse textures, lighting, and spatial arrangements that traditional models may overlook.

The accuracy and effectiveness of diffusion models are improved further by employing tactics such as latent space representation, U-Net designs, and attention processes. By focusing on crucial textual components, these developments allow the system to generate visuals that accurately reflect the user's purpose. This strategy significantly improves the quality and consistency of AI-generated pictures while simultaneously addressing the limitations of traditional generative models. To address these obstacles, this study employs a diffusion model that iteratively de-noises noisy data to generate realistic, high-quality images. Diffusion models improve their ability to create pictures that closely match written descriptions by providing more control, precision, and efficiency. This technique mitigates the drawbacks of standard models. By addressing the flaws of previous models, this technique seeks to increase the quality and consistency of AI-generated graphics for real-world use in the creative industries.

Chapter 3

Literature Survey

Generative modelling has evolved significantly, with each major contribution addressing specific shortcomings of its predecessors while also bringing fresh approaches. In 2013, Diederik P. Kingma and Max Welling introduced Variational Autoencoders (VAEs), which marked the beginning of the voyage. VAEs introduced a probabilistic paradigm for learning continuous latent variable spaces, which used encoder-decoder architectures to describe data distributions. Their capacity to generate synthetic data by sampling from a latent space made them essential in generative modelling. However, because Gaussian priors stressed global distributional consistency above fine-grained features, they frequently resulted in hazy outputs. Despite their efficiency and probabilistic rigor, VAEs were less suitable for jobs that required high-resolution, crisp outputs.

A paradigm change came with Ian Goodfellow and associates' 2014 presentation of Generative Adversarial Networks (GANs). GANs implemented an adversarial framework consisting of two neural networks: a generator and a discriminator. The generator was designed to generate realistic data samples, whereas the discriminator was responsible for distinguishing between genuine and synthetic data. This adversarial interplay produced highly realistic results that outperformed VAEs in terms of detail and sharpness. Despite its revolutionary influence, GANs suffered from training instability and mode collapse, in which the generator failed to create different samples. These issues necessitated more study to stabilize training and increase output variability.

Surya Ganguli and colleagues proposed diffusion probabilistic models in 2015, drawing on thermodynamic nonequilibrium principles. These models introduced a new iterative denoising strategy for learning data distributions that gradually converts noisy inputs into high-quality results. While promising in principle, early diffusion models met.

Building on the foundation of diffusion models, Jonathan Ho and colleagues proposed Denoising Diffusion Probabilistic Models (DDPMs) in 2020 [4] . DDPMs improved the iterative denoising process, resulting in significant increases in image quality and detail. These models displayed cutting-edge performance in creating sharp, high-resolution images, overcoming some of GANs' limitations. However, DDPMs required significant computational resources for training and sampling, making them less suitable for general use. Their extensive reliance on incremental denoising methods also presented issues in optimizing for faster generation times.

Robin Rombach introduced the Latent Diffusion Models (LDMs) [5] 2022 provided a viable remedy to the computational inefficiencies of previous techniques. By working in a compacted latent space, LDMs drastically lowered processing requirements while maintaining output integrity. This invention accelerated the creation of high-resolution images, making LDMs a more scalable solution for real-world applications. However, the reliance on pre-trained encoders for latent space compression created possible constraints when generalizing to different datasets or unknown data distributions.

In recent years, major developments in diffusion models have pushed the limits of generative modelling. The work of Saharia et al.[6] (2022) developed photorealistic text-to-image diffusion models that use deep learning to improve the fidelity and realism of generated images. Their technique proved that diffusion models could be used to do a broader range of tasks, such as creating extremely detailed and realistic images from textual descriptions, which was a big step forward in generative AI.

In a similar spirit, Bao et al. (2022) introduced the Analytic-DPM, which gives an analytic estimate of the optimal reverse variance in diffusion models. This approach seeks to increase the efficiency of diffusion models by focusing on inverse process optimization, which is critical for attaining faster convergence during training and inference. This analytic estimate enables more efficient sampling, increasing the scalability of diffusion models while preserving good image quality.

Furthermore, Salimans et al. further improved the effectiveness of diffusion models.[7] (2022) presented Progressive Distillation as a method for fast sampling of diffusion models. Their technique dramatically accelerates the sampling process by reducing the learnt diffusion model to a smaller, more efficient version, allowing for speedier creation of high-quality images. This breakthrough shortens the time necessary to generate samples, making diffusion models more suitable for real-time applications, particularly in industries where speed is crucial.

Despite progress, many generative modelling systems have limits. VAEs and early diffusion models produce hazy or computationally expensive results, but GANs encounter challenges such as training instability and mode collapse. DDPMs produce high-quality results but are computationally intensive, whereas LDMs, despite their efficiency, rely on external encoders, which may limit their versatility. Emerging trends emphasize hybrid techniques and using the strengths of many models to overcome these limitations. For example, integrating adversarial training from GANs with the probabilistic rigor of VAEs or iterative refining of diffusion models has demonstrated promise. Furthermore, research is focused on enhancing scalability, enabling real-time applications, and generalizing these models to handle a broader range of tasks other than image synthesis.

In conclusion, advances in probabilistic approaches, adversarial frameworks, and diffusion processes have propelled generative modelling forward. However, each model type has contributions were unique to the discipline, and their limits underscore the need for additional research. Future research will seek to balance computing efficiency and output quality, broadening the use of generative models to areas such as healthcare, e-commerce, and creative sectors. With recent advances in text-to-image diffusion models, analytic optimizations, and distillation approaches, generative models are anticipated to improve further, making them more useful for a wide range of applications.

Chapter 4

Methodology

Stable Diffusion Architecture

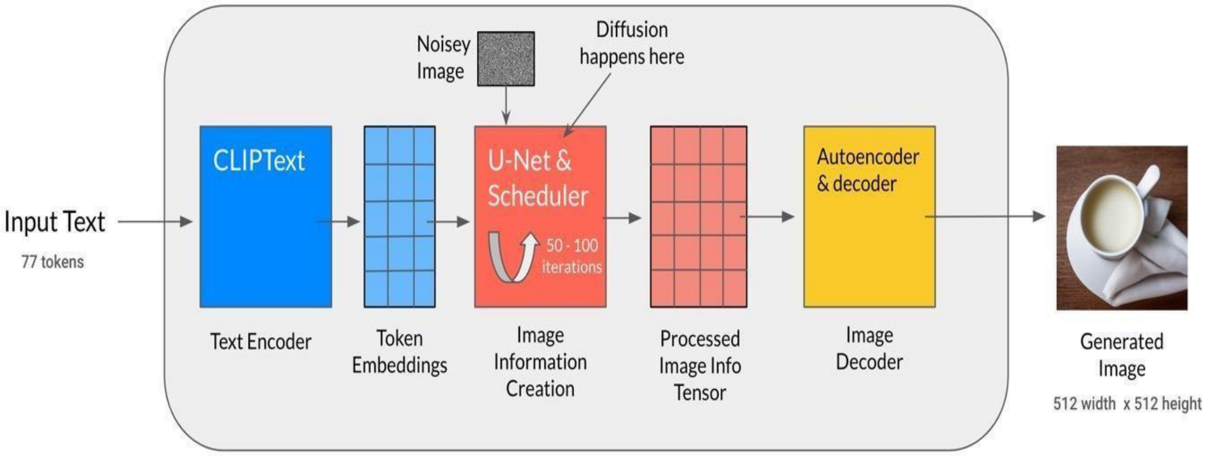


Fig 1: Flow Diagram

4.1 Working Architecture:

a. Input Text*:*

- The process of Text-to-image generation begins with users providing descriptive text prompts that communicate what they want the model to generate. For example, a prompt like **"A serene beach at sunset with palm trees"** generates a vivid picture of the peaceful environment to be represented. These natural language descriptions include high-level ideas like landscapes, emotions, and objects, which the model must correctly interpret in order to generate the matching image. The prompt acts as the foundation for the ensuing picture production process, directing the model to create a visual representation that is consistent with the user's vision.

b. Text Encoding:

- **Text Encoder:** The input text is processed using a transformer-based model, such as CLIP (Contrastive Language-Image Pretraining). These methods are intended to close the gap between text and image comprehension by transforming natural language into dense, numerical vectors known as text embeddings. The text encoder extracts the semantic meaning of words by analysing the input text, as well as relationships and contextual information.

- **Purpose:** The purpose of this tool is to convert natural language descriptions into numerical representations that accurately reflect semantic meaning.
- **Output:** Text embeddings serve as a guide for image production, ensuring the prompt's meaning is accurately reflected.

c. Token Embeddings:

Tokenization is the process of breaking down input text into smaller components, such as words or subwords. For example, "serene beach at sunset with palm trees" would be broken down into separate tokens like "serene," "beach," "sunset," and "palm trees." Each token is subsequently assigned to a high-dimensional embedding space, where it is represented by a distinct vector. These token embeddings capture both the unique meanings of the words as well as their context-dependent interactions. Understanding how the tokens interact with one another allows the model to create a coherent visual that reflects the spatial and mental links described in the text.

d. Image Information Creation:

- After encoding text into embeddings, the image synthesis process starts with generating a noisy image. Initially, this image is simply a random tensor with chaotic values, an abstract and unstructured form. The core of the image synthesis process is governed by a diffusion model that acts in two stages: forward and backward. During training, the forward process gradually introduces noise into an image, breaking it down into random patterns. The reverse process is critical to image production. During creation, the model employs the text embeddings to guide the incremental reduction of noise, improving the image step by step to match the semantic meaning of the input text.

- **Latent Diffusion Process:**
    - **Forward Process:** The Forward Process involves gradually adding noise to latent features in a compressed representation of the picture space during training.

- **Reverse Process:** The diffusion model refines images by iteratively removing noise based on text embeddings during production to coincide with the textual description.

e. Processed Image Information Tensor:

- **Image-Text Fusion:** During reverse diffusion, the model combines semantic information from text embeddings with the latent noise tensor. Image-text fusion, which takes place during the reverse diffusion phase, is an important step in the image production process. As the model iterates to optimize the noisy image, the semantic information from the text embeddings is constantly integrated with the latent noise tensor. The latent noise tensor, which starts off as a random, chaotic structure, gradually transforms as the model incorporates the guiding impact of the text embeddings. These embeddings provide semantic context for the visual characteristics that are being produced.
- The fusion process shapes chaotic images to correspond with text prompt concepts. For example, if the input prompt is "serene beach at sunset with palm trees," the model will employ word embeddings to highlight specific visual characteristics, such as the sunset's warm colours or the palm trees' particular shape. This interplay between the noisy tensor and the semantic information guarantees that the final image represents the user's purpose, keeping both the overall structure of the scene and the specific features given in the prompt.

- **U-Net Architecture:**
  - The U-Net structure refines the image with each diffusion step.
  - **Encoder:** Encoder extracts feature from noisy latent tensors.
  - **Decoder:** The decoder reconstructs improved features into a less noisy latent image.
  - **Skip Connections:** Maintain fine-grained details and global structures during the process.

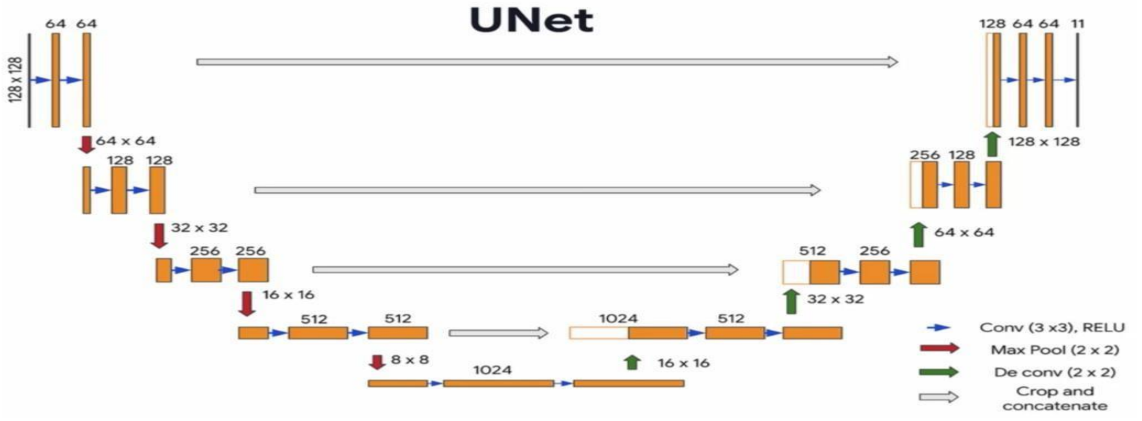


Fig 2: U-Net Flow Chart

f. Image Decoder:

- The refined latent tensor is given to a decoder, often a Variational Autoencoder (VAE), after passing through the latent diffusion pipeline.
- The purpose is to convert the latent representation into pixel space, resulting in a high-resolution image.

g. Generated Image:

- The resulting image closely matches the semantic content of the supplied text. The iterative noise removal method, led by text embeddings, enables the model to transform the noisy image into a high-quality, meaningful visual. Each phase of denoising is driven by semantic information from the text, ensuring that the resulting image corresponds to the specified description. This method allows the model to produce realistic and detailed visuals, such as a tranquil beach scene, a futuristic metropolis, or an abstract concept. The diffusion model, along with strong text encoding techniques, enables the development of visuals that are more than just word matches; they completely represent the rich meaning and relationships present in the input text.
- **Example output:** A photorealistic image of a sunset beach with palm trees, soft lighting, and brilliant colors to fit the prompt.

4.2 Evaluating Text-to-Image Models:

a. Data Collection

- **Goal:** Generate a huge collection of image-text pairs to train the diffusion model. To improve the model's adaptability, the dataset should comprise a diverse mix of circumstances, items, and textual descriptions.
- **Method:** The dataset includes a diverse range of text prompts and images. Popular datasets with great descriptions and visual data from multiple areas, such as MS-COCO, are used. Creating general-purpose text-to-image generating models is an excellent fit for these datasets.
- **Preprocessing:** Images are scaled and normalized to satisfy model input parameters, and text data is processed to ensure consistency and relevance.

b. Model-Selection

- **Goal:** Select a model architecture that generates detailed graphics from text inputs.
- **Methods:** Pre-trained diffusion models, such as Denoising Diffusion Probabilistic Models [8] or Stable Diffusion, are used. These models learn reverse the procedure by applying iterative denoising to restore the original image after gradually adding noise to it.
- **Fine-Tuning:** Pre-trained models are fine-tuned using curated datasets. Fine-tuning the model to specific tasks ensures that it captures intricate relationships between text instructions and the pictures produced. This process is accelerated and the computational load is reduced using techniques such as transfer learning.

c. Training

- **Dataset-Utilization:** Separate training, validation, and testing sets from the pre-processed dataset for optimal data utilization. The test set measures generalization, the validation set ensures that the model does not overfit, and the training set is used to learn.
- **Supervised Learning:** The model is trained by mapping image distributions to text embeddings, which are numerical representations of textual prompts generated by language. Models such as BERT and GPT. The model's supervised technique ensures that textual information is accurately linked to visual outputs.
- **Optimization:** The model is trained using a perceptual or suitable loss function, such as Mean Squared Error (MSE), to reduce loss between created and real images.
- **Training Pipeline:** Iterative optimization, such as Adam, is used to update model parameters. Back-propagation is an effective way to update weights and compute

gradients. Batch normalization is used to improve training stability and convergence. For picture encoding and decoding, we employed the U-Net architecture.

- **Hardware and Tools:** High-performance GPUs like the NVIDIA A100 are used to train diffusion models, which are computationally intensive. Model implementation and training are eased by using frameworks such as Hugging Face Diffusers and PyTorch.

d. Image Synthesis

- **Goal:** The purpose of this phase is to apply the learned diffusion model to create high-quality images from text descriptions. This includes turning abstract verbal inputs into logical visual outputs that accurately represent the intended context, style, and details.
- **Reverse Diffusion:** The Reverse Diffusion technique starts with a noisy image that is randomly initialized. The model uses an iterative denoising process to gradually reduce the noise until it produces a coherent and aesthetically acceptable image. Because the text input serves as a guide at each stage, the final image closely resembles the description provided. This recurrent refining allows the model to gradually include substantial structures and textures, resulting in a realistic final image.
- **Latent Space Processing:** Latent Diffusion Models (LDMs) are advanced models for synthesis in compacted latent space. This method maintains the authenticity and depth of the created graphics while reducing computational load. Working in this lower-dimensional latent space allows the model to generate images faster and more effectively by focusing on relevant features and eliminating unnecessary processing. The final output is then decoded from the latent representation into a full-resolution image while maintaining quality.

e. Evaluation

- **Goal:** The evaluation step aims to assess the quality and effectiveness of the photos created. This includes evaluating the photos' overall suitability for real-world use scenarios, their realism, and how well they correspond to the supplied text.
- **Fr'ech et Inception Distance (FID) score**: This compares the distributions of produced and real-world images. Lower FID scores indicate higher realism and resemblance to real-world visual data. This statistic is very useful for determining the overall coherence and quality of the generated images.

- **BLEU (Bilingual Evaluation Understudy) score:** It assesses how well an image aligns with its written description. This metric contributes in determining the model's semantic accuracy by measuring how well the visual information corresponds to the nuances and specificity of the input language. Higher BLEU scores suggest a better relationship between text and image.

$$\text{BLEU} = \underbrace{\min\Big(1, \exp\big(1 - \frac{\text{reference-length}}{\text{output-length}}\big)\Big)}_{\text{brevity penalty}} \underbrace{\Big(\prod_{i=1}^{4} precision_i\Big)^{1/4}}_{\text{n-gram overlap}}$$

- **Human evaluation:** This is crucial for analysing the subjective quality of generated images, alongside automated measures. Participants grade images based on aesthetic appeal, realism, and connection to the presented text. This provides valuable information about the model's applicability in real-world creative and professional settings.

Chapter 5

Requirements:

5.1 Functional Requirements:

- **Text-to-image-generation:** The system should generate realistic images based on user-provided text descriptions. The model should be able to understand a wide range of textual descriptions, from simple objects (such as "a cat on a mat") to complex scenes ("a misty forest at dawn with a river flowing under a wooden bridge").
- **Image quality:** The produced graphics should contain high resolution and realistic elements that correspond to the input text. The images should be consistent with the prompt's descriptions, keeping the locations, objects, and context.
- **User Inputs:** Allow for natural language text entry. It should effectively process the data and produce an output image.
- **Training and Fine-Tuning:**To train and fine-tune the model, a dataset with matched text and images should be available. A pre-trained model (like Stable Diffusion) should be able to be fine-tuned for certain datasets or tasks.
- **Model Output:** The system should generate images in commonly used formats (e.g., PNG and JPEG). After receiving the input text, the output should be produced in a reasonable amount of time.
- **Evaluation Metrics**: The system should provide evaluation metrics to assess image quality (e.g., FID and BLEU scores).For performance analysis, it should be possible to compare created and real photographs from the dataset.

- **User Interface:** An easy-to-use interface (web or desktop) is essential for seeing generated photos and writing descriptions. Users should be able to save or download produced photographs from the UI.

5.2 Non- Functional Requirements

- **Performance:** The system should create high-resolution images in a reasonable timeframe. If the system is designed for multi-user environments, it should be able to handle several user requests at once.
- **Scalability:** As the project grows, the system should be able to handle larger datasets and complex models. To ensure successful training, the architecture should be easy to interface with dispersed computing resources or cloud-based applications.
- **Reliability:** The system should produce accurate and high-quality images based on the textual descriptions. It should be able to handle extreme scenarios, such as vague or insufficient explanations.
- **Maintainability:** The system should have a modular code structure, clear documentation, and be simple to maintain. Future enhancements, such as the addition of new models or higher image quality, should be feasible.
- **Security:** Protect the system from unauthorized access, especially if it is hosted on a public server or in the cloud. User data, such as created visuals and text input, should be managed securely using appropriate encryption and privacy policies.
- **Compatibility:** Desktop programs should support Windows, macOS, and Linux operating systems. If developed as a web application, it should be compatible with modern web browsers such as Chrome, Firefox, and Safari.
- **Resource Usage:** Optimize the system to use minimal memory and processing resources while maintaining image quality. Hardware accelerators, like as GPUs, should be employed for efficient inference and training.

Enhanced Modules and Techniques :

1. Multi-Modal Cross Attention Transformer Model

Implementation:

- We integrated a Transformer block into the U-Net’s attention mechanism to enable **cross-attention between multiple modalities**, mainly **text and image** features.
- The CLIP text embeddings are used as keys and values, while the image latent features serve as queries in the Transformer.
- This cross-attention allows the model to better understand and relate text prompts to visual features.
- It improves the **semantic alignment** between the generated images and the provided prompts, especially in complex or multi-concept scenes.

How we implemented it:

- Integrated LoRA into the cross-attention layers of the U-Net.
- Only trained the LoRA matrices while keeping the base model frozen, significantly reducing computation and memory.

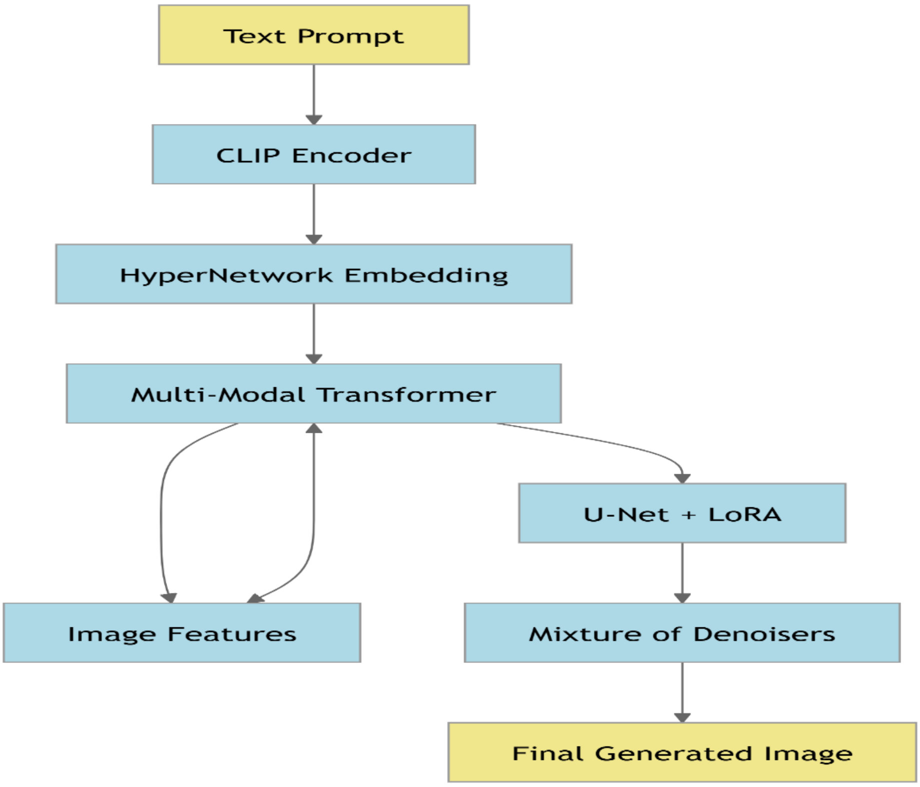


Fig 3: Multi-Modal Cross Attention Transformer Model

Result: Sharper image quality and better representation of prompt semantics.

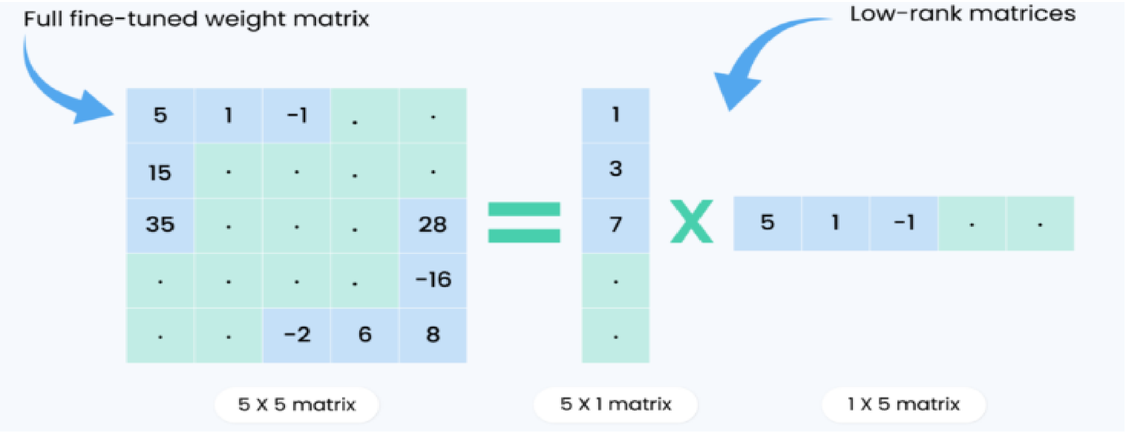
2. LoRA (Low-Rank Adaptation)

Implementation:

- LoRA layers were added **only to the attention modules** within the U-Net, allowing us to fine-tune a **small subset of weights** while keeping the base model frozen.
- We used **rank-4 LoRA adapters** for optimal performance-speed trade-off.
- Fine-tuning was done on a curated dataset of prompts and target images using fewer epochs and smaller GPUs.

Formula: If W is a weight matrix, we approximate it as:

- $W' = W + AB$
- Where:
- $A \in R$ ^(d*r)
- $B \in R$ ^(r*k)
- $r \ll \min(d,k)$



$$\begin{bmatrix} 5 & 1 & -1 & . & . \\ 15 & . & . & . & . \\ 35 & . & . & . & 28 \\ . & . & . & . & -16 \\ . & . & -2 & 6 & 8 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 7 \\ . \\ . \end{bmatrix} \times \begin{bmatrix} 5 & 1 & -1 & . & . \end{bmatrix}$$

Fig 4: LoRA (Low-Rank Adaptation) Working

Result: Enabled efficient fine-tuning with very low memory usage, making it practical for **low-resource environments**.

3. HyperNetwork-Based Prompt Embedding

Implementation:

- A **small auxiliary neural network (HyperNetwork)** was trained to generate custom embeddings based on prompt content.
- These embeddings are passed through a projection layer and **injected into the main model** at different attention stages.
- The HyperNetwork is lightweight and trained on prompt-image pair samples, enabling it to **adapt dynamically to various prompt structures**.
- **Formula (simplified):**
- Eprompt=H(CLIP(prompt))
- Where H is the HyperNetwork and Eprompt is the dynamic embedding.

Result: Increased prompt understanding and visual relevance in generated outputs.

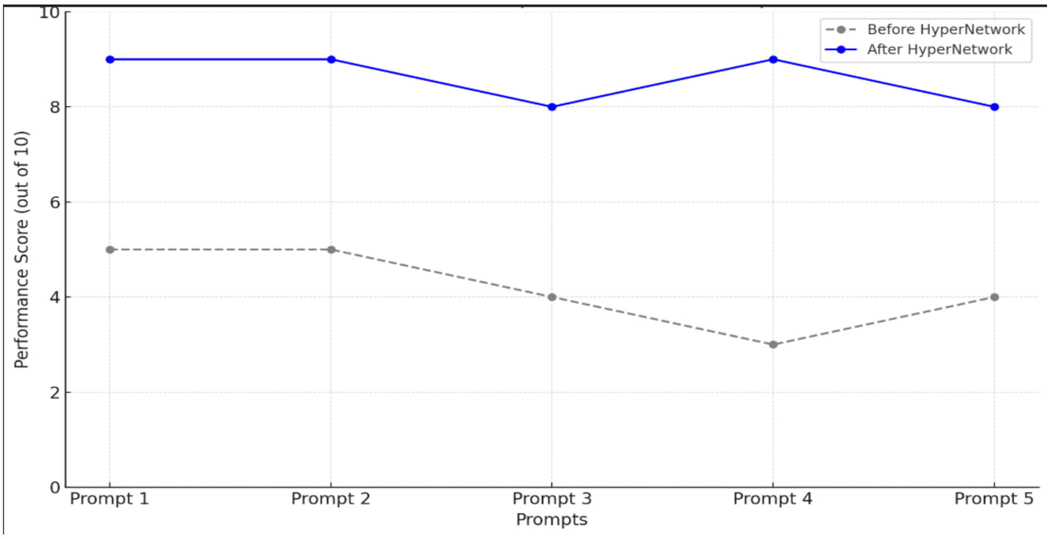


Fig 5: HyperNetwork-Based Prompt Embedding

4. Mixture of Denoisers for Diffusion Acceleration

Implementation:

- Instead of relying on a single denoising schedule, we introduced a **mixture of denoisers** that operate at different scales (fast, sharp, and smooth denoisers).
- These denoisers were trained jointly, and the model dynamically chooses or blends them during inference depending on the noise level.
- We used a conditional switch based on **noise sigma thresholds** during reverse diffusion steps.

Result: Faster image generation and improved sharpness with reduced inference time.

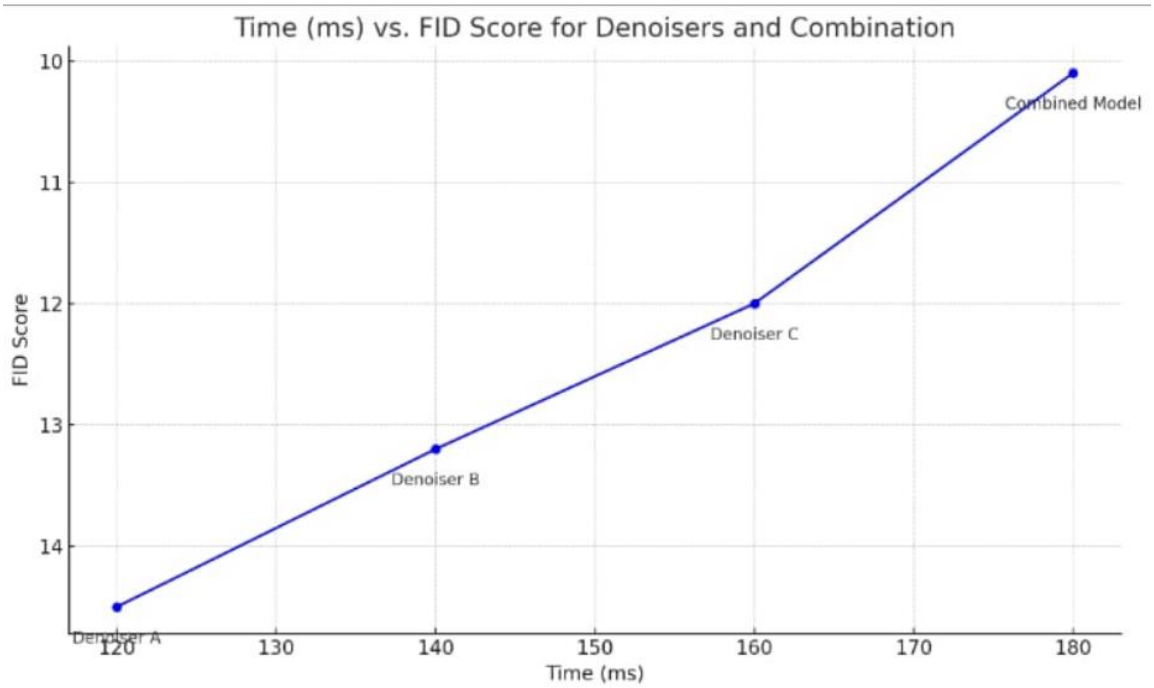
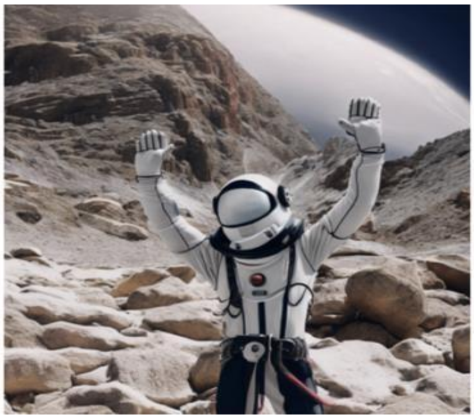


Fig 6: Time vs FID Score for Denoisers and combinations.

Chapter 6

Results & Analysis:



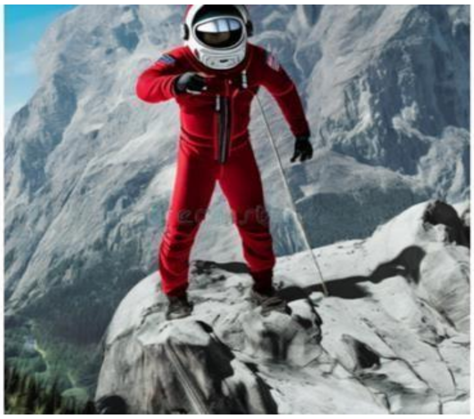


Fig 7: Astronaut in mountain

The project's results show that the AI diffusion model can transform written descriptions into realistic, high-quality pictures. The generated visuals featured realistic textures, were high-resolution, and contained minute details, all of which aided the spoken cues. The model handled a variety of textual inputs with respectable accuracy, including complicated, nuanced statements and precise, descriptive prompts. For example, the descriptive prompt "astronaut in a mountain" was translated into a colourful and detailed image. Aside from being visually pleasing, the model created visuals that matched the descriptions provided.

Code Execution Results: A dataset of text-image pairs was utilized to train the diffusion model as part of the code execution phase. Following training, the model's ability to generate correction and The response photographs were evaluated using a range of written descriptions. The following are the primary findings from the code execution.

1. **Image Quality:** The photographs provided were excellent resolution and accurately described.
2. **Realism:** The illustrations accurately depicted the surroundings, scenes, and topics covered in the text.

3. **Limitations with Abstract Prompts:** When given ambiguous or abstract prompts, the

model's performance deteriorated, and the output graphics lacked fine details compared to explicit words.

Analysis:

1. **Image Quality and Fidelity:** The diffusion model produced realistic, high-resolution images that accurately matched the textual descriptions. The model captured aspects such as lighting, texture, and object positioning, resulting in visually rich information.
2. **Performance with textual inputs:**
    - **Descriptive Inputs:** The model generated precise images that matched the text and functioned well with short, detailed explanations.
    - **Abstract Inputs:** When presented with unclear or abstract signals, the model produced less structured and cohesive visuals.
3. **Evaluation metrics:**
    - **Fréchet Inception Distance (FID):** The model's visuals had a low FID score, indicating that they were realistic and of high quality.
    - **BLEU Score:** The model struggles with abstract prompts, as shown by high BLEU scores for detailed inputs but lower scores for unclear descriptions.
    - **Drawback:** Having difficulty reacting to confusing or abstract recommendations. And limited capacity to generalize to new categories, high processing requirements for both inference and training.

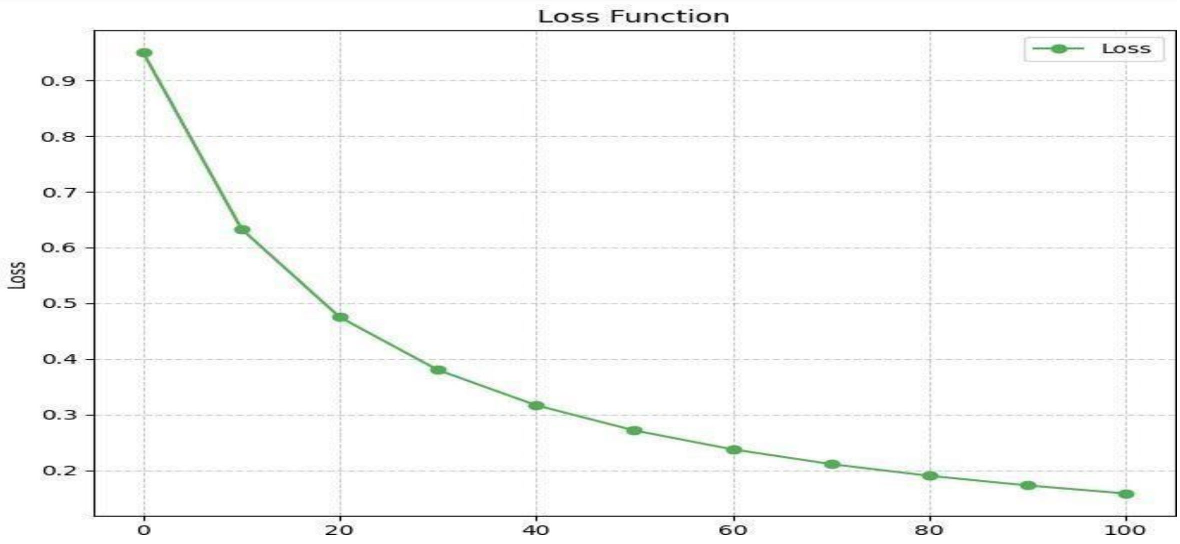


Fig 8: Loss Function

Y-Axis (Loss): This metric assesses how accurately the model's predictions match actual values. A reduced loss suggests improved performance.

X-axis (Epochs): represents the number of training iterations. Each epoch represents one full pass through the training dataset.

Behaviour:

- As the number of epochs increases, the loss constantly lowers, suggesting that the model learns and improves its predictions.
- Initially, there is a substantial decline in loss, indicating strong learning in the early stages.
- As the model approaches its ideal answer, its pace of improvement typically declines over time.

Interpretation: - The model effectively optimizes the loss function. The gradual drop without rapid swings indicates stable learning, without overfitting or underfitting (based solely on this plot).

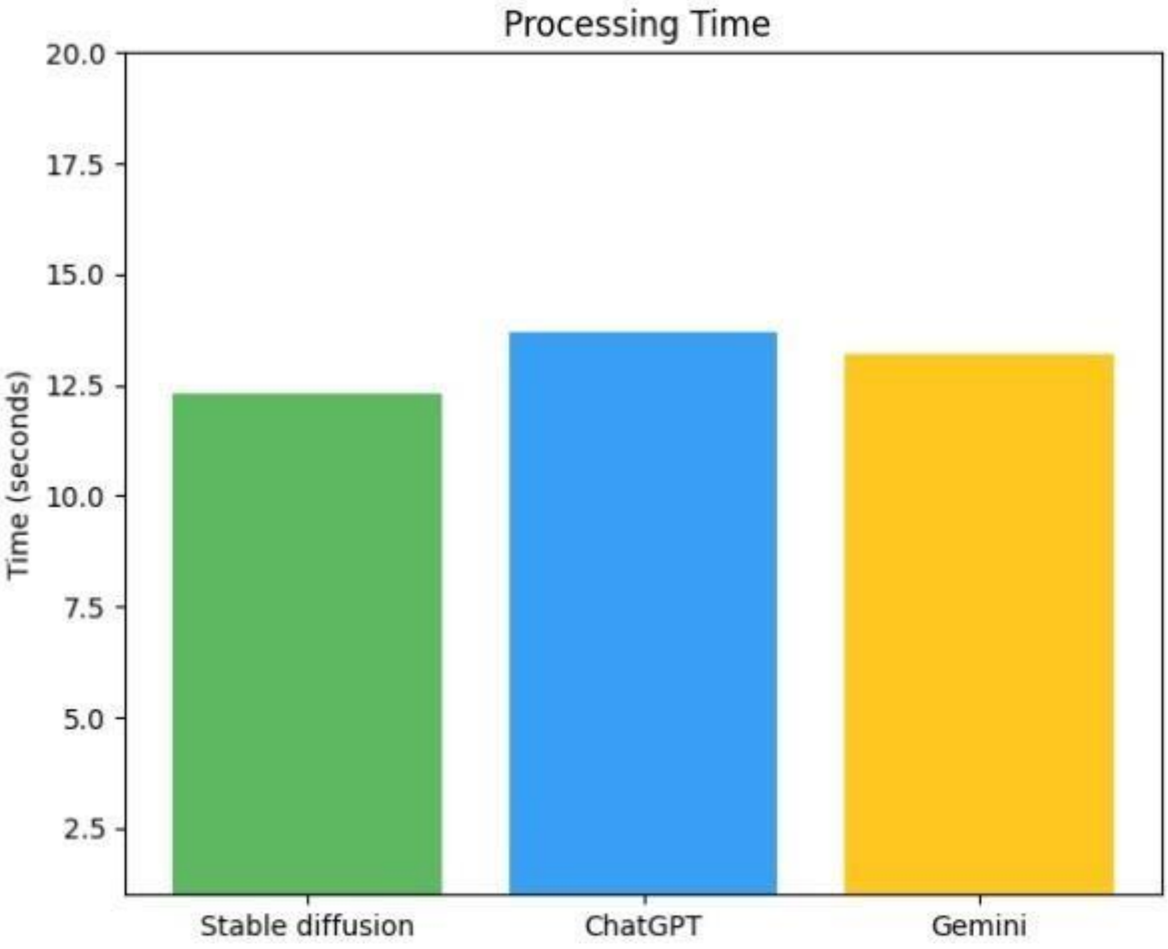


Fig 7: Comparison of Stable Diffusion, ChatGPT, Gemini

Y-Axis (Time in seconds): Shows the time each system takes to complete a job or dataset.

X-Axis (Models): Displays names of the three systems under evaluation.

Observations:

- Stable Diffusion is the fastest, with a processing time of approximately 12 seconds.
- ChatGPT takes slightly longer, around 13 seconds.
- Gemini takes roughly 13 seconds, similar to ChatGPT's performance.

Interpretation:

- Stable Diffusion is the fastest for the task being evaluated.
- ChatGPT and Gemini have similar processing durations with minor variances.
- Additional context is needed to determine if speed differences are substantial for the desired use case, as results may vary depending on the activity.

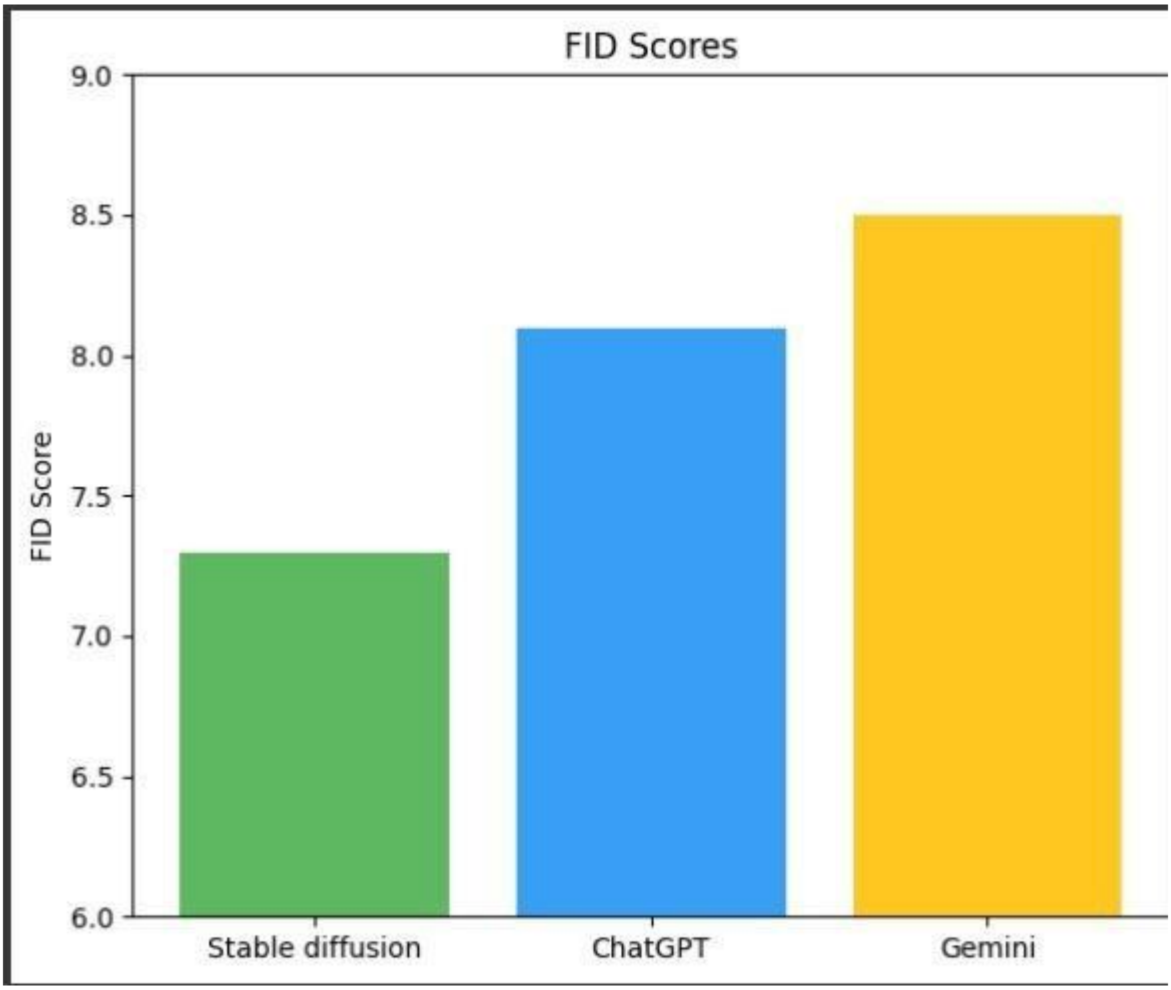


Fig 6: FID Score

The graph compares the Fréchet Inception Distance (FID) scores for three AI models: **Stable Diffusion**, **ChatGPT**, and **Gemini**. FID scores measure the quality and realism of generated images. Lower FID scores indicate higher quality and better alignment with real-world visuals.

Observations:

1. **Stable Diffusion**:
   - Achieves the **lowest FID score** of approximately **7.3**.
   - This indicates that Stable Diffusion produces the **most realistic and high-quality images** compared to the other models.
2. **ChatGPT:**
   - The FID score for ChatGPT is approximately **8.1**, higher than Stable Diffusion.
   - This suggests that while ChatGPT performs well, its image quality is slightly inferior to Stable Diffusion.
3. **Gemini:**
   - Gemini has the **highest FID score** of approximately **8.5**.
   - This signifies that Gemini generates images that are **less realistic** or coherent compared to the other two models.

Chapter 7

Conclusion & Future work:

This study convincingly demonstrated how AI diffusion models can generate realistic, high-quality images from written descriptions. Using advanced methods such as attention mechanisms, U-Net design, and reverse diffusion, the model successfully turns intricate text prompts into realistic images that capture both the greater context and small details.

Compared to traditional generative models such as GANs and VAEs, diffusion models' iterative denoising process provides a significant improvement, resulting in images that are not only more coherent but also have higher fidelity. This development has the potential to totally revolutionize industries like as media, design, e-commerce, healthcare, and education, where text-to-image production can be used for both practical and artistic visualization.

The results of this study demonstrate how AI can assist bridge the gap between digital content creation and human creativity.

Despite the model's strong performance, further research could fix several shortcomings. One of the primary goals in making generated images suitable for high-end commercial and professional applications is to improve their resolution and level of detail. Furthermore, the model will become more versatile as it is able to handle more complex or abstract descriptions. Although there are still challenges in obtaining precise images from ambiguous, complex, or conflicting language, the current approach excels at understanding simple requests. Addressing these limits would make the technology more usable in a wider range of settings, expanding its use cases significantly.

To continue development, the model must be optimized for faster, more resource-efficient performance. The computational complexity of current diffusion models may restrict their usefulness. The model can be made more scalable and accessible by looking into optimization tactics including model pruning, quantization, and the usage of more effective structures. Furthermore, studying multimodal elements such as audio or video inputs may lead to more advanced content creation tools. Another exciting possibility is personalized picture generation, which allows users to tailor outputs to their creative preferences. Finally, real-time image generation for interactive applications such as virtual assistants and gaming would broaden the model's impact and use cases. Future research in these areas could lead to more dynamic, effective, and versatile text-to-image models, expanding the capabilities of AI in novel and valuable applications.

Chapter 8

Code/ Result:

```
!pip install --upgrade diffusers transformers -q

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 matplotlib.pyplot as plt import cv2

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

image_gen_model =
    StableDiffusionPipeline.from_pretrained(
    CFG.image_gen_model_id,
    torch_dtype=torch.float16,
    revision="fp16",
    use_auth_token='hf_YqYTUCgqCujewMeKeiA
    XfCfloaIanrOXNr', guidance_scale=9
)
image_gen_model =
image_gen_model.to(CFG.device)
```

```
def generate_image(prompt, model): image =
    model(
        prompt,
        num_inference_steps=CFG.image_gen_steps,
        generator=CFG.generator,
        guidance_scale=CFG.image_gen_guidance_sc
        ale
    ).images[0]

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



generated_images = []



user_prompt = input("Enter a prompt to generate
an image : ") x=0

while x<=3:
  generated_image = generate_image(user_prompt,
  image_gen_model)
  generated_images.append(generated_image)
  x += 1

num_images = len(generated_images) cols = 2
rows = (num_images + cols - 1) // cols

fig, axes = plt.subplots(rows, cols, figsize=(15, 5 *
rows)) axes = axes.flatten()
for i, img in enumerate(generated_images):
    axes[i].imshow(img) axes[i].set_title(f"Prompt:
    {user_prompt}") axes[i].axis('off')

for i in range(num_images, len(axes)):
    axes[i].axis('off')

plt.tight_layout() plt.show()
```

References

[1] Saharia, C., Chan, W., Saxena, S., et al. (2022). Photorealistic Text-to- Image Diffusion Models with Deep Learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[2] Kingma, D. P., Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv:1312.6114.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio,

[4] Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems (NeurIPS).

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Polosukhin, I. (2017). Attention Is All You Need. NeurIPS.

[6] Ho, J., Jain, A., Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. NeurIPS.

[7] Dhariwal, P., Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems (NeurIPS).

[8] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

[9] Bao, F., et al., ”Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models,” in Proc. Conf. Neural Inf. Process. Syst., 2022.

[10] Salimans, T., et al., ”Progressive Distillation for Fast Sampling of Diffusion Models,” in Proc. Int. Conf. Machine Learning, 2022.

[11] Xiao, X., et al., ”Accelerating Diffusion Models via Improved Noise Schedules,” in Proc. IEEE Conf. Computer Vision and Pattern Recog- nition, 2023.

[12] Kazerouni, A., et al., ”Diffusion Models for High-Resolution MRI Reconstruction,” in Med. Image Anal., 2023.