A Major Project Report

On

IMAGE GENERATOR HARNESSING STABLE DIFFUSION

Submitted for partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

In

ELECTRONICS AND COMMUNICATION ENGINEERING

SUBMITTED BY

P. AMULYA	**21271A0405**
K. BHAVANI	**21271A0408**
D. JAHNAVI	**21271A0417**
N. SATHWIK	**21271A0440**

Under the guidance of

Mr. J. Ramesh

Assistant Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

JYOTHISHMATHI INSTITUTE TECHNOLOGY AND SCIENCE

(AUTONOMOUS)

(Approved By AICTE, accredited with NAAC'A' Grade, NBA and Affiliated to JNTUH)

2021-2025

JYOTHISHMATHI INSTITUTE OF TECHNOLOGY & SCIENCE (AUTONOMOUS)
Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad
Accredited with NAAC 'A' Grade | ISO 9001:2015 Certified
UG - CSE,ECE,EEE Accredited by National Board of Accreditation, New Delhi

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

This is to certify that the project work entitled **“IMAGE GENERATOR HARNESSING STABLE DIFFUSION”** is a Bonafide work carried out by **P. Amulya (21271A0405), K. Bhavani (21271A0408), D. Jahnavi (21271A0417), N. Sathwik (21271A0440)** in partial fulfillment of the requirements for the degree of “**BACHELOR OF TECHNOLOGY**“ In “***Electronics and Communication Engineering”*** from the Jyothishmathi Institution of Technology and Science, (Autonomous) during the academic year 2021-25.

No part of this report has been submitted elsewhere for award of any other degree

Submission for major project viva voce examination held on__________

Project guide	**Head of the Department**
Mr. J. RAMESH	**Dr. N. UMAPATHI**
Assistant professor	**Professor & Head**

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

The development of the project though it was an arduous task, it has been made by the help of many people. we pleased to express our thanks to the people whose suggestions, comments, criticisms greatly encouraged in betterment of the project.

We would like to express our sincere thanks and profound gratitude to **DR. T. ANIL KUMAR**, principal, **Jyothishmathi Institute of Technology and Science,** for his support, guidance and encouragement in the course of our project.

We also thankful to the Head of the Department **Mr. N. UMAPATHI**, Professor & Head Electronics & Communication Engineering Department for providing excellent infrastructure and a nice atmosphere for completing this project successfully.

We would like to express our sincere gratitude and indebtedness to our project guide **MR. J. RAMESH,** Assistant Professor**,** for his valuable suggestions and interest throughout the course of this project.

We highly thankful to the Project Coordinator**, D. RAVIKIRAN BABU,** Associate Professor for their valuable suggestions, encouragement, and motivations for completing this project successfully.

We thankful to all other faculty members for their encouragement.

We convey our heartfelt thanks to the lab staff for allowing us to use the required equipment whenever needed.

Finally, We would like to take this opportunity to thank our family for their support through the work. We sincerely acknowledge and thank all those who gave directly or indirectly their support in completion of this work.

P. AMULYA (21271A0405)
K. BHAVANI (21271A0408)
D. JAHNAVI (21271A0417)
N. SATHWIK (21271A0440)

DECLARATION

This is to certify that the work reported in the present project entitled **"IMAGE GENERATOR HARNESSING STABLE DIFFUSION"** is a record of work done by us in the partial fulfilment for the award of the degree of ***Bachelor of Technology*** in ***Electronic & Communication Engineering, Jyothishmathi Institute of Technology and Science (Autonomous),*** affiliated to JNTUH, Accredited By NAAC and NBA, under the guidance of **Mr. J. Ramesh, Assistant Professor,** ECE Department. We hereby declare that this project work bears no resemblance to any other project submitted at Jyothishmathi Institute of Technology and Science (Autonomous) or any other university/college for the award of the degree. The conclusion and results in this report are based on our own.

P. AMULYA (21271A0405)

K. BHAVANI (21271A0408)

D. JAHNAVI (21271A0417)

N. SATHWIK (21271A0440)

ABSTRACT

In recent years, the advancement of artificial intelligence has led to remarkable progress in generating realistic images from textual descriptions. This project introduces “Stable Diffusion”, an innovative text-to-image synthesis model that achieves photorealistic image generation through a unique iterative refinement process. Stable diffusion employs a stepwise approach, gradually enhancing a random noise image while aligning it with the given text prompt. Text to image synthesis refers to the method of generating images from the input text automatically. This iterative process continues until convergence, yielding high-quality images that faithfully represent the text description. By utilizing the Stable Diffusion AI model for image generation, our system can generate realistic and accurate facial images of suspects based on input text descriptions, improving the accuracy and efficiency of suspect identification by law enforcement agencies.

The image generator powered by Stable Diffusion is an advanced AI tool that creates high-quality, visually striking images based on textual descriptions. This model leverages deep learning techniques to generate abstract and creative visual representations from simple prompts, allowing users to explore a wide range of artistic styles, from surreal landscapes to geometric patterns. By interpreting text input, Stable Diffusion can produce images that convey emotions, concepts, or moods in an abstract form, often blending colors, shapes, and textures in ways that are unexpected and thought-provoking. It provides a powerful platform for artists, designers, and creators, offering endless possibilities for digital art and conceptual imagery, all while maintaining a high level of detail and creativity**.**

LIST OF FIGURES

TABLE OF CONTENTS

Page No

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER-1

INTRODUCTION

In recent years, artificial intelligence has made significant strides, particularly in converting textual input into highly realistic images. This project introduces Stable Diffusion, a cutting-edge model designed for generating images from text. It stands out by using an iterative refinement method to transform random noise into detailed, photorealistic visuals based on textual descriptions .The model is trained on an extensive and diverse image dataset and uses a pre-trained text encoder to guide image creation. Through a progressive enhancement process, Stable Diffusion starts with a noise-filled image and gradually adjusts it to align with the input prompt. This continues until a final, high-quality image is produced that accurately reflects the given text.Stable Diffusion demonstrates strong performance across various subjects, including human figures, animals, nature scenes, and abstract compositions. Its ability to interpret and visualize complex text enables it to create expressive, detailed images that go beyond basic illustration . One notable area where text-to-image generation has great promise is in education. By converting phrases into relevant images, the technology can help children grasp language concepts more effectively, enhancing learning through visual association. In the domain of forensic science, artificial intelligence offers powerful tools. It can assist investigations by processing and interpreting large volumes of visual and textual data. AI can detect irregularities, recognize faces, identify key objects, and track movement in videos. Additionally, natural language processing enables analysis of communications such as emails, messages, and social media—to extract potential evidence.

AI can also analyze extensive data sets to find patterns or connections that may not be immediately visible to human investigators and even forecast possible future incidents. In cyber forensics, AI tools can sift through digital records to uncover signs of cybercrime.However, integrating AI into forensic processes raises several ethical and legal challenges. These include concerns over individual privacy, data protection, biases in algorithmic decisions, and whether AI-derived evidence is permissible in court. There’s also the risk of misuse, especially in surveillance or misinformation through AI-generated content. Addressing these challenges requires strict ethical guidelines, legal regulations, and continuous human oversight to ensure responsible use of the technology.Artificial Intelligence (AI) is playing a transformative role in forensic sketching by helping create realistic facial images of suspects based on descriptions given by witnesses. This process uses intelligent algorithms that translate verbal or written inputs into digital sketches that resemble potential suspects. Additionally, AI can produce age-progressed images of missing individuals and compare generated images with existing databases to find possible matches. These capabilities not only reduce the time and effort required but also improve the detail and accuracy of visual suspect representations.However, the use of AI in this field comes with challenges. There are concerns about fairness and precision, as AI models may misinterpret subtle human features or reflect biases based on the gender or ethnicity of the subject. This can lead to inaccurate outcomes or unjust profiling. Beyond image generation, AI is also capable of analysing visual and textual evidence detecting faces, recognizing patterns, monitoring movements, and extracting insights from digital communication such as emails, chats, and social media posts.

Traditional methods for suspect identification often rely on models like generative adversarial networks (GANs) or transformer-based systems. In contrast, the proposed system utilizes a stable diffusion-based image generation model, aiming to provide law enforcement with a faster and more reliable tool for creating suspect images from text-based descriptions. The system is composed of four modules: input processing, the diffusion model, image creation, and a user interface. By converting descriptive narratives into visual content, this system mirrors the way people naturally imagine stories, making the identification process more intuitive and effective.The proverb "a picture is worth a thousand words" underscores the profound impact that visual information can have in conveying meaning. When humans read textual narratives, they often construct mental images that enrich understanding and enhance emotional engagement. Replicating this imaginative capability in machines specifically, generating realistic images from textual descriptions remains a challenging and significant task in artificial intelligence. This task, known as text-to-image synthesis, represents a key milestone toward building systems with human-like perceptual and generative abilities.The field has evolved rapidly with advances in deep learning. Early work such as AlignDRAW pioneered text-to-image generation but produced low-quality visuals. Subsequent models, like text-conditional GANs, introduced end-to-end trainable architectures capable of converting text directly into images. These models, however, were largely constrained by small datasets and struggled to produce diverse or highly detailed results.To overcome these limitations, autoregressive models like DALL·E (OpenAI) and Parti (Google) began leveraging large-scale datasets, significantly improving output quality. Nonetheless, their sequential nature led to high computational costs and susceptibility to error accumulation during generation.More recently, diffusion models have emerged as the state-of-the-art approach in this domain, delivering highly realistic and coherent images from text prompts. Their success has triggered a wave of research and social media interest, leading to a proliferation of methods built upon diffusion-based architectures.Despite the growing body of work, there is currently no survey that focuses solely on diffusion-based text-to-image generation. Existing reviews either broadly cover diffusion models across multiple domains or focus exclusively on earlier GAN-based methods, leaving a clear gap in the literature.

1.1 BACKGROUND ON DIFFUSION MODEL:

Diffusion Models (DMs), also referred to as diffusion probabilistic models, represent a class of generative models grounded in Markov chains and trained through variational inference. Their fundamental concept involves learning to reverse a gradual noising process—termed diffusion—which perturbs data over time. Once trained, the model can generate new samples by reversing this process, effectively denoising random noise into structured data.A key breakthrough in this domain came with the introduction of the Denoising Diffusion Probabilistic Model (DDPM) in 2020. This work significantly advanced the field and triggered a surge of research interest in diffusion-based generative techniques. To build a solid understanding of DDPM, it is essential to first examine prior foundational methods and then explore how DDPM functions in the context of unconditional image generation.In addition, this section highlights the concept of guidance, which plays a critical role in extending diffusion models to conditional generation tasks. This is especially relevant for text-conditional diffusion models, which underpin modern text-to-image synthesis systems.

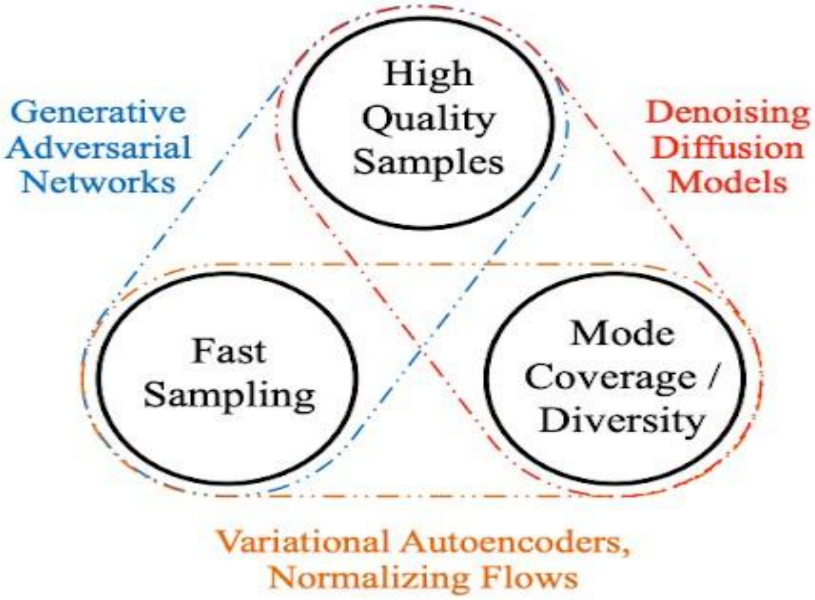

High Quality Samples
Generative Adversarial Networks
Denoising Diffusion Models
Fast Sampling
Mode Coverage / Diversity
Variational Autoencoders, Normalizing Flows

Fig 1.1: Types of Diffusion Models

1.2 Development before DDPM:

The emergence of Denoising Diffusion Probabilistic Models (DDPM) can be traced back to two key earlier contributions: Score-Based Generative Models (SGM) in 2019 and Diffusion Probabilistic Models (DPM), which first appeared in 2015. To gain a clear understanding of DDPM, it is crucial to first examine the principles behind DPM and the role of Stochastic Differential Equations (SDEs) in these models.Experimental results across several datasets have validated the effectiveness of DPM in capturing and modeling intricate data distributions. While DPM forms the foundational theory behindDDPM, the latter refines and enhances the original approach, resulting in more efficient implementations and better performance.The DPM framework was the first to model data distributions by reversing a Markov diffusion process, which transitions data into a simpler, more tractable Adistribution. In particular, the forward process in DPM gradually deforms the complex data distribution into a more basic one, such as a Gaussian distribution. The model then learns how to reverse this transformation, recovering the data from the noise added during the forward process.

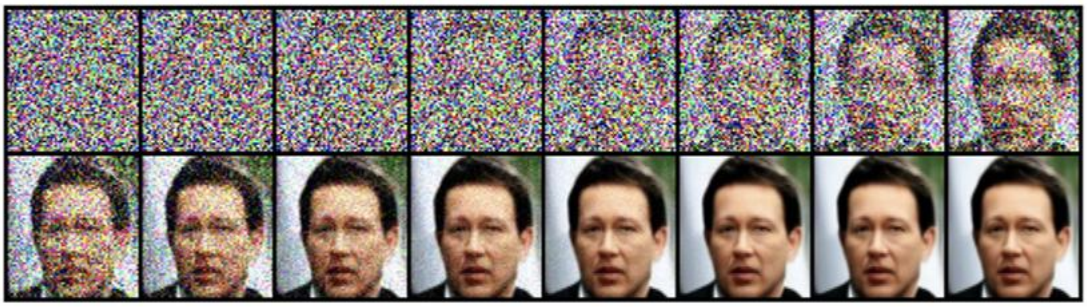


Fig 1.2: Denoising Diffusion Probabilistic Models

1.3 Source-based Generative model(SGM):

Various techniques have been explored to enhance Score-Based Generative Models (SGM). In SGM, the data is perturbed by adding Gaussian noise with varying intensities. The model uses the gradient of the log-probability density as the score function and generates samples by progressively reducing the noise levels. During training, SGM learns by estimating these score functions for distributions of noisy data.Although the motivations behind SGM and DDPM differ, both models share a similar optimization goal during training. In fact, under certain parameterizations, DDPM can be considered equivalent to SGM in terms of training dynamics. An improved version of SGM has also been proposed, which is designed to scale to high-resolution images.

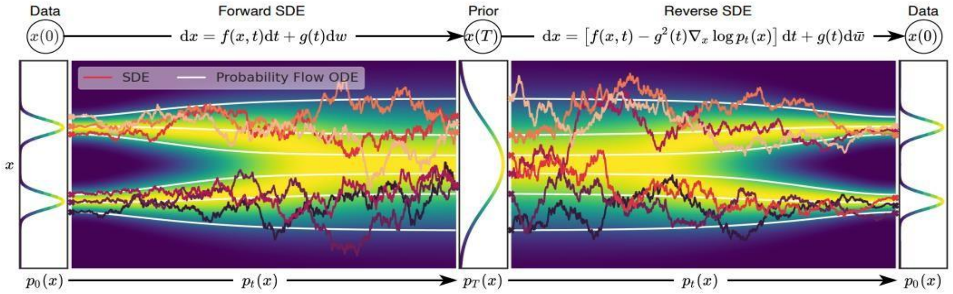

Data
Forward SDE
Prior
Reverse SDE
Data
$x(0)$
$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w$
$x(T)$
$\mathrm{d}x = [f(x,t) - g^2(t)\nabla_x \log p_t(x)]\,\mathrm{d}t + g(t)\mathrm{d}\bar{w}$
$x(0)$
SDE
Probability Flow ODE
x
$p_0(x)$
$p_t(x)$
$p_T(x)$
$p_t(x)$
$p_0(x)$

Fig 1.3: Score-based Generative model

Guidance in diffusion-based image synthesis:

The use of labels has been shown to significantly enhance the quality of image synthesis. Early research on Generative Adversarial Networks (GANs) demonstrated that incorporating class labels can improve the generated images. For instance, Conditional GAN introduced the concept of feeding class labels as an additional input layer to the model, contributing to better image generation. Further advancements, such as the work in, utilized class-conditional normalization statistics to refine the image synthesis process. Additionally, AC-GAN explicitly incorporated an auxiliary classifier loss, enabling more control over the generated images.

In these GAN-based methods, labels play a dual role they either serve as conditional inputs or guide the image generation through an auxiliary classifier. Building on these successful strategies, a recent study applied both class-conditional normalization and auxiliary classifiers to diffusion models, extending the benefits of labels to this new paradigm. To clarify the ways in which label information is utilized, we adopt the definitions for two types of models in diffusion-based image generation

Conditional Diffusion Model: This model incorporates additional information, such as class labels or text descriptions, as inputs during training, enhancing its ability to generate images based on specific conditions

Guided Diffusion Model: this approach, the model is trained with class-specific gradients, which, for example, can be derived from an auxiliary classifier. These gradients are then used during the sampling process to guide the generation towards desired outputs

CHAPTER-2

LITERATURE SURVEY

“Generative Adversarial Text-to-Image Synthesis - Scott Reed, Zeyanep Akata , Xinchen Yan.”

In this they introduces a method for generating realistic images from text descriptions using generative adversarial network (GANs). GAN-based architecture with a combination of text and image encoders.

“Text to photo realistic image synthesis with stacked Generative Adversarial Networks - Han Zhang, Tao Xu , Hongsjeng.”

In this it represents the proposes a stacked GAN architecture for generating high resolution images from text descriptions.

“Image Generation from Text with Transformers - Patrick Esser, Robin R. Selvaraju Marc Areli. ”

In this Transformer-based architecture for image synthesis. The paper explores textto-image generation using transformer-based models.

“Creating Images from Text - Alec Radford, Karthik Narasimhan, Tim Salimans.”

In this A large-scale transformer model with a focus on text-image generation. They introduces DALL_E, a model capable of generating diverse images from textual descriptions.

“DF-GAN: Deep Fusion Generative Adversarial Networks - Zhangetal.”

In this DF-GAN introduces a one-stage GAN architecture that simplifies the text-to-image generation process. By eliminating the need for complex multi-stage designs, DF-GAN employs a Deep Fusion Block (DFBlock) to effectively merge text and image features, resulting in high-quality, semantically consistent images. This approach addresses limitations in previous models related to training complexity and image-text alignment.

“AttnGAN :Attentional Generative Adversarial Networks - Xuetal.”

In this AttnGAN incorporates attention mechanisms to refine image generation by focusing on specific words in the text description during different stages of image synthesis. This allows the model to generate more detailed and semantically relevant images, improving upon earlier GAN-based approaches that lacked fine-grained control over image features.

“VQGAN-CLIP: Vector Quantized GAN with CLIP Guidance - Esseretal.”

In this VQGAN-CLIP combines the generative capabilities of Vector Quantized GANs with the semantic understanding of CLIP (Contrastive Language–Image Pretraining). This integration enables the generation of high-fidelity images that are semantically aligned with the input text, offering a balance between image quality and textual relevance.

"DALL·E 2: Hierarchical Text-Conditional Image Generation - Rameshetal."

In this DALL·E 2 builds upon its predecessor by introducing a two-step process: first, generating a CLIP image embedding from a text prompt, and then using a decoder to produce the final image. This hierarchical approach enhances image quality and diversity, allowing for more accurate and creative interpretations of textual descriptions.

"Imagen: Photorealistic Text-to-Image Diffusion Models - Sahariaetal."

In this Imagen leverages large pre-trained language models to better understand text prompts and employs diffusion models to generate high-resolution, photorealistic images. The model demonstrates state-of-the-art performance on various benchmarks, highlighting the effectiveness of combining advanced language understanding with powerful image generation techniques.

"Make-A-Scene: Scene-Based Text-to-Image Generation - Gafnietal. "

In this Make-A-Scene introduces a novel approach that allows users to provide both textual descriptions and scene layouts as inputs. This method offers greater control over the composition and spatial arrangement of generated images, enabling more precise and user-guided image synthesis.

"GLIDE: Guided Language to Image Diffusion for Generation and Editing - Nicholetal."

In this GLIDE explores the use of diffusion models for both image generation and editing tasks, guided by textual prompts. The model exhibits strong performance in generating images from text and allows for nuanced edits to existing images based on new textual inputs, showcasing versatility in text-to-image applications.

"CogView: Chinese-Oriented Text-to-Image Generation - Dingetal. "

In this CogView focuses on generating i.mages from Chinese text prompts, addressing the challenges of multilingual text-to-image synthesis. By adapting Transformer-based architectures to handle Chinese language inputs, CogView expands the applicability of text-to-image models to non-English languages.

CHAPTER-3

METHODOLOGY

T. Q. Chen et al. explored the stable diffusion process, describing it as a type of stochastic process that models the random movement of particles within a medium. This natural phenomenon, commonly observed in heat conduction, fluid dynamics, and diffusion, provides valuable theoretical underpinnings for its application in image generation. Their insights highlight the potential of stabilized diffusion techniques for producing coherent and high-quality visuals.

Daniel Grathwohl et al. conducted a comparative analysis of Generative Adversarial Networks (GANs) and diffusion models in the context of image synthesis. Their findings demonstrate that diffusion models exhibit superior stability and consistency, outperforming GANs in terms of image quality and reliability, which underscores the growing preference for diffusion-based approaches in modern generative tasks.

In a related contribution, Alaluf et al. introduced the HyperStyle method, which uses hypernetworks to perform StyleGAN inversion for real-world image editing. This approach inspired the idea of enabling fine-grained modifications to generated images by altering details while maintaining identity and coherence an essential feature for practical applications like suspect sketching or portrait editing.

Brock et al. advanced the field with their work on large-scale GAN training, achieving state-of-the-art performance in natural image synthesis through the training of high-capacity models. Their success demonstrated the importance of scale and data diversity in producing photorealistic outputs. Similarly, the SWAPGAN technique introduced a strategy of progressive refinement beginning with low-resolution image synthesis and gradually increasing detail which provided a useful framework for high-fidelity generation.

Karas et al. proposed a training paradigm known as progressive growing, where both the generator and discriminator of GANs are expanded in size during training. This method improves image quality, training stability, and diversity by focusing computational effort where it's most needed over time.

Another foundational contribution comes from Kingma et al., who developed the Variational Autoencoder (VAE) a neural generative model capable of learning compact latent representations of high-dimensional data. By maximizing a lower bound on the data log-likelihood, VAEs support tasks such as data generation, compression, and unsupervised learning, making them crucial to latent space modeling in modern architectures.

Radford et al. presented Deep Convolutional GANs (DCGANs), emphasizing unsupervised representation learning. Their architecture demonstrated how meaningful visual features could be learned without labeled data, enabling the model to generate high-quality, semantically coherent synthetic images aligned with the input prompt. Collectively, these works offer critical insights into the development of robust, flexible, and efficient generative systems. They not only illuminate the strengths and limitations of various models such as the training instability often observed in GANs or the heavy data requirements but also informed the design of our system. Specifically, the ability to modify specific image details for refinement and feedback-driven updates stems from the innovations discussed in these studies, which collectively guide the ongoing evolution of image synthesis methodologies.

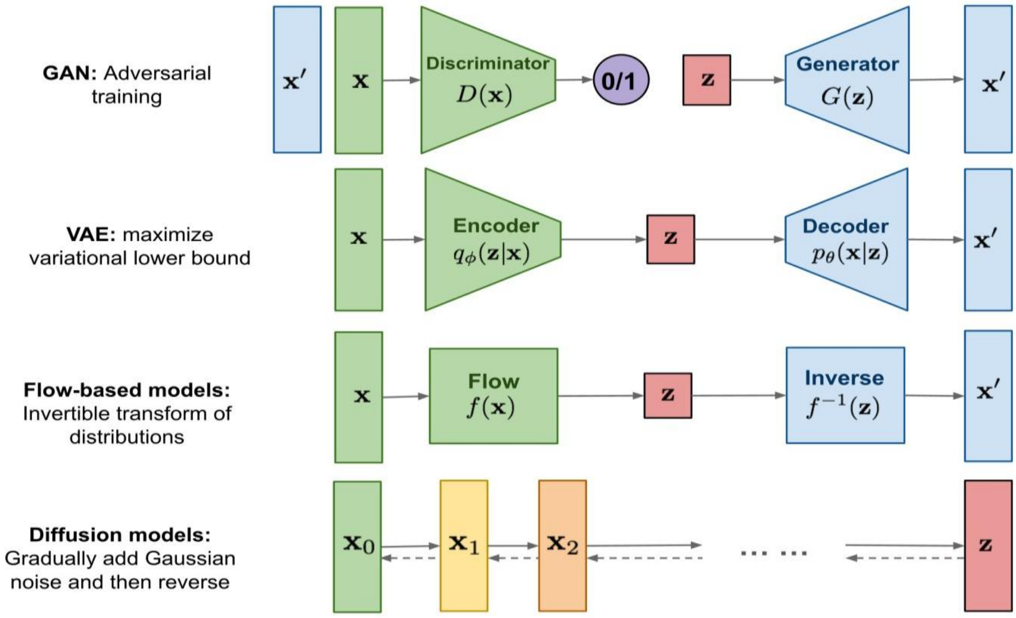

GAN: Adversarial training
$\mathbf{x}'$
$\mathbf{x}$
Discriminator $D(\mathbf{x})$
0/1
$\mathbf{z}$
Generator $G(\mathbf{z})$
$\mathbf{x}'$
VAE: maximize variational lower bound
$\mathbf{x}$
Encoder $q_\phi(\mathbf{z}|\mathbf{x})$
$\mathbf{z}$
Decoder $p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$
Flow-based models: Invertible transform of distributions
$\mathbf{x}$
Flow $f(\mathbf{x})$
$\mathbf{z}$
Inverse $f^{-1}(\mathbf{z})$
$\mathbf{x}'$
Diffusion models: Gradually add Gaussian noise and then reverse
$\mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{x}_2$
... ...
$\mathbf{z}$

Fig 3.1: Generative Models

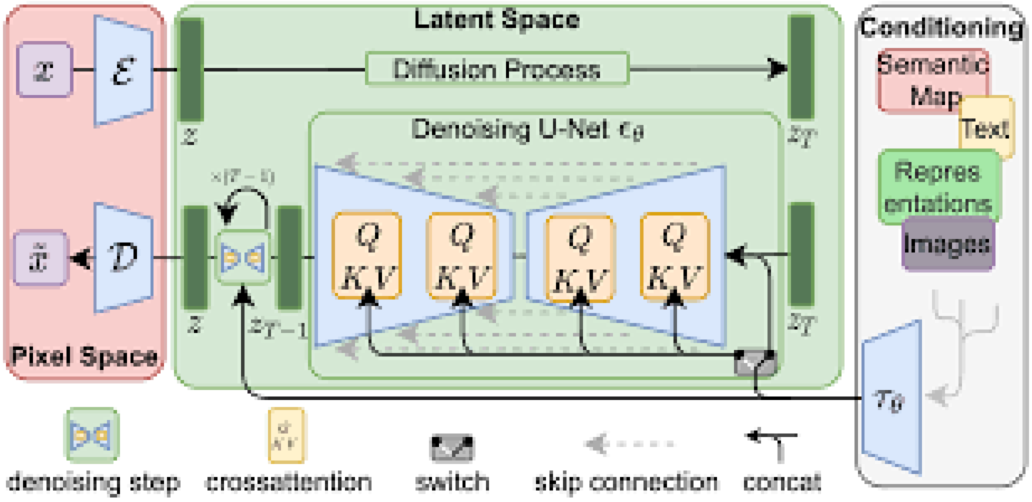


Fig 3.2: Diffusion Model

Generative models for synthesis of images Generative modelling faces unique difficulties due to the high dimensionality of pictures. Networks of Generative Adversaries(GAN) provide effective examining of high goal pictures with adequate perceptual quality, however, they are trying to tune and have trouble capturing the complete data distribution. In contrast, likelihood-based techniques priorities accurate density prediction, making optimisation more compliant. High resolution pictures may be synthesised well using variational autoencoders (VAE) and flow-based models , but sample quality is not on par with GANs. A sequential sampling procedure and computationally costly designs limit the resolution of the pictures that autoregressive models (ARM) can produce, despite their good performance in density estimation. Maximum-likelihood training uses a disproportionate amount of capacity to model the scarcely perceptible, high-frequency features that are present in pixel-based representations of pictures, leading to lengthy training timeframes. Many two-stage techniques model a compressed latent image space with ARMs rather than raw pixels in order to scale to higher resolutions. Recent advancements in sample quality and density estimation have been made by Diffusion Probabilistic Models (DM). When these models' neurological underpinnings are implemented as UNets, they naturally suit the inductive biases of image-like data, which gives rise to their generative capacity. When are weighted goal is used for training, the best synthesis quality is often attained. In this present circumstance, the dissemination Model is comparable to a lossy blower and considers the compromise of pressure proficiency for picture quality. Nevertheless, the disadvantage of evaluating and improving these models in pixel space is low inference speed and very large training costs. Although improved sampling techniques and hierarchical sampling can help to some extent with the former, at approaches. Preparing on high-goal picture information generally expects to ascertain costly angles. We address the two downsides with our proposed LDMs, which work on a compacted inactive space of lower dimensionality. Image Synthesis in Two Stages Several studies have focused on using a two step strategy to combine the benefits of various techniques into more effective and performant models in order to reduce the drawbacks of individual generative approaches. Auto-regressive models are used by VQ-VAEs to develop an expressive prior over a discretized latent space. By studying a combined distribution across discretized picture and text representations, extend this method to text-to-image creation. In contrast to VQ-VAEs, VQGANs scale autoregressive transformers to bigger pictures using a first stage with an adversarial and perceptual goal. The overall performance of such techniques is, nonetheless, obliged by the enormous pressure rates important for viable ARM preparing, which adds billions of teachable boundaries, and less pressure comes at the punishment of a high computational expense. Because to our proposed LDMs' convolutional backbone, which scales more easily to larger dimensions latent spaces, such compromises are avoided. So, we are allowed to choose the amount of pressure that best intercedes between learning major areas of strength for a phase, without giving the generative dissemination model an excess of perceptual pressure, while yet guaranteeing high constancy reconstructions. While there are strategies to become familiar with an encoding/unravelling model couple with a result based earlier either mutually or independently.

CHAPTER-4

EXISTING METHOD AND PROPOSED METHOD

4.1 EXISTING METHOD:

Architecture:

The system architecture relies on Generative Adversarial Networks (GANs) and Transformer-based models, combining the strengths of adversarial training with attention mechanisms. GANs, composed of a generator and a discriminator in a competitive setup, are adept at producing visually plausible images. Transformer-based components, often inspired by models like CLIP or DALL·E, are used to interpret textual inputs and inform the image synthesis process. However, the hybrid architecture increases the complexity of both model training and inference, which may limit real-time deployment in resource-constrained environments.

Image Quality:

While capable of generating diverse and creative visuals, the image output often suffers from lower fidelity, particularly in facial features, hands, and fine textures. Common issues include asymmetry, missing elements, or distorted proportions. These imperfections frequently require manual post-processing using tools like Photoshop or image inpainting networks to achieve production-quality results.

Customization:

Customization capabilities are constrained. Fine-tuning specific features such as precise facial expressions, clothing details, or environmental elements is challenging without access to prompt engineering tricks or internal model weights. Users must rely on vague prompt manipulation, which often yields inconsistent results. Lack of controllable parameters or modular editing tools limits use in professional design workflows or iterative concept development.

Text-Image Alignment:

The model exhibits difficulty interpreting and translating complex, nuanced, or abstract descriptions into visually accurate representations. For example, prompts containing idiomatic expressions, emotional subtext, or detailed multi-step actions may lead to visual outputs that are misaligned with user intent. This limitation stems from the model's reliance on large-scale pretraining data, which may not fully capture rare or context-sensitive linguistic patterns.

Narrative Complexity:

Handling multi-character scenes or scenarios with layered narratives such as a dynamic story involving multiple interacting individuals, evolving actions, or distinct settings—is particularly problematic. The model tends to simplify scenes or misrepresent relationships and actions between subjects. Spatial coherence, interaction realism, and continuity across characters are common weak points, making such systems suboptimal for storyboarding or comic generation without significant human intervention.

Computational Load:

The combined use of GANs and Transformers results in substantial computational demands, especially during training. The adversarial setup necessitates frequent discriminator-generator updates, while Transformers consume significant memory for attention mechanisms. As a result, generation times are slower, especially for high-resolution outputs or complex prompts. This makes the model less suitable for interactive applications or deployment on standard consumer hardware.

Training Stability:

GANs are notorious for training instabilities, including mode collapse (where the generator produces limited variety), vanishing gradients, and oscillatory behavior between generator and discriminator. These issues complicate training and often lead to inconsistent outputs, especially across different training sessions or datasets. Despite numerous advancements like Wasserstein loss and progressive growing, full stability remains elusive in many GAN-based systems.

Feedback Integration:

Incorporating user feedback—such as refining an image based on critiques or adjusting individual elements without starting from scratch—is not natively supported. This is primarily because current architectures do not support iterative image editing with preservation of context. Most refinements require retraining or complete regeneration from modified prompts, which limits usability in iterative design processes or user-in-the-loop workflows.

Resolution Control:

Generating images at very high resolutions (e.g., 4K or above) often introduces artifacts, such as blurring, aliasing, or patch inconsistencies. Upscaling methods, like super-resolution GANs (e.g., ESRGAN), can alleviate this to some extent, but they are often detached from the core model and may not preserve fine semantic details. Native high-resolution generation remains a challenge due to memory constraints and limited dataset examples during training at such scales.

4.2 PROPOSED METHOD

The proposed system leverages the power of Stable Diffusion, a state-of-the-art generative AI model, to generate facial images of suspects based on textual descriptions provided by witnesses. By integrating this advanced model with real-time feedback mechanisms, the system addresses critical challenges in traditional suspect identification methods, particularly those relying on manual sketches by forensic artists or older GAN-based architectures.

Key Innovations and Advantages

- **Real-Time Witness Feedback**: The system allows for iterative refinement of generated images through direct input from witnesses, improving the likeness and accuracy of suspect portrayals.

- **Automation of Manual Tasks**: By reducing the dependency on forensic sketch artists for every case, the system frees up valuable human resources for more complex investigative work.

- **Enhanced Accuracy**: Stable Diffusion produces high-fidelity images that are perceptually realistic and aligned with textual descriptions, offering better results than GAN-based systems, which often suffer from facial distortions and lack of consistency.

System Architecture

The system is composed of four major modules:

1. **Input Processing Module**

   Parses witness narratives or textual descriptions, extracting key facial attributes and converting them into a structured prompt suitable for the diffusion model.

2. **Diffusion Model Engine**

   Utilizes a latent diffusion model that transforms the text prompt into a compressed latent space, then gradually denoises it to produce a realistic image. This model benefits from separating compressive encoding from the generative process, drastically improving computational efficiency

3. **Image Generation Module**

   Renders high-resolution facial images from the latent representations. Advanced super-resolution techniques are optionally applied to enhance the output for investigative use.

4. **User Interface Module**

   Provides an intuitive, interactive interface for law enforcement officers and witnesses to review, refine, and approve suspect images in real time.

Performance and Efficiency Improvements

- The latent space transformation reduces the computational cost associated with pixel-level function evaluations, which are common in conventional diffusion models.
- By using an autoencoder to map the image space to a lower-dimensional perceptual space, the system maintains visual fidelity while achieving significant gains in speed and energy efficiency.
- Compared to existing GAN-Transformer-based systems, our model:
  1. Requires fewer resources
  2. Produces more coherent and accurate images
  3. Offers better alignment with narrative text inputs

Psychological and Cognitive Foundations

Humans naturally visualize stories as mental images while listening or reading. This system aligns with this cognitive mechanism by translating descriptive narratives into images, helping witnesses more effectively communicate what they saw. As a result, it supports:

- Cognitive recall reinforcement
- Higher confidence in identification
- Reduced ambiguity in suspect image generation

Challenges Addressed

- **Existing Limitations**:
  - GANs and transformers often produce artifacts or inconsistent features
  - Require large GPU memory and are slow in high-resolution rendering

- Lack iterative feedback integration

➢ **Our Solution**:

- Stable Diffusion addresses training stability and output fidelity
- The autoencoding approach lowers computational demand
- Real-time witness interaction loop enhances usability and result precision

CHAPTER-5

BLOCK DIAGRAM

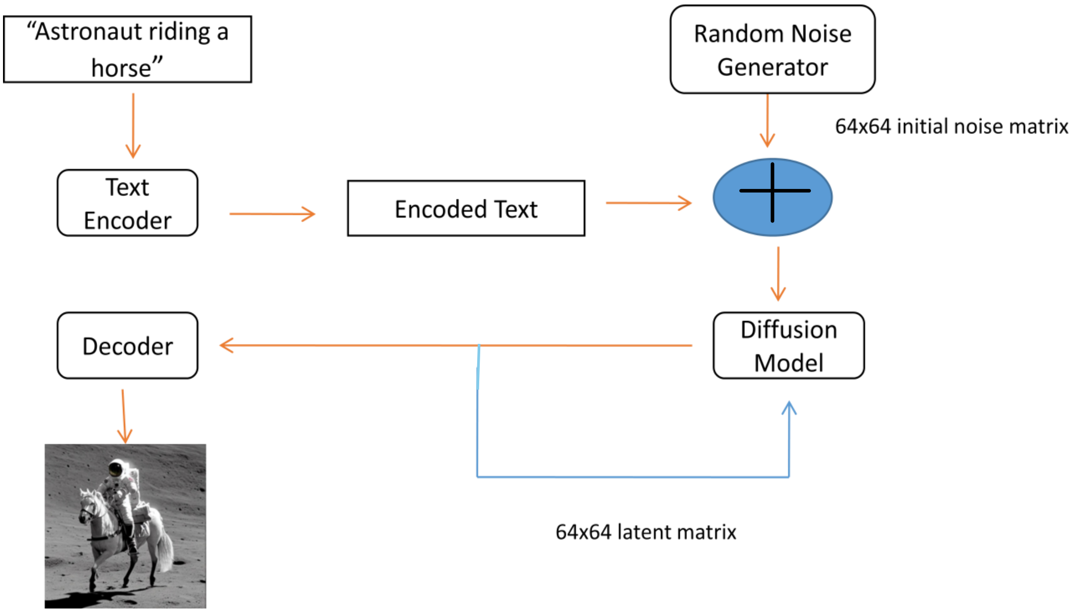


Fig 5.1: Block Diagram

This diagram illustrates the workflow of a text-to-image generation system using a diffusion model, specifically resembling the architecture used in models like Stable Diffusion. Here's a step-by-step breakdown of each component and the process:

1. Text Input

- **"Astronaut riding a horse":**

  This is the text prompt provided by the user. It serves as the input description for the image to be generated.

2. Text Encoder

- The text is processed by a Text Encoder (e.g., CLIP or a Transformer-based encoder), which converts the natural language input into a numerical representation called Encoded Text.
- This encoding captures the semantic meaning of the prompt.

3. Random Noise Generator

- A 64x64 matrix of random noise is created. This serves as the starting point for the image generation process.
- The model gradually transforms this noise into a meaningful image through denoising steps.

4. Conditioning the Diffusion Process

- The Encoded Text and the initial noise matrix are combined and fed into the Diffusion Model.
- The encoded text conditions the diffusion process so that the final image reflects the input description.

5. Diffusion Model

- This is the core generative engine. It takes the noisy image and progressively refines it through a series of denoising steps, guided by the text encoding.
- The result is a 64x64 latent matrix, which is a compressed representation of the image in latent space.

6. Decoder

- The latent matrix is then passed to a Decoder (typically a Variational Autoencoder (VAE) decoder), which transforms it into a final image in pixel space.
- The output image visually represents the original text prompt.

7. Final Output

- The final image (in this case, an astronaut riding a horse) is generated and presented to the user.

Summary

This architecture shows how a text-to-image diffusion model works:

- It starts with a text prompt
- Encodes the text
- Combines it with random noise
- Uses a diffusion model to denoise and generate a latent image representation
- And finally decodes that latent image into a realistic visual.

CHAPTER-6

IMPLEMENTATION

In this model we will have to import the dependencies that are required for the creation of this system and such will be stable diffusion library, Tensorflow, pytorch, OpenCV, fastAPI, tensorflow.js, react, Nodemon, node.js. The image generating module uses the Stable Diffusion AI model to create facial images of suspects based on input text descriptions. It pre-processes the input text to remove unwanted characters and generate a latent vector representation of the facial features that will be created. The decoder module then maps this vector back to the image space to produce a realistic image of the suspect. The generated image can be refined based on real-time feedback. This module offers a reliable and efficient way for law enforcement agencies to identify suspects. The Diffusion module is a key component of the Stable Diffusion AI image generation process. It gradually smooths the image over multiple iterations by applying noise and blending it into the image using a diffusion process. The amount of noise is reduced during each iteration, resulting in a smoother image with important features and patterns. The Diffusion module can handle large datasets and is useful for various tasks, such as image generation, manipulation, classification, and object detection. The encoder module takes the input image and encodes it into a lower-dimensional latent space, where the image features can be more easily manipulated and processed. This process is done to reduce the complexity of the input image, making it easier to work with and generate new images. The decoder module takes the latent representation and decodes it back into the original image space, producing an output image that is a reconstruction of the original input image. This module is responsible for generating high-quality images that are similar to the original input image. The prior module imposes a prior distribution on the latent space, encouraging the encoded features to be smooth and structured. This helps to ensure that the generated images are coherent and consistent with the input image. The noise module adds Gaussian noise to the input image at each iteration, helping to regularize the diffusion process and prevent overfitting.

This process helps to ensure that the generated images are diverse and not just reproductions of the original image. The training module trains the entire model end-toend, optimizing the model parameters to minimize the difference between the reconstructed output image and the original input image. This process involves updating the model weights and biases to improve the accuracy and quality of the generated images. The training process is repeated multiple times until the model achieves the desired level of performance The User Interface module provides a GUI for inputting text descriptions and displaying generated facial images. It takes user input and communicates with image generating module to generate a facial image of the suspect. The generated facial image is displayed in the user interface, allowing for real-time feedback to improve the accuracy of the generated image. This module is essential for efficient and accurate input of text descriptions and providing a visual representation of generated facial images, improving efficiency and accuracy of the suspect identification process.

Improving model architectures:

On the choice of guidance: Beyond the classifier-free guidance, some works have also explored cross modal guidance with CLIP. Specifically, GLIDE finds that CLIP-guidance underperforms the classifier-free variant of guidance. By contrast, another work UPainting points out that lacking of a large-scale transformer language model makes these models with CLIP guidance difficult to encode text prompts and generate complex scenes with details. By combing large language model and cross-modal matching models, UPainting significantly improves the sample fidelity and image-text alignment of generated images. The general image synthesis capability enables UPainting to generate images in both simple and complex scenes. On the choice of denoiser. By default, DM during inference repeats the denoising process on the same denoiser model, which makes sense for an unconditional image synthesis since the goal is only to get a high-fidelity image.

In the task of text-to-image synthesis, the generated image is also required to align with the text, which implies that the denoiser model has to make a trade-off between these two goals. Specifically, two recent works point out a phenomenon: the early sampling stage strongly relies on the text prompt for the goal of aligning with the caption, but the later stage focuses on improving image quality while almost ignoring the text guidance.

Therefore, they abort the practice of sharing model parameters during the denoising process and propose to adopt multiple denoiser models which are specialized for different generation stages. Specifically, ERNIE-ViLG 2.0 also mitigates the problem of objectattribute by the guidance of a text parser and object detector, improving the fine-grained semantic control.

Sketch for spatial control:

Despite their unprecedented high image fidelity and caption similarity, most text-to-image DMs like Imagen and DALL-E2 do not provide fine-grained control of spatial layout. To this end, SpaText introduces spatio-textual (ST) representation which can be included to finetune a SOTA DM by adapting its decoder. Specifically, the new encoder conditions both local ST and existing global text. Therefore, the core of SpaText lies in ST where the diffusion prior in trained separately to convert the image embeddings in CLIP to its text embeddings. During training, the ST is generated directly by using the CLIP image encoder taking the segmented image object as input. A concurrent work proposes to realize fine-grained local control through a simple sketch image. Core to their approach is a Latent Guidance Predictor (LGP)that is a pixelwise MLP mapping the latent feature of a noisy image to that of its corresponding sketch input. After being trained (see for more training details), the LGP can be deployed to the pretrain text-to-image DM without the need for fine-tuning.

Textual inversion for concept control:

Pioneering works on text-to-image generation rely on natural language to describe the content and styles of generated images. However, there are cases when the text cannot exactly describe the desired semantics by users, e.g., generating a new subject. In order to synthesize novel scenes with certain concepts or subjects, introduces several reference images with the desired concepts, then inverts the reference images to the textual descriptions. Specifically, inverts the shared concept in a couple of reference images into the text (embedding) space, ”pseudo-words”. The generated ”pseudowords” can be used for personalized generation. Dream Booth adopts a similar technique and mainly differs by fine-tuning (instead of freezing) the pretrained DM model for preserving key visual features from the subject identity.

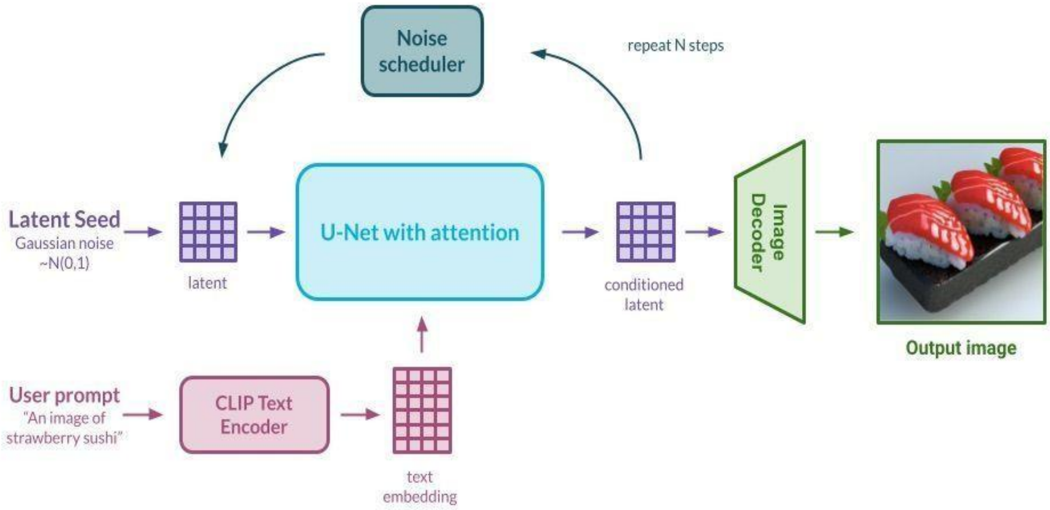


Fig 6 : Prompt is generated into Image

6.1 Text-guided creative generation :

Visual art generation:

Artistic painting is an interesting and imaginative area that benefits from the success of generative models. Despite the progress of GAN-based painting , they suffer from the unstable training and model collapse problem brought by GAN. Recently, multiple works present impressive painting images based on diffusion models, investigating improved prompts and different scenes. Multimodal guided artwork diffusion (MGAD) refines the generative process of diffusion model with multimodal guidance (text and image) and achieves excellent results regarding both the diversity and quality of generated digital artworks. In order to maintain the global content of the input image, DiffStyler propose a controllable dual diffusion model with learnable noise in the diffusion process of content image. During inference, explicit content and abstract aesthetics can both be learned with two diffusion models. Experimental results show that DiffStyler achieve excellent results on both quantitative metrics and manual evaluation.

6.2 Video generation and story visualization:

Text-to-video:

Since video is just a sequence of images, a natural application of text-to-image is to make a video conditioned on the text input. Conceptually, text-to-video DM lies in the intersection between text-to-image DM and video DM. Regarding text-to-video DM, there are two pioneering works: Make-A-Video adapting a pretrained text-to-image DM to text-to video and Video Imagen extending an existing video DM method to text-to-video. Make-A-Video generates high-quality videos by including temporal information in the pretrained text-to-image models, and trains spatial super-resolution models as well as frame interpolation models to enhance the visual quality. With pretrained text-to-image models and unsupervised learning on video data, Make-A-Video successfully accelerates the training of a text-to-video model without the need of paired text-video data. By contrast, Imagen Video is a text-to-video system composed of cascaded video diffusion models. As for the model design, Imagen Video points out that some recent findings (e.g., frozen encoder text conditioning) in text-to-image can transfer to video generation, and findings for video diffusion models(e.g., v-prediction parameterization) also provides insights for general diffusion models. Text-to-story generation.

6.3 3D generation:

3D object generation:

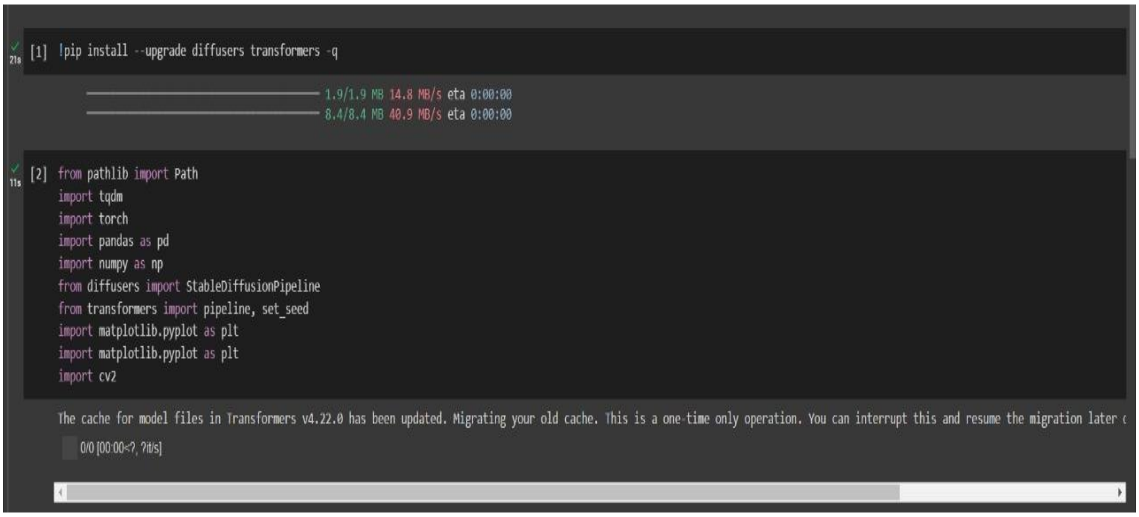
The generation of 3D objects is evidently much more sophisticated than their 2D counterpart, i.e., 2D image synthesis task. Deep Fusion is the first work that successfully applies diffusion models to 3D object synthesis. Inspired by Dream Fields which applies 2D image-text models (i.e., CLIP) for 3D synthesis, DeepFusion trains a randomly initialized NeRF with the distillation of a pretrained 2D diffusion model (i.e., Imagen). However, according to Magic3D, the low-resolution image supervision and extremely slow optimization of NeRF result in low-quality generation and long processing time of DeepFusion. For higher resolution results, Magic3D proposes a coarse-to fine optimization approach with coarse representation as initialization as the first step, and optimizing mesh representations with high-resolution diffusion priors. Magic3D also accelerates the generation process with a sparse 3D hash grid structure. 3DDesigner focuses on another topic of 3D object generation, consistency, which indicates the crossview correspondence. With low-resolution results from NeRF-based condition module as the prior, a two stream asynchronous diffusion module further enhances the consistency, and achieves 360-degree consistent results.

Models of Latent Diffusion:

Dissemination Models are probabilistic models made to progressively denoise a regularly dispersed variable to become familiar with an information circulation p(x), which is comparable to learning the contrary course of a decent Markov Chain of length T. The best models for picture blend utilize a reweighted variety of the variational lower limit on p(x), which is like denoising score-coordinating. Idle Portrayal Generative Demonstrating We presently approach a successful, low-layered dormant space in which high-recurrence, imperceptible data are preoccupied away utilizing our prepared perceptual pressure models made out of E and D. This space is more appropriate for probability based generative models than the high layered pixel space since it permits them to I focus on the pivotal, significant parts of the info and (ii) train in an extensively lower layered, computationally undeniably more productive environment. In contrast to previous work that used autoregressive, attention-based transformer models in a highly compressed, discrete latent space, we may benefit from our model's image-specific inductive biases. Diffusion models, like other generative models, may theoretically represent conditional distributions of the kind p. (z|y). This opens the door to regulating using inputs like text, semantic maps, or other image-to-image translation tasks in the synthesis process and may be accomplished with a conditional denoising autoencoder ϵΘ (zt, t, y). Nevertheless, integrating the generating potential of DMs with conditionings other than class names or obscured varieties of the information picture is at this point a neglected field of concentrate with regards to picture combination.

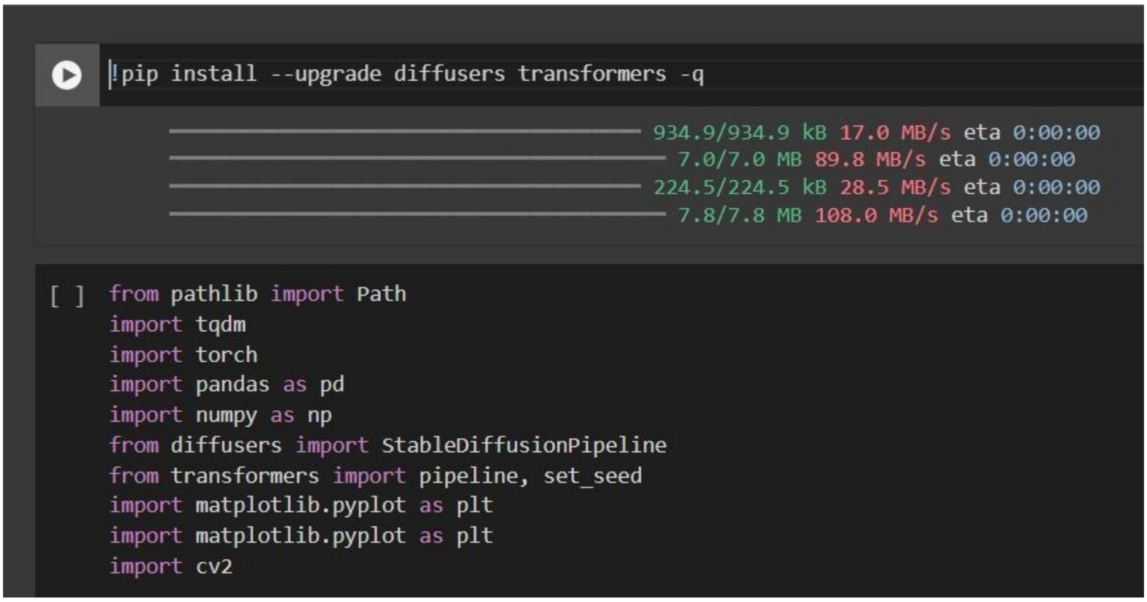
CHAPTER-7

IMPLEMENTED SCREENSHOTS



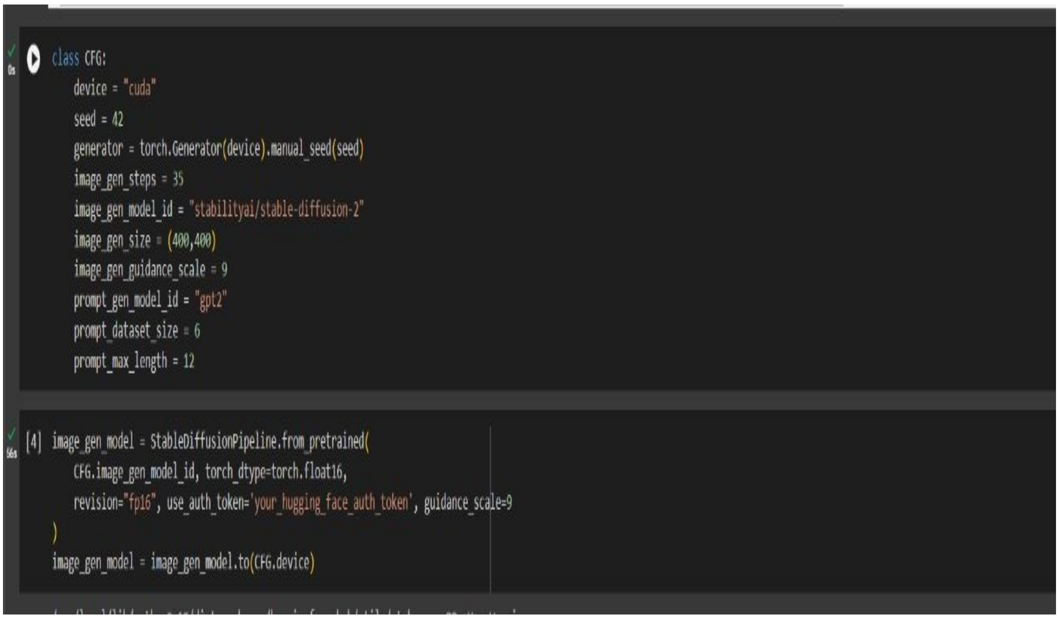
```
21s [1] !pip install --upgrade diffusers transformers -q

        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.9/1.9 MB 14.8 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 8.4/8.4 MB 40.9 MB/s eta 0:00:00

11s [2] 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

        The cache for model files in Transformers v4.22.0 has been updated. Migrating your old cache. This is a one-time only operation. You can interrupt this and resume the migration later
         0/0 [00:00<?, ?it/s]
```

Fig 7.1: Pip Install



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

        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 934.9/934.9 kB 17.0 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.0/7.0 MB 89.8 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 224.5/224.5 kB 28.5 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.8/7.8 MB 108.0 MB/s eta 0:00:00

[ ] 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
```

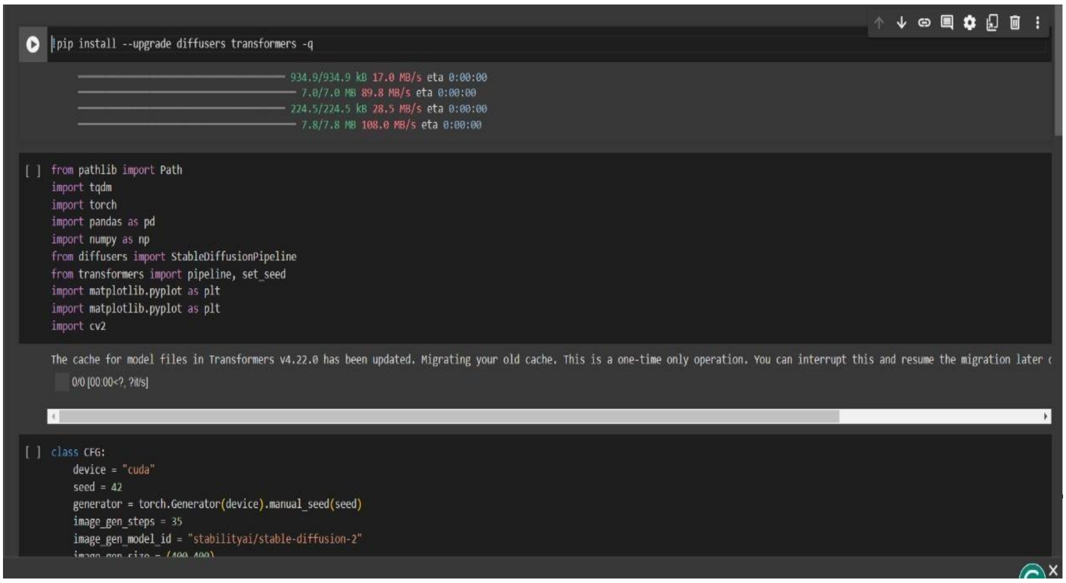
Fig 7.2: Installing pip



```
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
```

```
[4] image_gen_model = StableDiffusionPipeline.from_pretrained(
        CFG.image_gen_model_id, torch_dtype=torch.float16,
        revision="fp16", use_auth_token='your_hugging_face_auth_token', guidance_scale=9
    )
    image_gen_model = image_gen_model.to(CFG.device)
```

Fig7.3: Install libraries

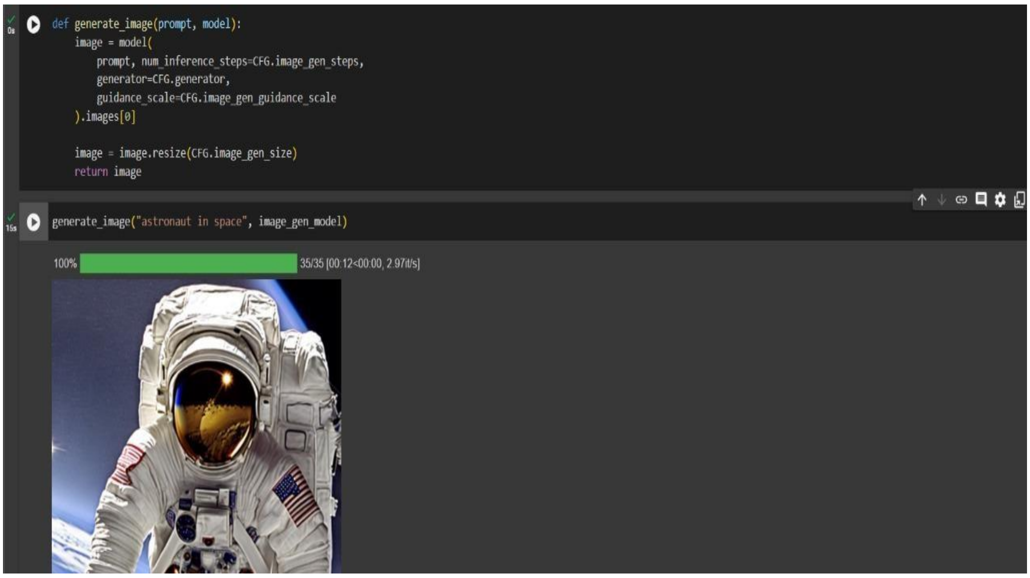


```
!pip install --upgrade diffusers transformers -q

     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 934.9/934.9 kB 17.0 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.0/7.0 MB 89.8 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 224.5/224.5 kB 28.5 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.8/7.8 MB 108.0 MB/s eta 0:00:00
```

```
[ ] 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

    The cache for model files in Transformers v4.22.0 has been updated. Migrating your old cache. This is a one-time only operation. You can interrupt this and resume the migration later c
    0/0 [00:00<?, ?it/s]
```

```
[ ] 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)
```

Fig 7.4: Import Libraries



```
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
```

```
generate_image("astronaut in space", image_gen_model)
```

```
100% 35/35 [00:12<00:00, 2.97it/s]
```

Fig 7.5: Loaded image with respective prompt.

CHAPTER-8

ADVANTAGES

1. Improved Image Quality

Improved image quality refers to enhancements in the visual fidelity and detail of the generated images. This includes:

- **Sharper details**: Clearer textures, defined edges, and reduced blurriness.
- **Better color accuracy**: More realistic and appealing color palettes.
- **Higher realism or stylistic accuracy**: Depending on the intended style (photorealistic, cartoon, etc.), the model better matches the target aesthetic.
- **Fewer artifacts**: Reduction in common generative errors like distorted faces, hands, or backgrounds.

This improvement results from better model training techniques, larger and more diverse training datasets, and architectural advancements in generative models.

2. Consistency

Consistency in image generation implies that:

- **Stylistic coherence** is maintained across multiple images or frames (important for animations or multi-image outputs).
- **Structural integrity** is preserved objects look correct and follow logical anatomy or perspective.
- **Prompt adherence** is strong, meaning the generated image accurately reflects the input description.

For example, if a user requests an image of a person with specific features or in a certain setting, consistency ensures those features are correctly represented every time.

3. Low Cost to Run Training and Interface

This refers to the efficiency and affordability of both:

- **Model training**: Using optimized architectures, such as transformer-based models with efficient attention mechanisms or diffusion models that require fewer training steps, reduces computational resource needs.
- **Interface operations**: Running the model for inference (image generation) becomes more accessible and affordable, making it viable for wider use in

consumer applications, mobile devices, or cloud-based platforms with budget constraints.

These cost reductions are crucial for scalability and making AI image generation accessible to more users and businesses.

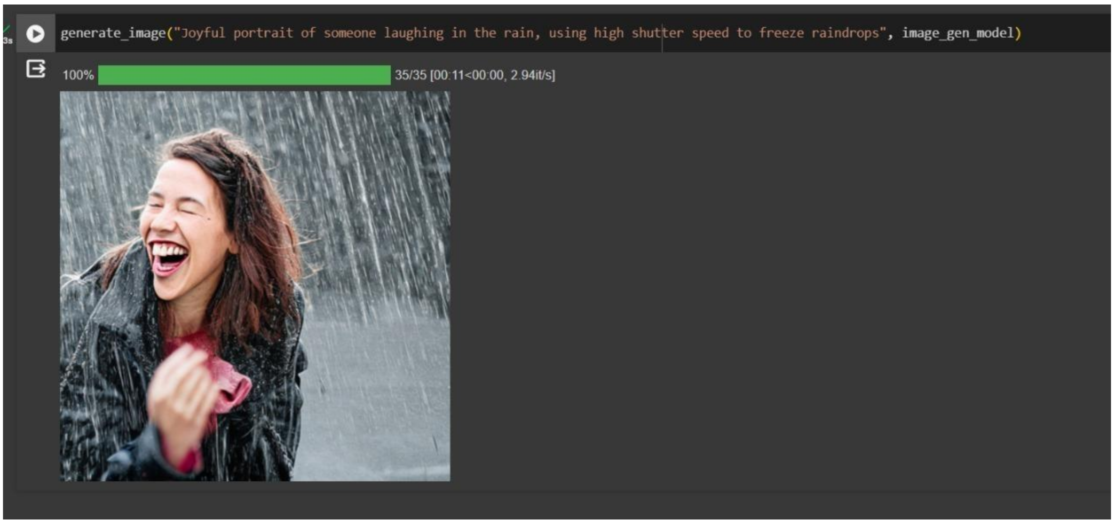
4. High Resolution Image Generator

This feature enables the generation of images with a high number of pixels (e.g., 4K, 8K) while preserving quality. Benefits include:

- **Detailed prints**: Useful for posters, product renders, or professional media.
- **Zooming without loss**: Users can zoom into the image without pixelation or degradation.
- **Better post-processing**: High-resolution images allow for more flexibility in editing, cropping, and using in different formats.
- High-resolution generation often involves multi-step diffusion processes or upscaling mechanisms built into the generation pipeline (like super-resolution models)

CHAPTER -9

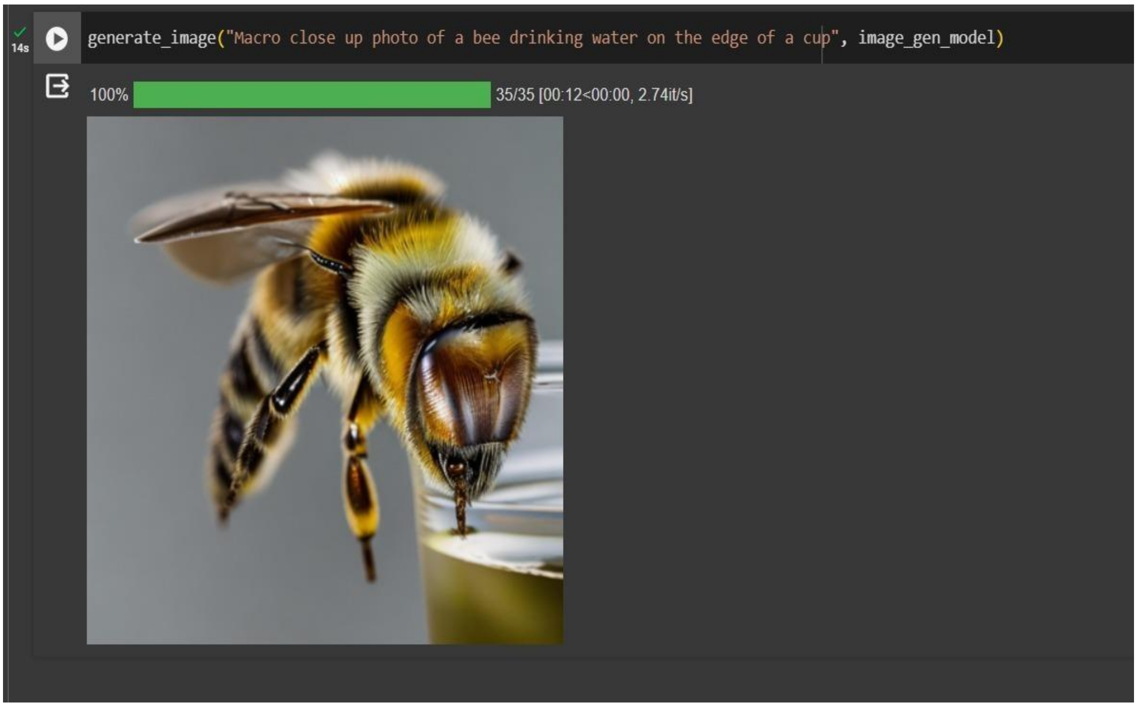
RESULTS



```
generate_image("Joyful portrait of someone laughing in the rain, using high shutter speed to freeze raindrops", image_gen_model)
```

```
100% 35/35 [00:11<00:00, 2.94it/s]
```

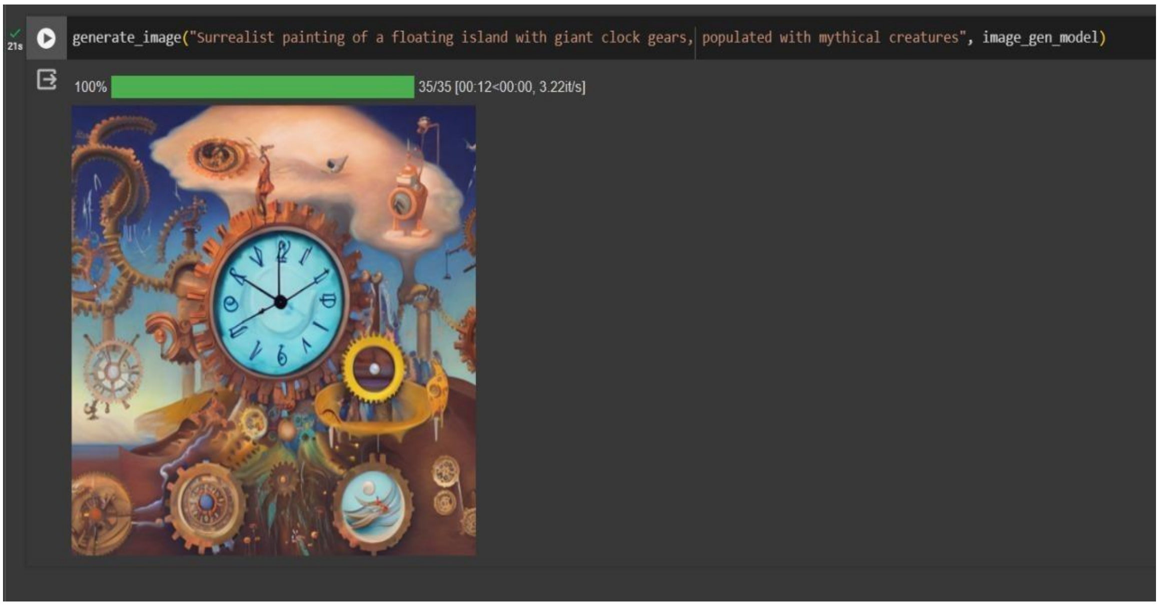
Fig 9.1: “Joyful portrait of someone laughing in the rain, using high shutter speed to freeze raindrops”.



```
generate_image("Macro close up photo of a bee drinking water on the edge of a cup", image_gen_model)
```

```
100% 35/35 [00:12<00:00, 2.74it/s]
```

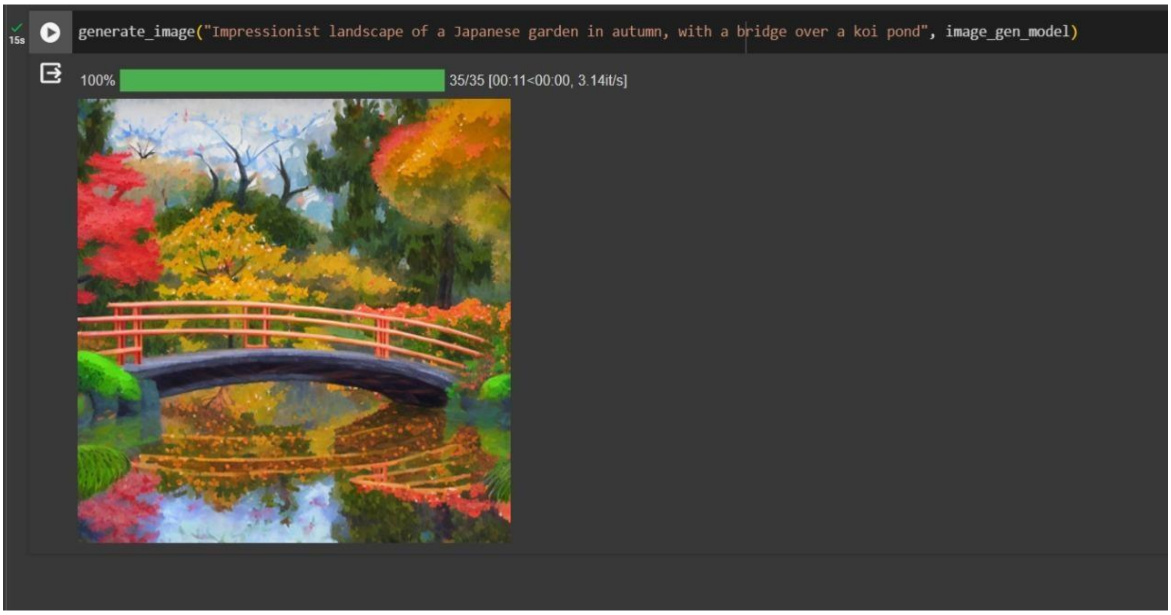
Fig 9.2: “Macro close up photo of a bee drinking water on the edge of a cup”.



```
generate_image("Surrealist painting of a floating island with giant clock gears, populated with mythical creatures", image_gen_model)
```

```
100% 35/35 [00:12<00:00, 3.22it/s]
```

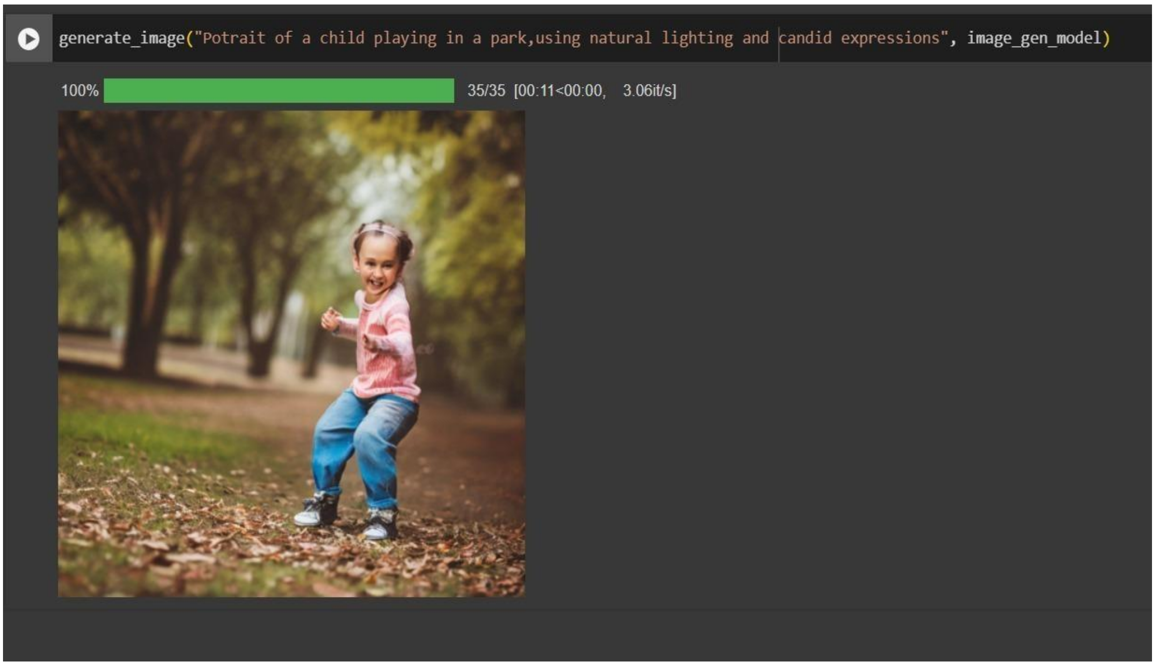
Fig 9.3: "Surrealist painting of a floating Island with giant clock gears, populated with mythical creatures".



```
generate_image("Impressionist landscape of a Japanese garden in autumn, with a bridge over a koi pond", image_gen_model)
```

```
100% 35/35 [00:11<00:00, 3.14it/s]
```

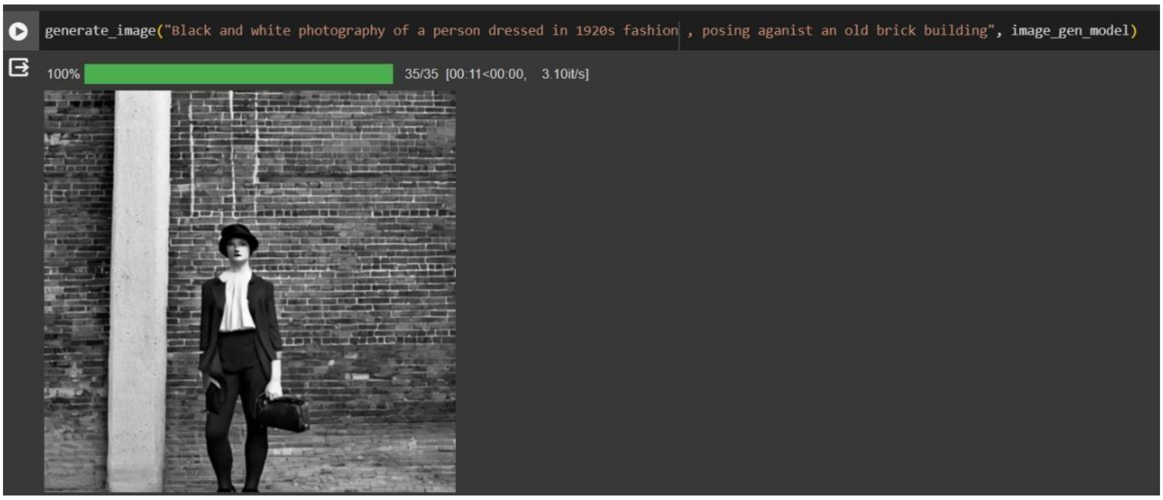
Fig 9.4: "Impressionist landscape of a Japanese garden in autumn, with a bridge over a koi pond".



```
generate_image("Potrait of a child playing in a park,using natural lighting and candid expressions", image_gen_model)
```

```
100% 35/35 [00:11<00:00,  3.06it/s]
```

Fig 9.5: “Portrait of a child playing in a park, using natural lighting and candid expression”.



```
generate_image("Black and white photography of a person dressed in 1920s fashion , posing aganist an old brick building", image_gen_model)
```

```
100% 35/35 [00:11<00:00,  3.10it/s]
```

Fig 9.6: “Black and white photography of a person dressed in 1920’s fashion, posing against an old brick building”.

CHAPTER-10

CONCLUSION& FUTURE ENHANCEMENT

CONCLUSION

The proposed system is a modern, cloud-based solution developed to generate high-quality facial images from textual descriptions, with a focus on forensic and investigative applications. It integrates three key components: a user-friendly interface for inputting witness descriptions, an advanced Stable Diffusion AI model for generating images, and a secure database for storing both input data and generated outputs. Designed with accessibility, scalability, and security in mind, the system offers a practical and efficient alternative to traditional suspect sketching methods.

At the heart of the system is the Stable Diffusion model, a state-of-the-art generative AI framework known for producing visually realistic and detailed images from text prompts. Unlike manual sketches that rely on an artist's interpretation, this AI model ensures objectivity and consistency, reducing human bias and increasing the accuracy of suspect representation. Evaluation reports and testing results show that the model can produce images closely resembling real individuals, enhancing its value in real-world investigations.

The intuitive interface allows law enforcement personnel, forensic artists, or even witnesses to easily input physical descriptions, which the AI processes in seconds to produce corresponding facial composites. All data, including textual input and generated images, is stored in a secure, encrypted cloud database that complies with modern privacy and data protection standards.

Cloud infrastructure also enables remote access, seamless updates, and scalable deployment across different law enforcement agencies. This flexibility, combined with the system's high accuracy and user-friendly design, makes it a powerful tool for improving the speed and reliability of suspect identification.

In summary, the system effectively combines artificial intelligence and cloud technology to deliver a practical, secure, and scalable solution for facial image generation from text, significantly advancing investigative capabilities in modern law enforcement.

FUTURE ENHANCEMENT

To further enhance the effectiveness of the proposed system, several strategic improvements can be pursued. One key direction involves incorporating advanced image processing and analysis techniques to increase the accuracy and realism of the generated facial images. Technologies such as facial symmetry correction, contextual background synthesis, and machine learning-based feature refinement can help produce more lifelike and representative visuals that align closely with witness descriptions. Another significant opportunity lies in broadening the system's scope beyond suspect identification. For example, it can be adapted to support missing persons investigations by generating age-progressed images or reconstructing appearances from limited data. Additionally, the system could assist in reconstructing crime scenes or visualizing partial evidence based on narrative descriptions, offering valuable visual aids for investigators.

Improving the system’s performance to achieve real-time or near real-time image generation is also critical, particularly for time-sensitive investigations where rapid decision-making is essential. This would require optimizations in both the AI model’s inference pipeline and the underlying computational infrastructure to minimize latency and resource consumption. Furthermore, integrating the system with existing law enforcement tools such as facial recognition software, biometric databases, and criminal record systems would create a unified investigative framework. Such integration would streamline workflows, enhance data cross-referencing capabilities, and improve overall identification accuracy. Collectively, these advancements would transform the proposed system into a versatile and powerful asset for a wide range of law enforcement and public safety applications.

REFERENCES:

1) Ankit Yadav1, Dinesh Kumar Vishwakarma2, Recent Developments in Generative Adversarial Networks: A Review (Workshop Paper),2020.

2) Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, “StackGAN: Text to Photorealistic Image Synthesis with Stacked Generative Adversarial Networks" in Rutgers University and Lehigh University August 2017.

3) Tao Xu, Pengchuan Zhang, Qiuyuan Huang Han Zhang,Xiaolei Huang, Xiaodong (2018).AttnGAN: Finegrained text to Image Generation with Attention Generative Adversarial Networks.

4) Patrick Esser, Robin R.Selvaraju, Marc’Areli (2021). Image Generation from text with Transformers.

5) Generative Models and the Stabilizing Diffusion" by T. Q. Chen, et al., (ICLR 2021) introduced a novel method for image generation using the stable diffusion process.

6) Diffusion Models Beat GANs on Image Synthesis" by D. Grathwohl, et al., (ICLR 2021) compared the performance of GANs and diffusion models for image synthesis.

7) A.I., S.: Stable diffusion public release, https://stability.ai/blog stable-diffusionpublic-release.

8) Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

9) Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. Proceedings of the International Conference on Learning Representations (ICLR).

10) Wu, J., Zhang, C., Xue, T., Freeman, B., & Tenenbaum, J. (2019). Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. Advances in Neural Information Processing Systems (NeurIPS), 82-92.

11) R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation

and maximization," in Proceedings of the International Conference on Learning Representations (ICLR), 2019.

12) S. D. McDermott and M. W. Mahoney, "Adaptive importance sampling for diffusionbased generative models," arXiv preprint arXiv:2102.02760, 2021.

13) A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in Proceedings of the International Conference on Learning Representations (ICLR), 2018.

14) Y. Zhang, Y. Zhang, J. Wen, and Y. Li, "Self-supervised learning for image synthesis and manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

15) C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

16) A. Brock, J. Donahue, and K. Simonyan, "Understanding and improving interpolation in autoencoders via an adversarial regularizer," in Proceedings of the International Conference on Learning Representations (ICLR), 2019.

17) A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Generative and discriminative voxel modeling with convolutional neural networks," arXiv preprint arXiv:1608.04236, 2016.