Implementation of Diffusion Model in Realistic Face Generation

By

Abidur Rahman
20101441
Faiyaz Al-Mamoon
20101499
Mohammad Nazmus Saquib
20101480
Farhan Bin Bahar
20101519
Mukarrama Tun Fabia
20101572

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

School of Data and Sciences
Department of Computer Science and Engineering
Brac University
January, 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Abidur Rahman

20101441

Faiyaz Al-Mamoon

20101499

Mohammad Nazmus Saquib

20101480

Farhan Bin Bahar

20101519

Mukarrama Tun Fabia

20101572

Approval

The thesis titled “Implementation of Diffusion Model in Realistic Face Generation” submitted by

1. Abidur Rahman (20101441)
2. Faiyaz Al-Mamoon (20101499)
3. Mohammad Nazmus Saquib (20101480)
4. Farhan Bin Bahar (20101519)
5. Mukarrama Tun Fabia (20101572)

Of Fall, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on January 09, 2024.

Examining Committee:

Supervisor:
(Member)

____

Dr. Md. Khalilur Rhaman

Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

____

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

____

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Ethics Statement

Our research work on Implementation of Diffusion Model in Realistic Face Generation prioritizes ethical considerations by ensuring privacy compliance, labeling biases and striving for uprightness in the development and usage of AI technologies. We strictly commit to abide by the legal obligations of our home country, university, workplace and local community.

Abstract

Realistic Face Generation has emerged as a compelling area of research in the field of Artificial Intelligence and it has gained massive attention through the people as it has significant usage in many sectors. Its application ranges from Facial Recognition Systems to Deepfake Detection. Our research focuses on adapting and fine-tuning the Diffusion Model specifically to the domain of Face Generation. We propose a Novel architecture that combines the Diffusion process with a Latent Space Model, enabling precise control over the generated faces' attributes such as age, gender, facial features etc. Furthermore, we are using a dataset having diverse facial images to train and evaluate the performance of our model. The work that has been done in this paper includes, using Diffusion Models in areas related to Realistic Face Generation with a goal of improving current infrastructures, as well as establishing new ones. Our research not only explores the theoretical underpinnings of Diffusion Models, but also extends its inquiry into their practical applications, encompassing mathematical computations, formulas, principles, and cutting-edge execution techniques tailored to the domain of Realistic Face Generation. This research looks into numerous sectors where the applications of this Realistic Face Generation technique can make the overall process more efficient. First of all, our work starts by analyzing the existing scholarly articles and papers on various types of Diffusion Models, their usage, and contribution to the world of Computer Science. We are examining some Diffusion Models with such details that inaugurates our theoretical base of the research. Furthermore, as we are trying to implement these models to generate faces and recognize faces, we are also addressing the influence of various parameters such as noise level, time constraints, quality of the images and many more. Moreover, we are taking the testing and learning phases into Deep Monitoring so that this nobel work should overcome the practical challenges related with the usage of Diffusion Models for Realistic Face Generation without breaking any ethical code or breaching data privacy. Additionally, we plan to use Deep Learning concepts for further face detection and recognition and find more use cases. In conclusion, our research advances the field of Face Generation by introducing and implementing the Diffusion Model as a powerful framework for generating highly realistic and diverse human faces and the results of our experiments highlights the models potential for applications in this area.

Keywords: Diffusion model, Artificial intelligence, face generation, Stable diffusion, deep learning, neural network, natural language processing

Dedication

We dedicate our thesis to our family and friends, who have helped us throughout the whole process and provided us with the inspiration we needed to successfully finish our paper.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, we would like to convey our profound gratitude to our supervisor Dr. Md. Khalilur Rhaman sir for his invaluable guidance, kind support and insightful feedback throughout the whole research process. We would also like to thank Sayantan Roy Arko, our research assistant, for providing valuable insights regarding the technical outlooks of authoring the article. Special thanks to our families, friends and friends of friends for their participation in our research survey which contributed to its completion.

Table of Contents

Declaration i

Approval i

Ethics Statement iv

Abstract v

Dedication vi

Acknowledgment vii

Table of Contents viii

List of Figures x

List of Tables xi

Nomenclature xii

1 Introduction 1

1.1 Background and Research Motivation 1
1.2 Research Objectives 2
1.3 Problem Statement 3

2 Literature Review 4

2.1 Text to Image 4
2.2 CLIP Transformation 5
2.3 Image Encoder 6
2.4 Text Encoder 6
2.5 Imagen 6
2.6 Imagic 6
2.7 Diffusion Model 7
2.8 Model Comparision 9
2.9 Finetuning Methods 9

3 Dataset 11

3.1 Dataset 11
3.2 Data Collection 12
3.3 Previous Attempts at Dataset Preprocessing 13

3.4 Dataset Preprocessing 14
3.5 Challenges Faced 16

4 Model **18**
4.1 Model Selection 21
4.2 Model Descriptions 22
4.2.1 SD Version1.5 23
4.2.2 SD Version2.0 23
4.2.3 SDXL 25
4.2.4 DALL-E-3 26

5 Model Training **28**
5.1 Model Training Environment 28
5.2 Fine Tuning Methods 29
5.2.1 DreamBooth 29
5.2.2 LoRA 31
5.2.3 Textual Inversion 33
5.2.4 Hypernetwork 33
5.3 DreamBooth Training 34
5.4 Challenges in Training 35

6 Results and Analysis **36**
6.1 Model Training and Generation 36
6.1.1 Model Selection & Parameter Tuning 36
6.2 Image Generation and Evaluation 37
6.2.1 Models under Evaluation 37
6.2.2 Quantitative Metrics 37
6.2.3 Quantitative Results 39
6.2.4 Qualitative Analysis 39

7 Conclusion **42**
7.1 Conclusion 42
7.2 Future Scope 42

References **46**

Bibliography **46**

List of Figures

1.1 General idea behind diffusion model [16] 2

2.1 CLIP Architecture 5

3.1 FFHQ (Flickr-Faces-HQ) Dataset 12
3.2 Defining Facial Attributes based on Landmarks 13
3.3 Work Flow 15
3.4 Facial Landmark Templates 16

4.1 U-Net Architecture 20
4.2 Workflow of Stable Diffusion 20
4.3 Official first samples by Stable Diffusion 21
4.4 Outputs from facial feature prompts by SD1.5 22
4.5 Stable Diffusion 2.0 23
4.6 SD1.5 VS SD2.0 24
4.7 SD 1.5 vs SD 2.0 24
4.8 SD 1.5 vs SD 2.0 25
4.9 Official examples of SDXL 26
4.10 Real Face Generation using DALL-E-3 27

5.1 Simplified Workflow of DreamBooth 30
5.2 Detailed Workflow of DreamBooth 30
5.3 Artistic self-portraits using LoRA 32

6.1 Sample Survey 40
6.2 Votes for each model on each image 40
6.3 Total votes for each model 41

List of Tables

5.1 Training Environment 28
5.2 DreamBooth VS FastDreamBooth 31
5.3 Basic comparison between Fine-Tuning Models 33

6.1 Score of different models 39

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$CLIP$	Contrastive Language-Image Pretraining
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
KID	Kernel Inception Distance
LDM	Latent Diffusion Model
$LoRA$	Low-Rank Adaptation
$LPIP$	Learned Perceptual Image Patch Similarity
MMD	Maximum Mean Discrepancy
SD	Stable Diffusion
$SDV1.5$	Stable Diffusion Version 1.5
$SSIM$	Structural Similarity Index Measure
UI	User Interface
VAE	Variational AutoEncoder
$ViLT$	Visual-Language Transformer
ViT	Vision Transformer

Chapter 1

Introduction

1.1 Background and Research Motivation

With the significant progress of computer graphics, AI and image processing techniques, the area of face illustration has observed remarkable advancement in recent years. Furthermore, the integration of cutting edge methods, including those developed by OpenAI and other contributors, has played a crucial role in elevating the refinement and capabilities within the sector of face illustration. In hindsight, face generation or illustration totally relied on 3D modeling, manual drawing and texture mapping. While these ways have been successful to a certain extent, they are time-consuming and usually need a significant amount of artistic skills. Moreover, they may struggle to accurately capture the difficulty and volatility of human facial features. Now, there are various apps and softwares like Amazon Recognition, Betaface, BioID etc. for face recognition, but very few applications or models that generate surrealistic faces. From a use case perspective, a realistic and accurate text to face model can help a lot of the presently in place infrastructures. One of such cases is finding a missing person. Through implementation of our research, the whole process can be made a lot more efficient. An individual can file a report and based on the description one provides, we will be able to get a realistic and accurate representation of the missing person. We can then run that data into facial recognition software and install these into CCTV feeds and locate them successfully. Thus making the whole process more efficient and a lot less time consuming.

To utilize the best of technology and solve the problem, researchers first brought Variational Autoencoders (VAEs) to generate images from textual descriptions. [22] But as the VAEs had some restrictions to generate detailed high quality images, GANs were introduced to the world in 2014. [22] While GANs were performing much better than the first approved idea of VAEs, GANs also had some limitations regarding model collapse, vanishing gradient issue, unstable training and many more. At last overcoming most of the problems that GANs faced, in 2020, Diffusion models were introduced. Since then they have gained noteworthy popularity in the community. The basic and general idea behind the diffusion model is that it is a generative model that takes input and gradually appends noise through it in many steps which is also known as the forward process. After that, a neural network works to regain the given data by removing the noise from the data which is also known as reverse diffusion process.[9] At the very beginning of the usage of diffusion models, these models were only used as image modification tools. As it evolved day by day, now

Midjourney, stable diffusion, Dall-E, Imagen are some of the well recognized models used to illustrate face.

From that researches built another version named stable diffusion model on the idea of diffusion models and initiated much more advancement to make the model more efficient and stable than others. Studies suggest that a stable diffusion model gives better outcomes than other models. Yet the quality of the result is much worse than models trained on. The main objective of this project is to create software using diffusion model applications and deep learning to illustrate face from available data and details provided.

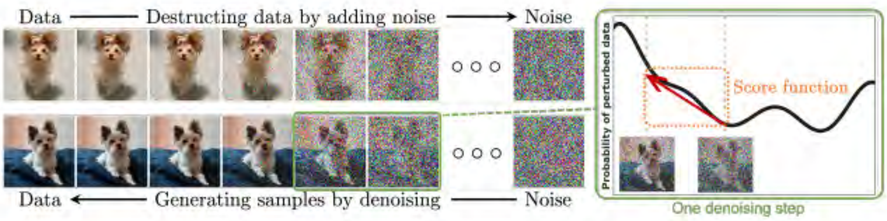

Data
Destructing data by adding noise
Noise
Data
Generating samples by denoising
Noise
Probability of perturbed data
Score function
One denoising step

Figure 1.1: General idea behind diffusion model [16]

1.2 Research Objectives

This research aims to investigate the usage of various diffusion models and to train stable diffusion with custom annotated dataset to create more accurate and realistic facial representations than the existing artistic generative models. The objectives of this project are:

- Study and review the existing models, their applications and contribution in the image processing and computer graphics field.
- Achieve an exclusive understanding of various models including dall-e-2, min-Imagen, stable diffusion and many more to analyze and capture the work sequence of the existing models.
- Implementing a framework that combines facial image datasets and diffusion based algorithms
- Overcome practical challenges and train an effective model that does not breach any kind of data privacy
- Gain performance to shorten the time-consuming part.
- Provide insights and recommendations for the responsible and efficient deployment of diffusion models in face illustration, considering their potential applications and limitations.

1.3 Problem Statement

The Traditional means of illustrating facial features accurately in a realistic way is a rather difficult and inconsistent field. Currently, we have a few models that help us get an output, but they mainly provide surrealistic or conceptual images instead of photo realistic, which results in inaccurate results. That is why we have decided to implement an intricate system that uses Stable Diffusion for illustrating the real face of a person. However, the bumps that we faced in our research are:

- **Insufficient Datasets:** One of the problems that we came across during our research process was the scarcity of relevant datasets that we need for training and testing diffusion models for illustration of facial features. Lack of a high quality dataset resulted in hindrance in development and ultimately produced highly inaccurate and inefficient outputs. So, a relevant, rich and diverse dataset that contains specific information like descriptive facial features in various lighting positions, poses, expressions, age, ethnicity, region based facial identifiers etc. is essential for smoothly conducting this type of research work.

- **Model Selection:** Another one of the core problems that we encountered is selection and designing of a diffusion model that is specifically catered towards our purpose of Face Illustration. There are several existing diffusion models like Denoising Diffusion Probabilistic Models (DDPMs) [4], Generative Adversarial Networks (GANs), Imagen, along with their many variants. All of them have their own strengths as well as weaknesses. Choosing one that specifically caters to our needs of producing realistic and expressive faces was a real challenge in and of itself.

- **Training, Efficiency and Optimization:** Training Diffusion model is a very computationally demanding task and requires a huge amount of resources as well as time. Training the model requires a lot of computational power and then producing a semi-accurate output is also demanding. Boosting the training efficiency and optimizing the whole procedure step by step, while also maintaining the quality of the generated illustration was a challenge that had to be mitigated. A few techniques we used in order to get over this is, regularization, data augmentation, parallel computing as well as various advanced optimization algorithms.

- **Ethical Considerations:** The ethical analysis of AI research, particularly face illustration, is essential. Addressing potential problems including bias, fairness, privacy, and the proper use of created graphics is crucial. Researching ethical guidelines, proposing strategies to reduce biases, and talking about the societal effects of using diffusion models for face illustration were important aspects of this research.

These are some of the significant research problems that guided us by laying a groundwork, so that we can properly make progress on the working methodology, experimentation and analysis. Addressing these challenges and coming up with workarounds for each of them ultimately resulted in advancement in this field which led us to a stronger, more efficient and more robust model in this domain.

Chapter 2

Literature Review

Recently text to image generation models have developed rapidly and gained a lot of popularity in both the scientific community and the common population. Our goal is to create such a model to generate realistic photos of human faces from text inputs of human likenesses to help in cases of human identification, criminal investigation etc.

2.1 Text to Image

Our research began with the paper “Zero-Shot Text-to-Image Generation” [7] that introduces us to DALL-E. It uses GANs to generate high-quality images from text inputs without any paired text-image training data and paves the way for generating diverse and creative images. To improve the quality of the images generated from textual descriptions, we found another paper titled ”Guiding Text-to-Image Diffusion Model Towards Grounded Generation”.[15] It proposes to use clip guided diffusion that utilizes CLIP mode to produce semantically consistent images. This method outperforms any existing model in grounding scores and quality. Another study “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding” [6] introduces Imagen, a text-to-image diffusion model that combines massive transformer language models like T5 with diffusion models to generate photorealistic images with a high level of language understanding. The study demonstrates that expanding the language model in Imagen has a greater impact on sample fidelity and picture-text alignment than expanding the image diffusion model. The research also introduces DrawBench, a new benchmark for text-to-image models, and compares Imagen with other advanced methods, such as VQ-GAN+CLIP, Latent Diffusion Models, GLIDE, and DALL-E 2. It concludes that Imagen outperforms other models in side-by-side comparisons for sample quality and image-text alignment, according to human raters. It shows its success in categories like generating actual texts in images which popular diffusion models have failed to do so as well as creating better semantics and accepting larger text prompts from users.

2.2 CLIP Transformation

CLIP is the abbreviation for Contrast Language Image Pre Training which is a multi-modal big model[21]. It illustrates both text and image data into a common latent space using dual encoder architecture. OpenAI is a research laboratory which actively works on cutting-edge deep learning techniques which tends computers to solve complex problems on its own by thinking like a human being [24]. OpenAI uses NLP to train billions of compilations of large datasets, complex parameters and contributes to better understanding resulting in better performance. GPT-3, one of the languages by OpenAI, is competent enough to recognize handwriting, convert text to image, recognize face and many more. CLIP is developed by OpenAI and pre-trained with a huge number of datasets. It is trained to understand the joint representation of both text and image. These two joint encoders are called Vision Transformer and Transformer Based Language Model[23]. Vision Transformer is used to encode images and Transformer Based Language Model for text. From the given textual description it chooses images with similar description and gradually moves out those that don't match. Because of joint representation training with vast datasets it is able to identify objects based on the textual descriptions that weren't included in pre-training as the shared embedding space permits CLIP to resemble the relation between text and image. This zero shot learning ability increased the efficiency and effectiveness of our work, which in turn, significantly streamlined our workflow and made our tasks more manageable. The working process of CLIP is mainly divided into three steps. Firstly, pre-training distinctively then creating categories of dataset and lastly prognosis of zero-shot.

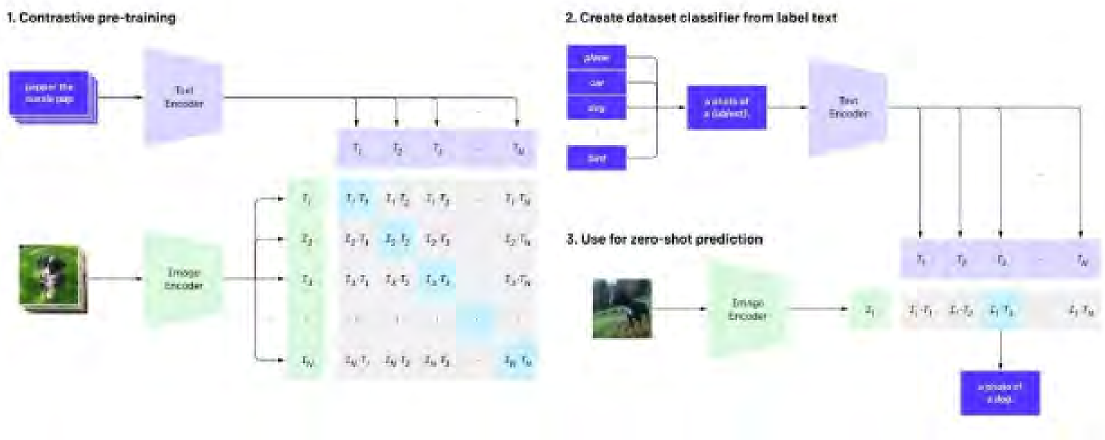

1. Contrastive pre-training
Text Encoder
Image Encoder
2. Create dataset classifier from label text
plane
car
dog
bird
Text Encoder
3. Use for zero-shot prediction
Image Encoder

Figure 2.1: CLIP Architecture
[23]

2.3 Image Encoder

Image encoder obtains essential characteristics from the image provided. After taking the image as input the encoder creates a vector illustration of high dimension. Usually architectures like CNN are used to extract features from images. It reduces the height and width of the image but the depth of the image grows simultaneously.

2.4 Text Encoder

Text encoder encrypts the denotation of the given textual description of the image. It takes the caption of the text as input and generates vector illustrations of high dimension. Usually transformer based architectures like BERT are used to operate the chain of texts.

2.5 Imagen

The study from the paper "Photo-realistic Text-to-Image Diffusion Models with Deep Language Understanding"[11] introduces Imagen, which is a text-to-image diffusion model that combines diffusion models with transformer language models. This helps to improve the quality of images generated from diffusion models. Imagen scores a FID score of 7.27 on COCO dataset without prior training that also corresponds with real human reviews of the similarity of generated images with the COCO dataset. According to human raters, Imagen performs better than other models in terms of sample quality and alignment. It also excels at enhancing semantics, and accepting longer text prompts. To further improve likelihoods on image density estimation benchmarks, the variational lower bound (VLB) expression can be reduced by effective noise schedule optimization and model improvements as cited in the paper "Variational Diffusion Models" [5]. In doing so the models outperform autoregressive models, demonstrate noise schedule invariance, and achieve fast optimization. It also utilizes a bits-back compression scheme for near-optimum lossless compression rates.

2.6 Imagic

The paper "Imagic: Text-Based Real Image Editing with Diffusion Models"[13] introduces Imagic, a text-to-image diffusion model for complex text-based semantic image editing. It enables sophisticated edits on real high-resolution images, including style changes, color changes, and object additions. The method combines text embedding optimization, model fine-tuning, and text embedding interpolation to achieve realistic image editing.

2.7 Diffusion Model

Midjourney Midjourney is a generative artificial intelligence that facilitates image from textual input. It can be accessed through only the discord servers by inviting bot to the certain server or texting bot. The bot will respond to the command "/imagine" followed by a prompt. It also allows the user to blend two images with the command "/blend" and suggests to the user how to shorten a long prompt in response to the "/shorten" command. Within a short time it gained popularity worldwide. A mid journey image won first place in a digital art competition at the 2022 Colorado State Fair called Théâtre D'opéra Spatial which is French for "Space Opera Theater ". Besides it has major drawbacks like comprehension "Sublime", cultural connection, inadequate credentials and job scarcity despite being creative, faster and diverse [11]. A research was conducted based on an Artistic Experiment and Architectural Experiment. The Artistic Experiment was basically engraving emotions through abstract art and the result showed that based on creativity, speed and sublime factor Midjourney was better than human beings. And based on the. Architectural Experiment which was basically portraying emotions in space, human mind output was better in creativity, uniqueness, relevance and practicality, though midjourney beaten human in speed. However it does not generate realistic images which is the main purpose of our research.

DALL-E: DALL·E model is a powerful AI model developed by OpenAI. It is designed to generate images from textual descriptions. The model was specifically designed to illustrate and modify visual contents based on textual prompts. It has been trained on a massive dataset of text and image pairs, allowing it to generate highly detailed and imaginative visuals. The DALL·E model is built on the GPT architecture. Its potential is showcased in a wide range of creative applications, which directly contributes to the evolving landscape of AI-generated visual content. DALL·E also has the ability to revolutionize various fields ranging from design and marketing to content creation.

Stable Diffusion Model: Stable Diffusion models are used as probabilistic frameworks for analyzing system evolution across various fields and has been pre-trained using different input data formats. The output of a stable diffusion model reflects the specified data of the training dataset and demonstrates a fidelity to the learned style in practical terms. This approach is particularly noticeable in image generation because the model generates content that is most likely to be representative of a specific category when trained on a set of images. The adaptability of stable diffusion models makes them useful tools for tasks that require data to be synthesized with a specific stylistic direction.

SD 1.5: Released in April 2022 but Stability AI, Stable Diffusion 1.5 belongs to a category of AI models known as latent diffusion models. This particular model has the ability to take a text prompt from you and utilize its comprehensive understanding of the world to produce an image that aligns with that prompt. Imagine it as an artist with extraordinary skills, capable of painting anything you can conceive simply by listening to your description. The process begins with a collection of random noise, which gradually transforms into an image through a series of steps guided by the text prompt you provided. It's akin to refining a blurry painting until it becomes a clear and distinct picture. This AI model is highly adaptable, easily controllable, and relatively simple to utilize. Although it is still under development and can be prone to mistakes, it is a powerful and impressive tool that has the potential to revolutionize the way we create art and communicate.

SD 2.0: Further improving on the existing 1.5, Stability AI released its predecessor, Stable Diffusion 2.0 in October, 2022. It offers a compelling upgrade for users seeking the best possible AI-generated imagery. Its improved image quality, higher resolution, more intuitive prompting, and additional features make it a versatile and powerful tool for artists, designers, and anyone interested in exploring the creative potential of AI. However, the increased computational demands and potentially limited access compared to version 1.5 are factors to consider. The Stable Diffusion Model XL is an advanced version of the diffusion model. It was released in July 2023 and it is designed to handle larger and more complex datasets. This model takes into account various factors that affect the spread of ideas or innovations, such as social influence, individual decision-making processes, and the size of the population. It provides a more comprehensive understanding of how information propagates and evolves within a society or system.

2.8 Model Comparision

The paper "Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion"[12] provided quantitative comparison among three popular systems about their ability to generate photorealistic faces. According to the most commonly used method for evaluating generative models, FID score, Stable diffusion generates better faces. This model is used to generate detailed images conditioned on text descriptions, and can be applied to other tasks such as inpainting, outpainting, and image translation. Whereas Midjourney is a model that synthesizes images from textual descriptions to generate surrealistic images and is popular among artists. And DALL-E-2 is a successor to DALL-E that can create more realistic images and combine concepts, attributes, and styles. The study evaluated the models based on both generated and real faces. For generated faces captions from the COCO dataset were used to synthesize images and ran them to MediaPipe face detector twice to detect faces and manually remove false positives which collected 15,076 generated faces, including 8,050 by Stable Diffusion, 6,350 by Midjourney, and 676 by DALL-E 2. And for real faces 13,233 faces were added from the LFW dataset and 30,000 real faces were extracted from the COCO training dataset. It was used to generate high quality images with high fidelity, diverse, and physically plausible faces. Results are averaged over 10 runs which shows Stable Diffusion scores a lower FID[less means better], but the quality of the generated faces is still much worse than the models specifically trained on portraits. DALL-E 2 performs worse than Stable Diffusion. All in all the paper concluded that though it might be a good idea to align the faces before computing the scores, Stable Diffusion generates better images than other models.

2.9 Finetuning Methods

DreamBooth DreamBooth is a deep learning model used to fine-tune text to image models. DreamBooth fine-tunes the whole diffusion model. Diffusion model can generate an image but to customize the image according to exact details in a specific subject DreamBooth is required [20]. It has 30 subjects of 15 different classes of which 9 are live subjects and rest of the 21 are objects[17]. The number of steps for DreamBooth training varies around 1400-2400. It actually depends on how many images are used to train. Overfitting steps may result in too many artifacts in output whereas underfitting will not provide the desired outcome. For training the image must be of dimension 512x512. File naming format like jpg or png is irrelevant here. Usually the number of class images required for DreamBooth is 200-300. Here a identifier called "SKS" implants distinctive features into the product domain. Two prompts: sample image prompt and negative prompt are required to generate an image. Negative prompts are what we do not need in our image like a cartoonish image can be a negative prompt since our goal is to generate realistic images whereas a sample image prompt contains information about the face required to generate. Although many papers suggest that 3-5 images are enough to train DreamBooth, minimum 25 images should be used to train.

LoRA LoRA is another model to fine-tune text to image models but it does not fine-tune the whole diffusion model rather it fine-tunes certain parameters. It is more suitable for a single subject. It is the abbreviation of Low-Rank Adaptation. The purpose of this design is to fine-tune wide-ranging models efficiently. But it has some drawbacks. As it is lightweight, fine-tunes only fixed parameters and often considered as a small stable diffusion model it is not suitable for diverse, accurate or highly quality outputs. In contrast DreamBooth fine-tunes the whole diffusion model and great for subject oriented results DreamBooth was our choice.

Chapter 3

Dataset

3.1 Dataset

Datasets play a pivotal role in the training of Artificial Intelligence (AI) models, serving as the fundamental elements that enhance their learning capabilities. Fundamentally, a dataset is a structured compilation of labeled information that allows AI algorithms to identify patterns and make forecasts. The performance and generalization ability of AI models across different applications are heavily influenced by the quality, diversity, and magnitude of the dataset. In other words, datasets are indispensable in training AI models to address a variety of challenges, ranging from image and speech recognition to natural language processing and decision-making. The importance of robust datasets lies in their ability to expose models to a wide range of scenarios, ensuring adaptability and reliability in real-world circumstances. The scenario is the same in case of Diffusion models too. Within the expansive and near-saturated landscape of generative AI models, Diffusion models stand out as a specialized selection of models that are focused on understanding information, innovation and trends within the networks. Diffusion models are trained in pairs of a wide variety of artistic images and their respective textual descriptions. But in case of our research, we are focusing on a specific field, namely, on face illustrations and for our own use, we need to choose a dataset which is exclusively based on pictures of real human faces and textual description of their facial features, such as the shape of their face, the color of their hair, size of their eyes and mouth etc.

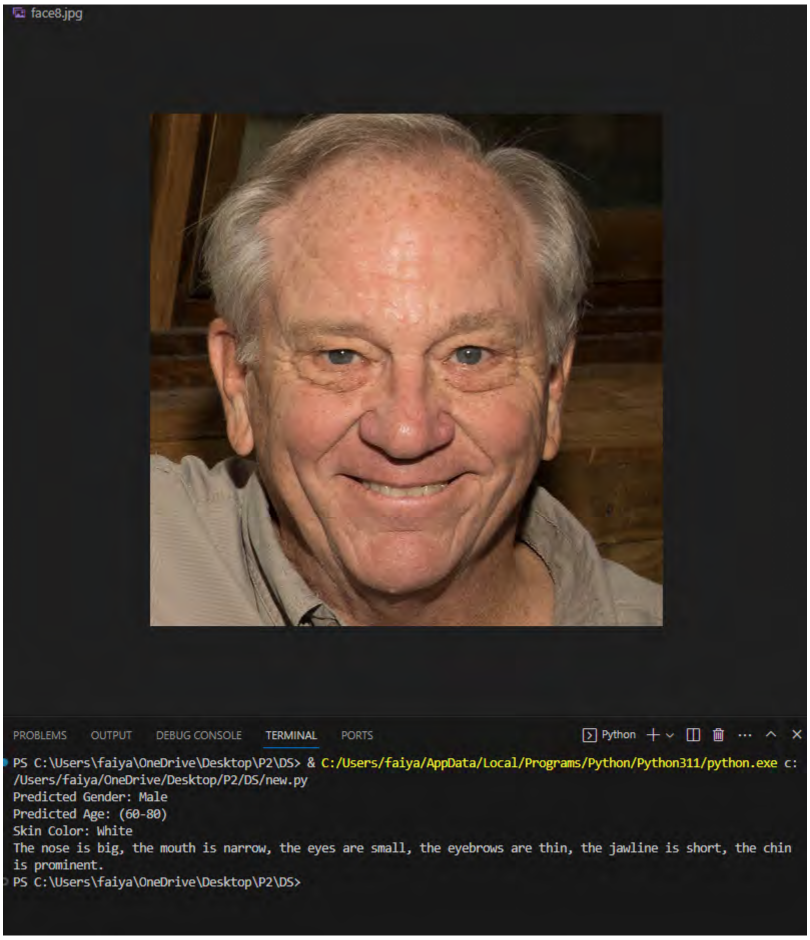
3.2 Data Collection

Dataset curation and selection plays a very vital role in the development of a robust AI model. The size of a dataset, the diversity of the contents inside it, quality and relevance of a dataset significantly impact the model's ability to generate and produce relevant outputs in real-world scenarios. The selection of a relevant and diverse dataset ensures that the model can learn from a collection of accurate, varied and representative examples. It also ensures that the model is independent of bias and is responsible for improving its adaptability. It is essential to select a dataset which has the perfect ratio of quality and quantity, tailored to one's specific need as it optimizes the learning process, as well as the overall performance of the system. As for our own use case, we need datasets that contained both human faces as well as textual descriptions about them, we tried to search for such datasets in places like Kaggle, datahub etc. and found multiple datasets containing images of real human faces only, like CelebA, CIFAR,MS COCO, ImageNet and many more. Unfortunately, we were unable to find any dataset that exactly catered to our needs. Among all the options, we chose Flickr-Face-HQ (FFHQ) as our preferred dataset due to its widespread use in fields like computer vision, face detection and machine learning. It consists of a diverse collection of high-resolution human face images belonging to different races, ethnicities, ages and genders. These images also portray people in various lighting conditions, with a variety of facial expressions, natural and real poses etc. which makes this dataset a natural choice for our goal. The reason for choosing this dataset over the others is because of the abundance of natural photos of normal, everyday people which was lacking in other datasets like for example CelebA, which mainly contains photos of celebrities that are conventionally appealing or attractive. We wanted a dataset that can train our model using facial features and expressions of real people in their everyday lives. And that is why we opted for FFHQ as opposed to the other options. From the FFHQ dataset, we selected 2500 photos and added textual descriptions to each of the images. This annotation provides a detailed explanation about the shape of a person's face, the size and shape of their nose, the color and size of their eyes, size of their mouth and so on.

Figure 3.1: FFHQ (Flickr-Faces-HQ) Dataset

3.3 Previous Attempts at Dataset Preprocessing

In order to overcome the obstacle of a lack of dataset, we attempted to create our own, which resulted in us taking multiple varied approaches to solve this problem. One of such approaches was to develop an algorithm using OpenCV and models like Caffemodel to use the numerical values or points from facial landmarks in order to define certain attributes regarding the eyes, nose, eyebrows, lips, ears, jawline, chin, skin tone, gender, age, etc.



```
PS C:\Users\faiya\OneDrive\Desktop\P2\DS> & C:/Users/faiya/AppData/Local/Programs/Python/Python311/python.exe c:
/Users/faiya/OneDrive/Desktop/P2/DS/new.py
Predicted Gender: Male
Predicted Age: (60-80)
Skin Color: White
The nose is big, the mouth is narrow, the eyes are small, the eyebrows are thin, the jawline is short, the chin
is prominent.
PS C:\Users\faiya\OneDrive\Desktop\P2\DS>
```

Figure 3.2: Defining Facial Attributes based on Landmarks

Additionally, we tried to further enhance our system to generate specific descriptions of worn attires that can be observed in the images. This included defining clothing items such as shirts, coats, or dresses that individuals in the images are wearing. Moreover, we tried to specify whether a person is wearing a hat and if not we tried to describe their hairstyle. Furthermore, we tried to determine whether individuals in the images are wearing glasses or not. These detailed descriptions would have contributed to building better semantics and understanding for the model. Since the dataset already contains facial landmarks generated using DLib, we only needed to run our algorithm on the metadata of the processed FFHQ dataset. Afterwards, we tried to diversify the predefined attributes from the algorithm properly. In order to achieve this goal, we used natural language generation (NLG). Lastly, we combined the facial descriptions with each of the respective images and prepared our new custom dataset which is to be divided or splitted into training, validation, and testing subsets, for model evaluation. However, during our actual implementation, it was noticed that we couldn't get a descriptive annotation of different facial features such as the shape of that landmark (for example- snub nose, pointed ears etc.). It was only able to give the size of any landmark (for example- short nose, long ears etc.), and even that wasn't completely accurate. What the algorithm did was basically draw a few points on certain landmarks on one's face and measured the Euclidean distance between each landmark, thus determining the length and size of certain facial features, such as mouth and eyes. And even those were not fully accurate all the time because of a few limitations such as the inconsistent distance of the face from the lens, inability to comprehend the features in case of side-profile images etc. This model also struggled to define the race/ ethnicity of the faces in the dataset. Overall, the algorithm was unable to provide us with descriptive and accurate annotations for the photos. Therefore, we search for alternative methods to complete the task.

3.4 Dataset Preprocessing

After selecting FFHQ as our primary dataset, we carried out the necessary preprocessing and exploratory data analysis on it. We first ensured that the dataset contained the images in the dimensions we wanted. FFHQ comes in 1024*1024 by default. First, we converted all the images that were going into our dataset into 512*512, which was our desired dimensions. Next we double checked to see whether it was correct or not. To do that, we developed a python script that checked all the images manually and guaranteed that the dimensions of each image were precisely (512*512). We also assessed the scaling and structure, examined metadata such as landmarks of each image, made sure that the face is the central aspect of the image etc. Since all our previous attempts were not up to the expected standard, we decided to label the data manually and curate our own custom dataset. While developing this new dataset, we selected FFHQ as our baseline and took pictures of individuals from it . During this image selection process, we made sure that each photo contained the face of only one person. This was done by using OpenCV facial detection, which removed images that contain more than one human face. We took extra measures to keep our data as streamlined and efficient as possible. For example, omitting photos of- people in very heavy makeup (cosplayers for example), people with unusual or absurd expressions on their face etc. as they are irrelevant to our use case. Then we proceeded to label each and every one of the selected images manually, describing in detail about their facial landmarks such as the shape of their nose, size of their eyes, color of their hair, shape of their face etc. We also took into account what the subject in the photo was wearing including dresses, sunglasses, earrings and other accessories. The textual descriptions were made as concise as possible and we implemented commonly used terms. In addition, We tried to make sure that these descriptions were devoid of any sort of adjectives and maintained impartiality. For instance, we avoided terms like beautiful, cute, handsome, ugly etc. as these words make the image subjective and biased. We tried to keep the data as factual and unbiased as possible. Moreover, ethical aspects like consent, privacy etc. were also carefully administered while implementing the popular FFHQ dataset. Our workflow is shown in the diagram next page:

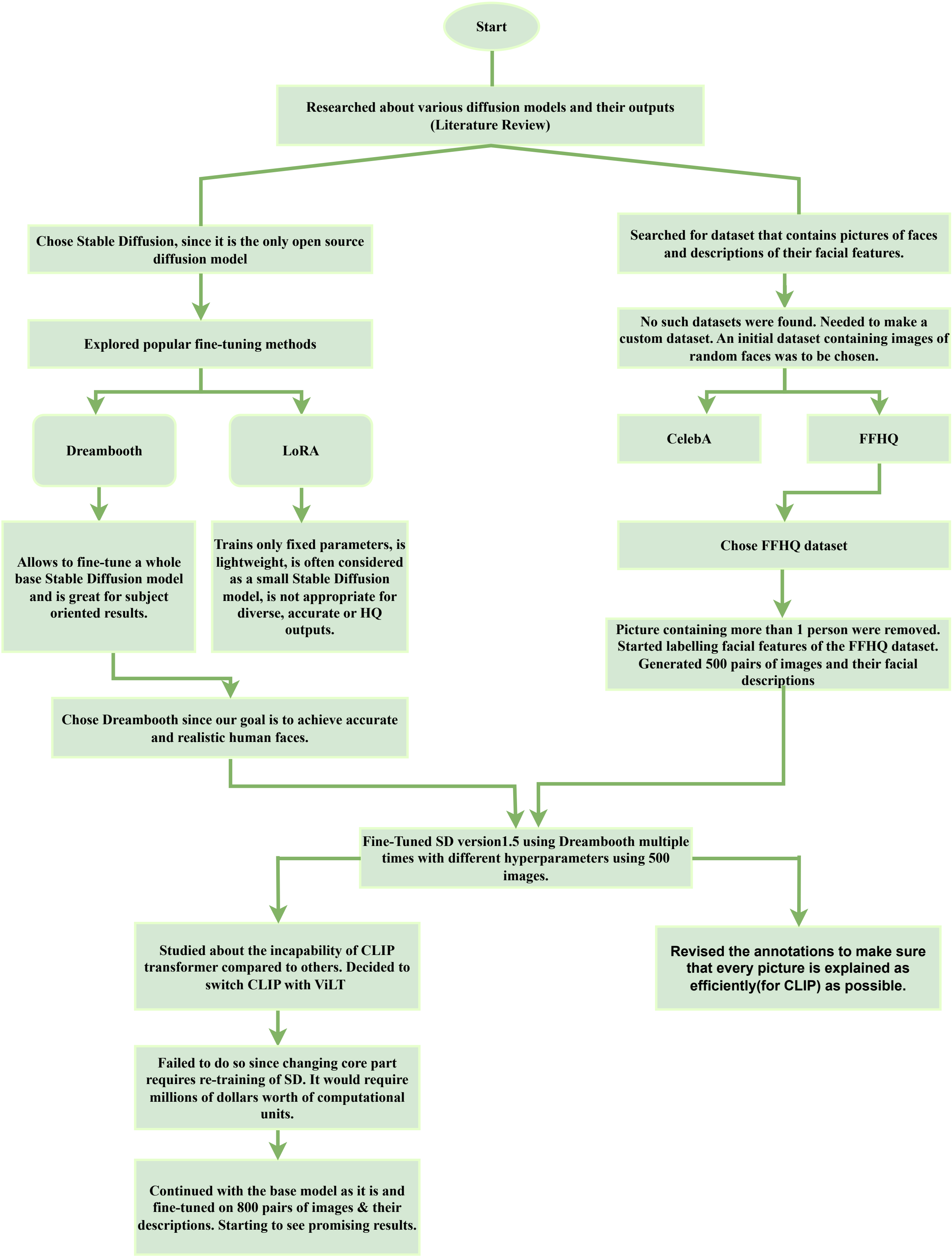

Start
Researched about various diffusion models and their outputs (Literature Review)
Chose Stable Diffusion, since it is the only open source diffusion model
Searched for dataset that contains pictures of faces and descriptions of their facial features.
Explored popular fine-tuning methods
No such datasets were found. Needed to make a custom dataset. An initial dataset containing images of random faces was to be chosen.
Dreambooth
LoRA
CelebA
FFHQ
Allows to fine-tune a whole base Stable Diffusion model and is great for subject oriented results.
Trains only fixed parameters, is lightweight, is often considered as a small Stable Diffusion model, is not appropriate for diverse, accurate or HQ outputs.
Chose FFHQ dataset
Picture containing more than 1 person were removed. Started labelling facial features of the FFHQ dataset. Generated 500 pairs of images and their facial descriptions
Chose Dreambooth since our goal is to achieve accurate and realistic human faces.
Fine-Tuned SD version1.5 using Dreambooth multiple times with different hyperparameters using 500 images.
Studied about the incapability of CLIP transformer compared to others. Decided to switch CLIP with ViLT
Revised the annotations to make sure that every picture is explained as efficiently(for CLIP) as possible.
Failed to do so since changing core part requires re-training of SD. It would require millions of dollars worth of computational units.
Continued with the base model as it is and fine-tuned on 800 pairs of images & their descriptions. Starting to see promising results.

Figure 3.3: Work Flow

3.5 Challenges Faced

Now the question may arise about the bias of the observer, and in order to mitigate that, we followed a set of strict predefined templates and only defined landmarks based on those previously set data.

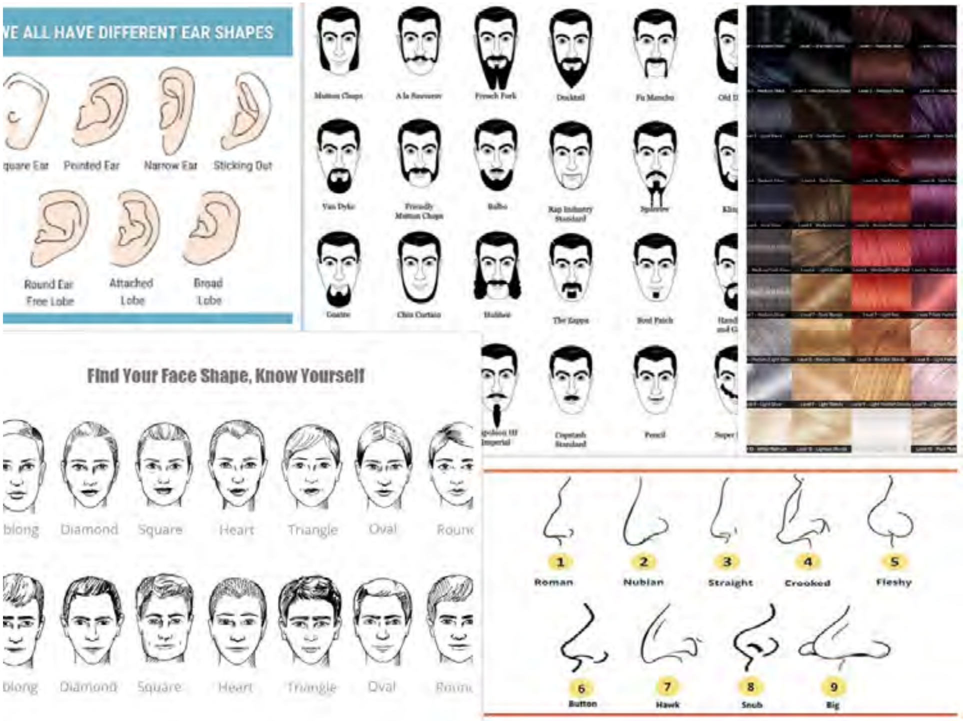


Figure 3.4: Facial Landmark Templates

During the process of dataset curation, we came across a number of obstacles which were stopping us from achieving our desired goals. Even though we tried to minimize the said obstacles as much as we could, still some of them remained unsolvable under current circumstances. We hope to overcome these obstacles in the near future. One of the significant obstacles we faced was trying to remove bias. We tried to not use any adjectives and only write factual data, but since our dataset was done manually by multiple people, naturally a little bit of subjectiveness and bias became apparent. Another one of such hindrances faced by us was the perception of everyone being different. Even though we tried to keep the data as factual as possible and used the same templates during annotation, a few data points such as the shape of someone's nose were rather subjective and differed from person to person. This in turn, caused dataset inaccuracy, which further resulted in the model not being able to correctly perceive and portray certain things. The manual annotation process also proved to be very lengthy and as a result of this we were able to only create a dataset of only 2500 data, which is low for training a diffusion model. Initially we did about a thousand data, but after fine tuning we gradually understood the type of prompt that is the most effective for our CLIP transformer, and thus we started fresh, which resulted in the low amount of collected data. Lastly, Our collected dataset was slightly lacking when it came to people of certain ethnicities such as Indians, Latin Americans etc., which resulted in the creation of bias within the model. Since the number of data of a certain field is relatively low, the model automatically creates stereotypes by taking into account the existing data. For example, when we instructed our model to illustrate the image of an asian woman, it defaulted to portraying her with a snub nose, or when instructed to do the same for a black male, it defaulted to portray him with curly hair. This limitation takes place due to a lack of diverse data. By adding more distinguished data for a few specific ethnicities or classes, our dataset can be made more diverse.

Chapter 4

Model

In this section, we will present our key components and architectural choices of our proposed system for realistic face generation. Our approach comprises several cutting-edge technologies including latent diffusion models with several significant factors like the U-net architecture, a pre-trained neural network by OpenAI that is CLIP, VAE and CLIP based conditioning. The selection of Stable Diffusion is a no-brainer when it comes to working with diffusion models since it is the most if not the only reliable diffusion model that is fully open source and available to the public. Other state of the art generative AI that is, diffusion models like DALL-E by OpenAI, Imagen by Google, Midjourney etc. are all large and complex models which are owned privately and are closed-source. To the public, some of these models can only be used to generate images from prompts that too after paying a hefty subscription fee. For our research, we explored all possible options and ended up with Stable Diffusion because we ultimately require a diffusion model that we can fine-tune on our own custom dataset without spending a fortune.

Stable Diffusion in its core, is a Latent Diffusion Model, commonly known as LDM. Compared to conventional Diffusion Models, LDM uses its component VAE, which is a popular neural network architecture which uses a combination of encoder and decoder. The encoder of LDM allows to extract crucial information and details from images and convert them into low dimensional representations of vectors of numbers which are also termed as latent codes. These vectors or latent codes are further sent to a subspace which is popularly known as the latent space. Latent space is a low dimensional mapping where each point represents an image. The images that have similar attributes and features, have their latent codes stored closer in the latent space. Thus, Stable Diffusion allows a much more compressed workflow compared to other popular diffusion models. For instance, in most cases it is observed that Diffusion Models directly work with pixels of images where noise is added to them gradually in t steps and in backward diffusion, the pixels of the noisy images are gradually removed and an output is acquired. As a result, the higher the resolution of the image, the more advanced computational units would have to be used. But in the case of Stable Diffusion, it solves this problem by converting images in a latent space. After the conversion, Gaussian noise is added to the latent codes or vectors gradually in small steps until we would get nothing but noise thus completing the forward diffusion.

Mathematically, forward diffusion takes place following the given equation:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{4.1}$$

Where:

- x_t: Image at step
- t_q: Transition probability of the diffusion process
- β_t: Noise schedule parameter controlling noise level at each step
- I: Identity matrix

During backward diffusion, denoising is done by using the U-Net and the VAE allows to up-sample or simply convert the denoised latent code into an actual image. However, it is also important to note that Stable Diffusion is not only a LDM but is arguably a hybrid model since it also uses CLIP. CLIP is a multimodal neural network model that comprises of an image as well as a text encoder. The image encoder is practically a Vision Transformer (ViT) whereas the text encoder is based on a text transformer architecture like BERT. What CLIP does is that it low dimensional vector representations of both an image and its respective textual description. Then it sends these latent codes to its own shared latent space where it aligns the image and textual vectors closely as a result of which it achieves a deeper understanding of the semantics behind the connection of an image and its textual description. Moving on, the usual forward diffusion takes place as we have already discussed. When a prompt is given by the user, backward diffusion takes place. With the help of an embedded prompt from the text encoder of CLIP, LDM has better guidance to denoise appropriately. Now, if we are to elaborate about the backward diffusion process that is primarily done by the U-Net, we must know what U-Net is. U-net architecture is based on convolutional neural network architecture, which is mainly known for its outstanding performance in image-text related tasks. It was highly used in medical imaging and is often seen in Generative AI nowadays. The design of the U-Net architecture resembles the letter 'U' which consists of an encoder and a decoder with skip connections and a bridge connecting the encoder-decoder blocks. Skip connections establish the connection between the two blocks in reality and help in the denoising process of the images. U-Net consists of convolutional layers with the help of which the encoder blocks down-samples the input image at first. Down-sampling is the extraction of high-level features and details in order to down-sample the original image. However, it is not the same as the encoder of VAE which is responsible for generating vector representations of the images. Decoder does exactly the opposite of the encoding method. That is, with the help of skip connections, the decoder takes the low-resolution features separated from the encoder part and up-samples them to try to recreate the data back to its original state. With the help of skip connections, the model can access both global and local features continuously. In our case, it is to be noted that Stable Diffusion uses the U-Net which consists of ResNet backbone for denoising images during the backward process.

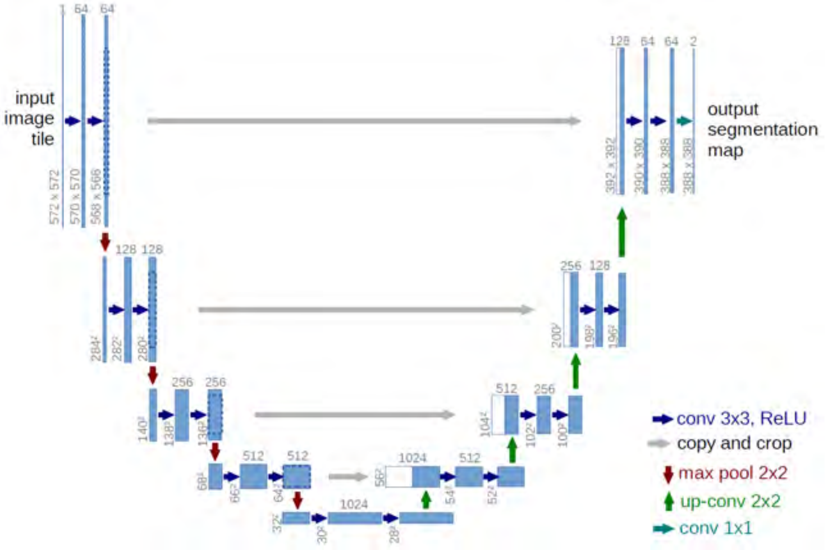


Figure 4.1: U-Net Architecture

Coming back to the working process of Stable Diffusion, as we have just discussed, the denoising process is mainly done using the decoder of U-Net, which is a fully convolutional network architecture consisting of encoder-decoder blocks. At first conditioning is done from the embedded prompt by CLIP's text encoder. The conditioned prompt is sent to the decoder of U-Net which gradually in small steps removes the noises. Again, as we know, U-Net has skip connections that enables cross-attention to create a far better understanding of features between texts and images. Ultimately, a clean vector representation is acquired which is finally converted to an actual image by VAE, which again, is a component of the LDM itself.

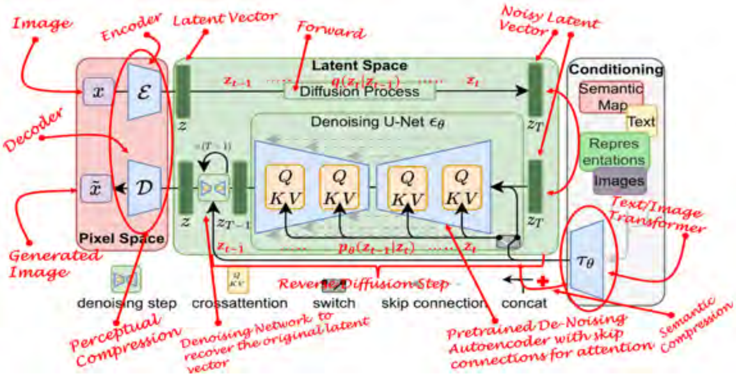


Figure 4.2: Workflow of Stable Diffusion

4.1 Model Selection

After targeting the usage of Stable Diffusion in our research, the deeper we got into the world of stable diffusion, the more we learnt about the different available “versions” or simply said, checkpoints. Although, people were excited about stable diffusion being the first and only open-source diffusion model, it was certain that the results or the generated images were of unrealistic, too artistic, low quality when compared to the existing models like Midjourney, DALL-E-2 etc.

Figure 4.3: Official first samples by Stable Diffusion

But it is noteworthy to say that the phenomenon of making the model fully public created sparks in the world of generative AI due to its infinite potential and possibilities. Ultimately, those possibilities started coming to life since the researchers started fine-tuning Stable Diffusion on large amounts of data which was from multiple subsets of the LAION-5B (en) dataset. They started from smaller amounts of data like the LAION-2B (en) and for each iteration, they created a version checkpoint like SD V1.1, SD V1.2, SDV1.3 etc. and so on. They also made sure to include FID scores and other evaluation metrics, testing and comparing it with the COCO 2017 dataset which is widely used for evaluating Diffusion Models. The most popular checkpoint or version released by Runway was the SD V1.5, which is still being widely used up to this date. Even we have chosen SD V1.5 as our base model to fine-tune with our own custom dataset consisting of real human faces along with textual descriptions of their facial features. Based on the requirement of low computational resources of SD V1.5 and on the hundreds of implementations of this model, we believe in its potential to achieve realistic faces even with our relatively small dataset. An exemplary image of a real human face generated by using this checkpoint is shown below.

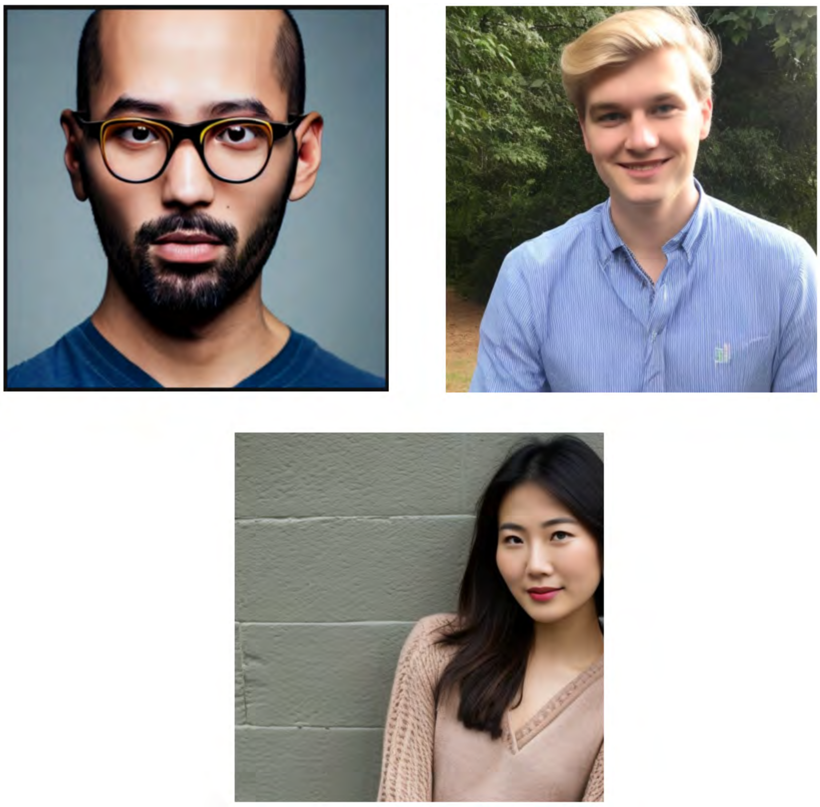


Figure 4.4: Outputs from facial feature prompts by SD1.5

4.2 Model Descriptions

We have already discussed the fact that we have selected Stable Diffusion Version 1.5. However, since we are highly positive on the capabilities of Stable Diffusion in achieving accuracy as well as realism in real face generation, it is a must that we compare with other available checkpoints. We have also chosen to compete with DALL-E-3, which is the latest Diffusion Model developed by OpenAI and is included with the highly popular LLM ChatGPT which enables users to try out text-to-image generations using DALL-E at a fair cost. We have also chosen to compare with further versions of Stable Diffusion. First and foremost, we have to show the improvements from our base SD V1.5. Secondly, we chose official releases of upgraded Stable Diffusion models that are the SD V2.0 and SDXL. It is to be noted that SDXL is the latest available Diffusion Model available by Stability AI which is recognized as one of the best Diffusion Models currently available in the world of Generative AI. We still find it important to include short descriptions of all the models that are going to be implemented, compared to and displayed in this research.

4.2.1 SD Version1.5

Released in August 2022, according to their descriptions on Hugging Face, their 5th iteration, that is the SD V1.5, enhances text-image alignment despite dropping 10% of text-conditioning to promote creativity rather than strictly sticking to prompts. While it not an official release from the founders of Stable Diffusion that is, from Stability AI, it still offers refined capabilities for research on diffusion model enhancements. We have already included exemplary pictures from the base model Stable Diffusion V1.5 where we put different prompts consisting of facial features and got slightly realistic outputs or generated images of human faces. It is quite extraordinary to see how better the results get after fine-tuning SD1.5 on our custom dataset which is in fact quite small compared to that of the amount of data the actual base checkpoint was fine-tuned on. It will only further prove the capability of Stable Diffusion models to accurately generate real human faces from prompts describing their facial features similar to how a real-life sketch artist would work.

4.2.2 SD Version2.0

In SD2.0 they re-trained the whole stable diffusion model instead of fine-tuning over an already existing checkpoint. They also replaced their text encoder to OpenCLIP which is a large-scale open-source version of the frozen pre-trained CLIP model that was used initially with Stable Diffusion. They also added x4 upscaling which boosts the quality of the images. Nonetheless, the accuracy in generating real faces were still quite low and many community forums believe the SD 2.0 is not optimal for real face generations.

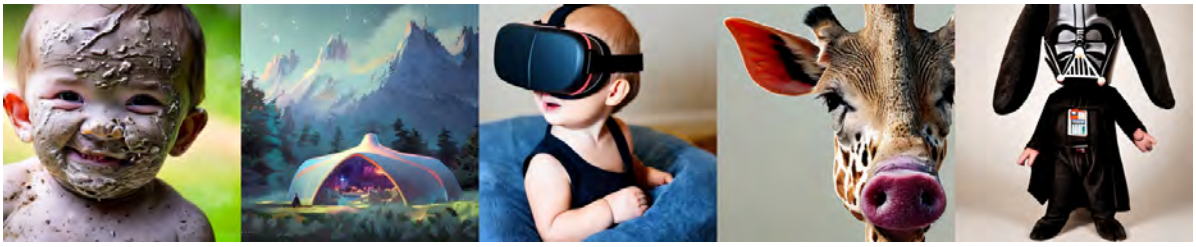


Figure 4.5: Stable Diffusion 2.0

Despite having higher quality, discussions and research were soon to point out the fact that SD V2.0 did not achieve accuracy for aligning the output along the prompt compared to that of SD V1.5. Forums and discussions were also keen on showing the cartoonish results in both of the models.

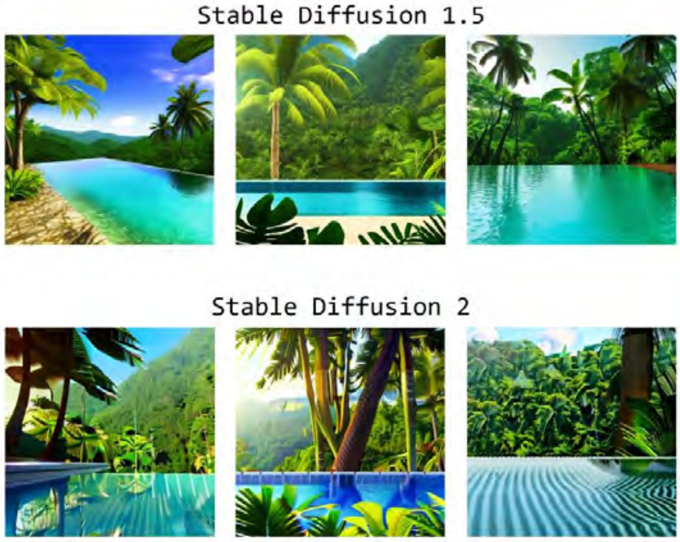


Figure 4.6: SD1.5 VS SD2.0

Stable Diffusion 2 released by Stability AI can be said to be more creative but lacks realism more than before. Here is a comparison between them generating images on the same prompts, acquired from an article by AssemblyAI:

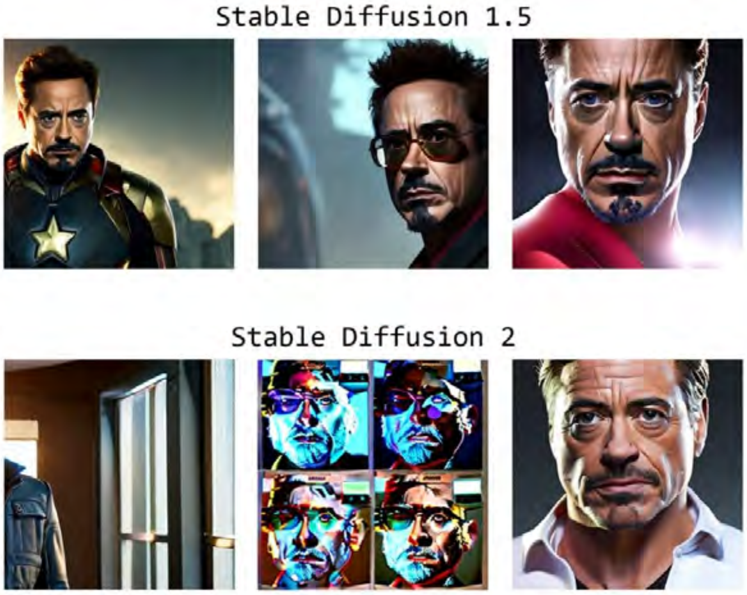


Figure 4.7: SD 1.5 vs SD 2.0
[10]

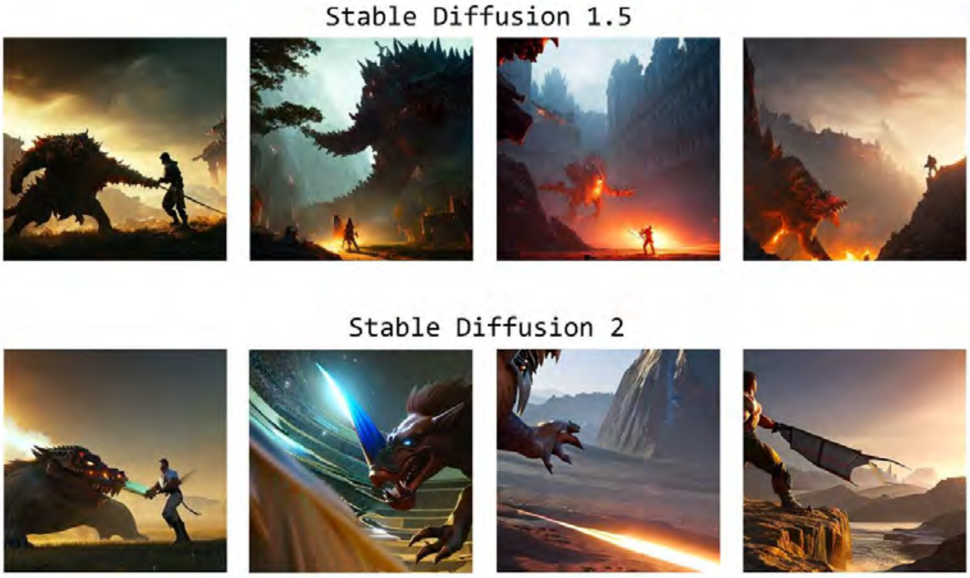


Figure 4.8: SD 1.5 vs SD 2.0
[10]

4.2.3 SDXL

Stable Diffusion XL is the latest running Stable Diffusion model that has been officially been released by Stability AI. It is being widely used in image generation tasks all over the world. It has a 3 times larger U-Net than the base stable diffusion model. Image quality has highly improved as well as image resolutions which were only 512 X 512 in SD 1.5 and SD 2.0. In SDXL the resolutions of outputs are 1024 X 1024. Moreover, SD1.5 and SD2.0 were both trained on 5 billion data whereas SDXL proudly claims to be trained on 6.6 billion images. It also claims to have achieved better textual conditioning and dictation. SDXL is commonly accepted as the model to achieve the most variance and vibrance in styles of image generations. It might seem like SDXL should be our optimal choice for our project. Unfortunately, the massive improvements of SDXL come at a cost, that is, it requires high amounts of computational resources to fine-tune on. Yet, we are still putting it up for a comparison which would hopefully raise positive expectations about the future implications of stable diffusion models like SDXL in accurate and realistic face generation.

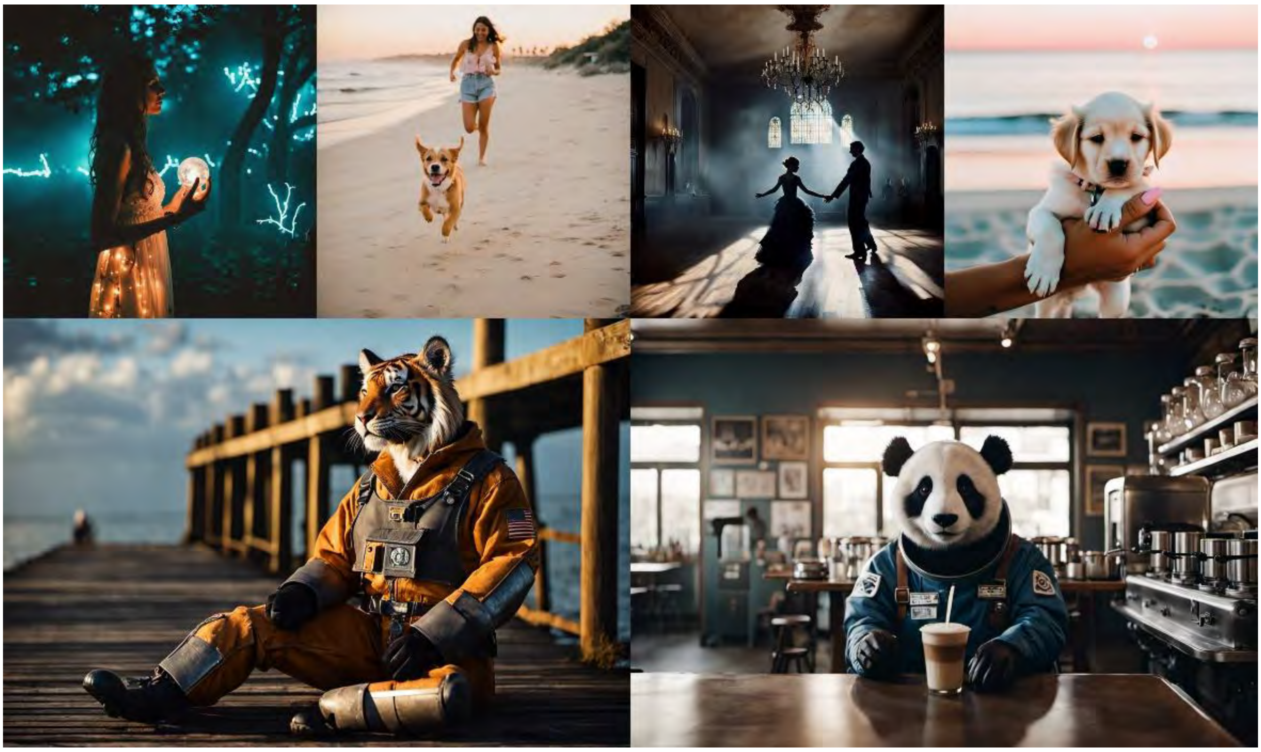


Figure 4.9: Official examples of SDXL

4.2.4 DALL-E-3

After the huge popularity of DALL-E-2, OpenAI introduced their latest, most powerful Diffusion Model that is DALL-E-3. Microsoft offers 15 free text to image generations powered by DALL-E-3 whereas ChatGPT4, i.e the premium version also allows the usage of DALL-E-3. Looking at the performance of DALL-E-3, it can be easily said that DALL-E-3 indeed seems to have higher quality of images. But one might argue that their pictures are too smooth or perfect to be realistic. OpenAI recaptioned their Dataset by building an Image captioner, fine-tuning the captioner and evaluating the dataset for re-training. Their main focus seemed to be in improving generated images by improving the captions or textual semantics. Nevertheless, it is important to show that even a free to use, open-source diffusion model like Stable Diffusion can compete in the field of realistic and accurate face generation if it is fine-tuned appropriately.

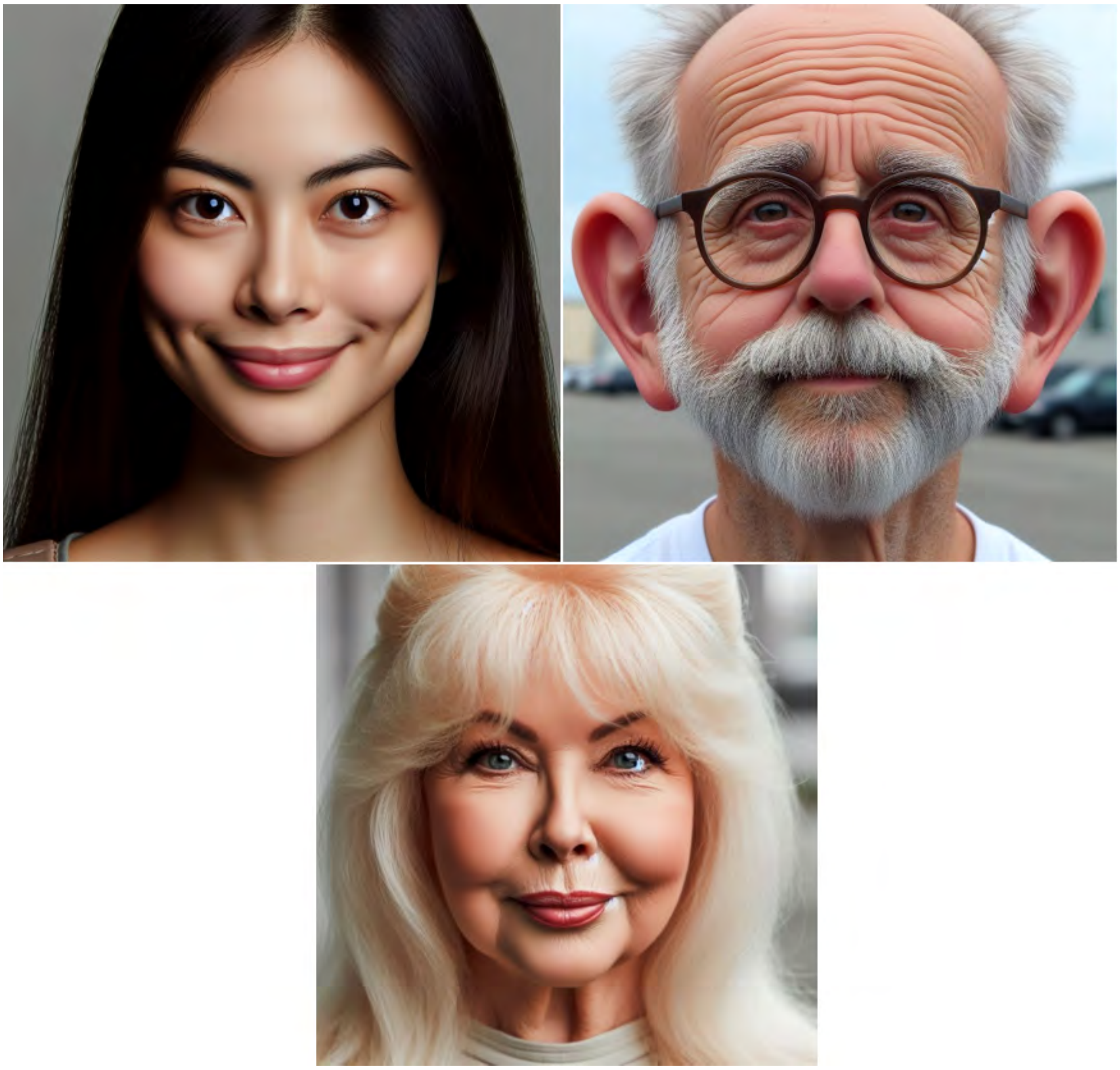


Figure 4.10: Real Face Generation using DALL-E-3

Chapter 5

Model Training

As we have previously discussed, our selected base model is Stable Diffusion V1.5. It's a fine-tuned checkpoint of the official first release of the Stable Diffusion Model. Unfortunately, we have to stick to "fine-tuning" instead of "training" our Stable Diffusion model from scratch. The sole reason behind this is that training a Diffusion Model from scratch, even for an optimized model like Stable Diffusion, requires hundreds of powerful GPU's with CUDA kernel support which ultimately would cost millions of dollars. This is the main reason why almost all of the general or common people are seen to be fine-tuning already released base models of Stable Diffusion instead of retraining the whole model from start. But it is noteworthy that it did not take much time for efficient fine-tuning research and procedures to be released specifically for Stable Diffusion. These models are discussed below.

5.1 Model Training Environment

As we have already mentioned, we have chosen Stable Diffusion Version 1.5 checkpoint as our base model. We will be fine-tuning the base model using DreamBooth, one out of a few popular fine-tuning methods available publicly for the sole purpose of fine-tuning stable diffusion models. To be more specific, we are using a newer implementation of DreamBooth, known as fastDreambooth that requires a GPU with at least16GB of VRAM. Since none of us fulfills the computational cost criteria for this, we chose Google Colab as our medium or environment to train SD V1.5 on a custom dataset comprising of FFHQ and textual descriptions.

Items	Parameters
OS	Virtual Linux-based environment (Google Colab)
GPU	Tesla4 (T4) 16GB VRAM
CPU	2 Virtual CPUs (VCPU)
RAM	13GB DDR4
Programming Language	Python 3.10.12
Machine Learning Library	PyTorch

Table 5.1: Training Environment

5.2 Fine Tuning Methods

In only a few years of time, multiple highly efficient methods of training or fine-tuning stable diffusion models or checkpoints have been released by the general users and researchers. These methods not only allow us to fine-tune a whole diffusion model, but also accomplishes it with very limited computational resources compared to the requirements of resources for training from scratch. This allows us to use consumer-based hardware to fine-tune Stable Diffusion V1.5 on our custom dataset without much issues. Some of the most popular fine-tuning methods have been discussed below including DreamBooth i.e the method that we followed for fine-tuning SD1.5 checkpoint. It is noteworthy that Automatic1111, a term that newcomers will often see alongside Stable Diffusion topics, is not a fine-tuning method. It is an advanced UI that allows loading any checkpoint Stable Diffusion models and lets users run text-to-image generations. Moreover, they do also have options for fine-tuning included in the UI but it uses either one of the given fine-tuning methods.

5.2.1 DreamBooth

DreamBooth is by far one of the most widely used methods for fine-tuning text-to-image diffusion models. Up to this date, DreamBooth methodology is being updated and implemented differently by different researchers and users to increase efficiency and accuracy of fine-tuning. It is mostly claimed that DreamBooth is best for fine-tuning Stable Diffusion models for single subject. DreamBooth allows two type of training prompts [17]. One being the “Class Prompts” which basically describes which large group the subject belongs to. For instance, if we were to fine-tune Stable Diffusion on a specific breed of dog, the class prompt would be “Photo of a Dog” or if the subject was a real person, the class prompt would be “Photo of a Person”. Then, DreamBooth allows the usage of Instance Prompts which altogether enables to enter a unique identifier token for the subject that will ensure that the fine-tuned Stable Diffusion explicitly generates an output based on its fine-tuning dataset. Simply put, if we are to elaborate the whole process of how DreamBooth works, it takes instance images for training on a subject, creates a unique identifier based on the class prompt of the subject and a unique token for the subject, then fine-tunes the U-Net components by encoding the unique identifier into the Stable Diffusion model. During testing, if the unique token is mentioned in the prompt, the unique identifier allows the Diffusion Model to accurately portray the visual features it had learnt from the instance images.

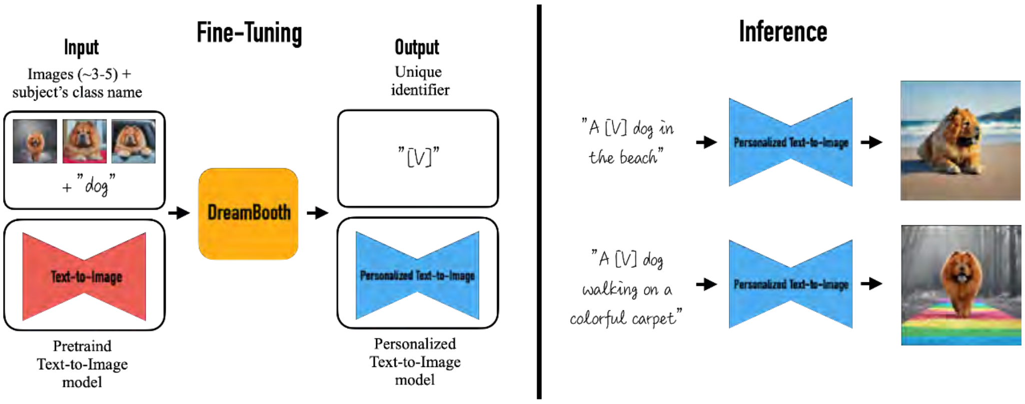


Figure 5.1: Simplified Workflow of DreamBooth

If we are to describe the working process in elaborate terms, at first the user provides pairs of images and instance prompts which helps to establish a unique identifier of the subject while the class prompts helps to fine tune a low-resolution text-to-image stable diffusion model by a method called "preservation loss" that enables to generate diverse images on that specific class based on the visual features of the given unique identifier or simply put, the subject. It then fine-tunes the super resolution components of Stable Diffusion with pairs of low- and high-resolution images from the dataset in order to maintain quality and fidelity. Although the paper of DreamBooth does not explicitly mention what the "Super Resolution Components" are, yet we are safe to assume it hints towards fine tuning the U-Net which is followed by a Super Resolution network that is responsible for upscaling and refining the outputs.

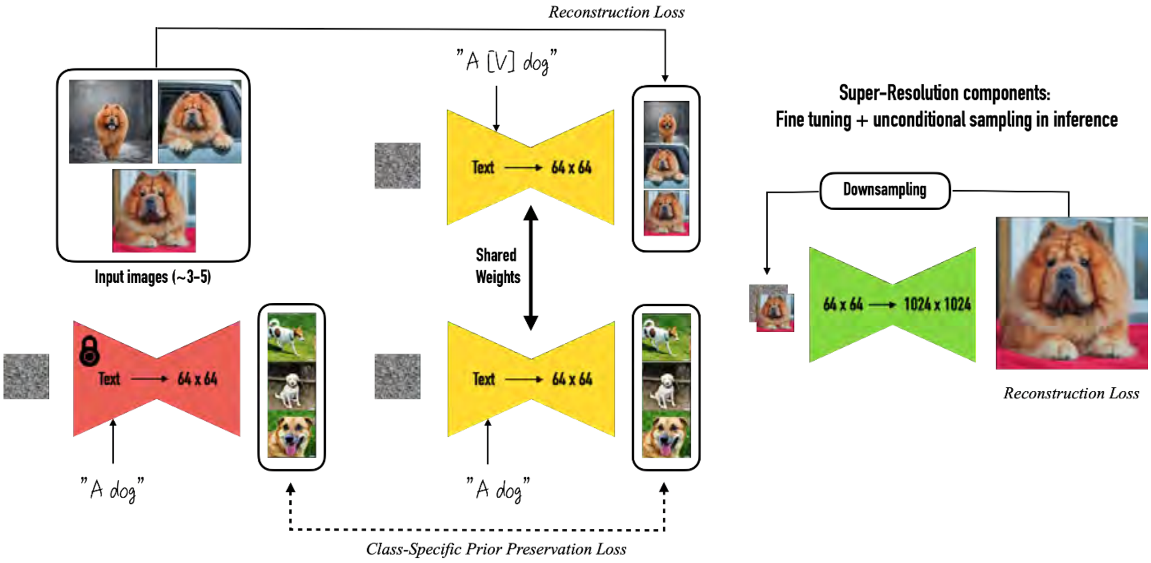


Figure 5.2: Detailed Workflow of DreamBooth

Now, a question might arise that DreamBooth is proven to be efficient for specific subject-driven image generation or personalization of stable diffusion models. How would that be effective for us, where our dataset consists of images of random people with random facial features? However, it is significant to state that multiple renditions or workarounds implementing the original DreamBooth method have been

Features	DreamBooth	fastDreamBooth
Training time	Slower	Faster
Memory requirement	Higher	Lower
Hardware requirements	Requires a GPU with atleast 30GB of VRAM	Can run on GPUs with 16GB of VRAM
Ease of use	More complex to set up	Easier to set up

Table 5.2: DreamBooth VS FastDreamBooth

released publicly which allows single captions or instance prompts for each and every training image. And it is safe to say that our class prompt is still “Photo of a person”. One such methodology that we have chosen to use is known as “Fast Stable Diffusion” or in our specific case, “fastDreambooth”. This process allows us to fine-tune any Stable Diffusion Checkpoint model at a significantly higher speed along with lower requirements for computational units. For instance, DreamBooth requires high-memory GPUs with VRAM of 30GB or more. On the other hand, fastDreambooth requires only 16GB VRAM GPUs which is now available for free from T4 GPUs in Google Colab Notebooks. FastDreambooth achieves this phenomenon by implementing a process called gradient checkpointing. Normally, DreamBooth stores all intermediate activations of different layers of the networks during backpropagation. What gradient checkpointing does is that instead of storing the activations, it dynamically recomputes parts of the network during backpropagation. Lastly, backpropagation is an important step in training neural networks which allows the network to improve by readjusting its internal parameters and weights. Additionally, fastDreamBooth also gives us the option to fine-tune the text encoder with each and every one of the instance prompts which allows the model to have a better understanding of the semantics and connections between the written facial features and the images. After fine-tuning we get a ckpt or safetensor format of our fine-tuned custom diffusion model.

5.2.2 LoRA

LoRA which is also elaborately known as Low-Rank Adaptation, is another widely popular method to fine-tune stable diffusion models. Unlike DreamBooth, LoRA does not fine-tune the entire U-Net of the targeted Stable Diffusion model. Instead, it uses its own additional low-rank layers in the U-Net thus creating its own impact in the denoising or backward diffusion process. As a result, in comparison to DreamBooth, it does not even fine-tune the entire Stable Diffusion model but only encodes its own layers in the denoising process. If we describe the process of LoRA in details, it works with the cross-attention layers of the U-Net which we know is responsible for denoising based on the conditioned embedded prompts. It analyzes the attention layers and compresses them in very small matrices. We can compare this phenomenon to that of latent codes done by the LDM without the latent space. These small matrices also contain the essential information about the actual cross-attention layers and these compressed, shortened layers, which we call its own low-rank layers, are fine-tuned on a custom dataset instead of the original. Once training is done, the low-rank layers or small matrices are recombined with the original cross-attention layers of the Stable Diffusion as a result of which the model does retain information about the dataset it was fine-tuned with. Due to this process, LoRA require substantially lower amounts of computing units compared to DreamBooth and is much faster in training as well as in generation. However, the impact or presence of Black Boxing is much higher in LoRA and it is not suitable for our work which requires accurate alignment with given textual prompts of facial features. Moreover, the quality or fidelity generated by LoRA are not as clear as DreamBooth. LoRA are still used at a huge rate for its success in generating strict subject specific creative images. It is widely accepted as the best process to fine-tune based on a specific subject. But we ignored this widely accepted fine-tuning process for the mentioned reasons. However, we still would like to show the plausible performance of LoRA in single subject generations. A single person's self-portraits in different styles generated by Stable Diffusion, fine tuned using LoRA where we can specific art styles of the images. It is to be mentioned that it is possible to generate realistic images using LoRA as well but it requires additional models like Realistic Vision 1.3, a model that contains external DreamBooth training. LoRA still can be looked as the second best option based on our availibility of computational resources.

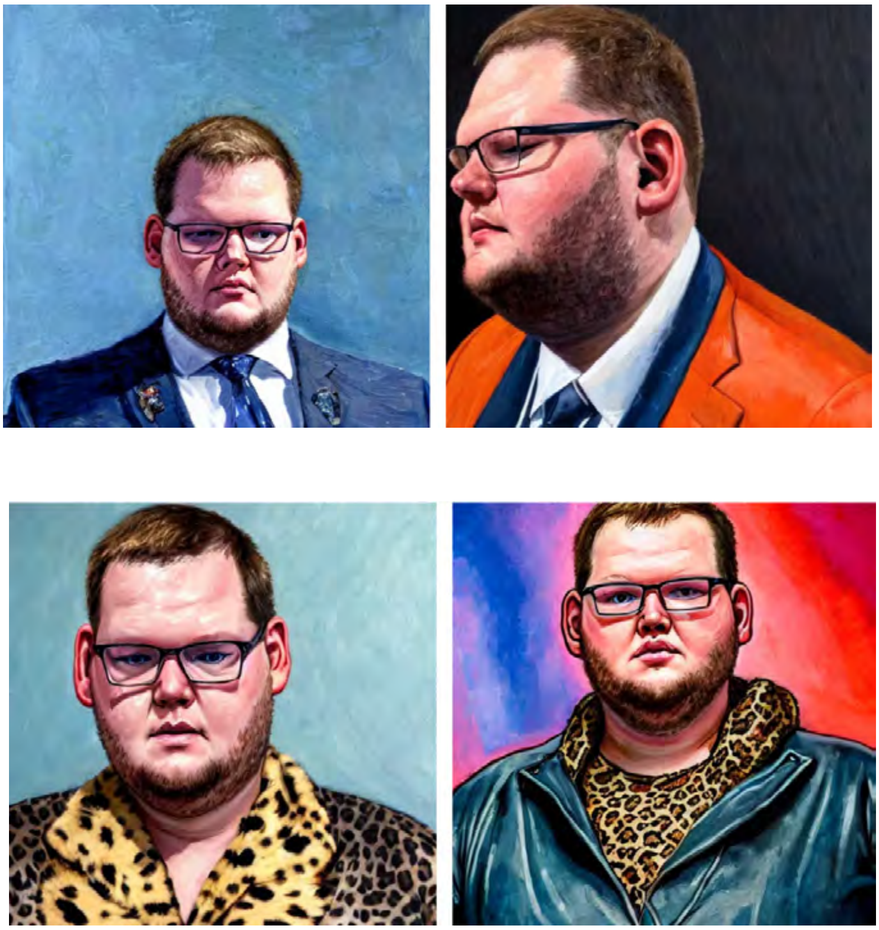


Figure 5.3: Artistic self-portraits using LoRA

5.2.3 Textual Inversion

We will not get too much into the details of the remaining two fine-tuning methods for Stable Diffusion since we did not find them fitting to our criteria. Textual Inversion is gaining popularity for its fine-tuning capabilities in the recent times. Using this method, one does not truly fine-tune the whole model but instead works at the shared latent space of the diffusion model. What it means is that, using Textual Inversion, latent embeddings of the training subject's prompt is sent to the latent space of the Diffusion Model which allows to create or produce images that align with the training subject since a more accurate image vector is chosen from the latent space due to fine tuning in the latent space. As a result, the underlying model or U-Net is not trained but the latent space is guided through the subject's unique keyword or embedded prompt. However, since textual inversion works in the latent space, it also requires absurdly high amount of computational resources which is not available for our research that is supposedly free of cost. Moreover, the outputs generated by using Textual Inversion are quite small in size which would reduce our demand of high quality and realism in the images. However, it might still have potential to accurately align the facial features of the prompt along with the generated images.

5.2.4 Hypernetwork

A company named Novel AI are the first to use Hypernetworks as a process to fine-tune Stable Diffusion checkpoints or models. In this process, a separate small neural network, which is known as the hypernetwork itself is used alongside the Stable Diffusion model to modify the generations and its styles based on a dataset. Similar to LoRA, this process modifies the U-Net of the Stable Diffusion and it basically includes its neural network to that of the U-Net of the Stable Diffusion. Specifically, it modifies the cross-attention denoising layers of the U-Net decoder. During training, it freezes the Stable Diffusion and only the small hypernetwork is trained. Since the neural network is small and linear, it takes a short amount of time to fine-tune using this process. Ultimately, combined with the U-Net of the Stable Diffusion model, it influences the backward diffusion process which is responsible for the image generation based on the given embedded prompts. The usage of Hypernetworks is relatively new and we have not explored its capabilities thoroughly for that reason. We chose to stick to already existing strong methodologies like DreamBooth or LoRA.

Feature	DreamBooth	LoRA	Textual Inversion	Hypernetworks
Target for fine-tuning	Entire U-NET	Compressed U-Net layers	Latent Space	A small neural network
Output Resolution	Same as input	Same as input	Smaller image	Same as inpur
Computational Costs	High	Moderate	Highest	Moderate

Table 5.3: Basic comparison between Fine-Tuning Models

5.3 DreamBooth Training

We are using an advanced custom implementation named FastDreamBooth. It has been already discussed in details in this paper in the section 5.2. In this section, we are going to discuss about the different hyperparameters set for training and some insights on what is being done. Firstly, DreamBooth pre-downloads all of the dependencies like libraries and modules required to run DreamBooth. Secondly, we get to choose any custom Stable Diffusion checkpoints which then downloads pre-trained base model that is going to be fine-tuned. Afterwards, we create a session for our fine-tuning in order to save the model in ckpt format after fine-tuning along with the instance images and external captions. DreamBooth saves the instance images and compresses them in a zip file to be used in training. Text files having the same file names as the instance images, contains the textual descriptions of the facial features of each and every image respectively. The text files are also zipped as captions.zip. Finally, the training process starts. We went through many trials and errors and tested out various values for the learning rates and learning steps for the U-Net. After many trials, we came to the conclusion that based on the relatively small size of our dataset, a learning rate of 2e-6 in 1200 steps were optimal for training the model on realistic faces. Many suggest more steps for larger dataset of images of real human faces which needs to be tested in our case as well. We implemented offset noise on our fine-tuning which is perfect for maintaining a similar style in the generations after fine-tuning. Since all of our images are candid portraits focusing on the actual face of the people, following that style in generations and leaving out intricate details for backgrounds are quite helpful. Training the text encoder is considered optional to many but since diffusion models lack appropriate detailed prompts for facial features, we also added learning rates and steps to the text encoder. 500 Steps are more than enough for very small datasets. For our dataset we chose 800 steps and the learning rate was set to 1e-6. The text encoder apparently overfits training data easily which would ultimately lose accuracy rather than increasing accuracy in aligning the images and the texts. As a result, we tried to stay with lower learning steps for the text encoder. Surprisingly, the training takes about 40 minutes while zipping the instance images and captions takes longer time for DreamBooth. For every step the loss was closely monitored to see whether there is a linear decrease/increase in values or whether there are absurd spikes. Generally, a non-linear curve might indicate towards the necessity to fix the hyperparameters and the fact that the learning was inefficient. However, in our case, the values of loss were different in each and every step and have random sparks instead of a linear decrease. But, we observed similar values in other Stable Diffusion fine-tuning and in evaluation our model still performed much better than the base 1.5.

5.4 Challenges in Training

Based on our research objective, the best way to implement our ideas would be to train a diffusion model like in our case, Stable Diffusion, on a dataset that almost objectively describes the facial features of the faces of the people in each and every image. Moreover, the dataset would have to be much larger as well. Unfortunately, we already know that training Diffusion Models is so costly that only large companies like Google or OpenAI can keep up with it. Even Stable Diffusion by Stability AI requires rigorous amounts of computational units to train from scratch which cost millions of dollars. Additionally, it is also not possible to change any internal components for our base Stable Diffusion Version 1.5 model. Because to make any core changes we would require to retrain the whole stable diffusion model. Attempts were made to replace the text encoder of CLIP for a newer and better textual transformer but as we said, it requires retraining the whole stable diffusion model. However, one recent research showed positive signs on using a very small LoRA checkpoint where they changed the transformer with BanglaBERT [18]. But it is to be remembered that one of our prime goals is to achieve high quality realism of the faces instead of cartoonish, creative outputs which would not be possible by a 64 X 64 resolution output. And without a small LoRA which can be treated as a small Stable Diffusion model, we cannot train on a dataset on consumer grade GPUs. Thus, we are aware of the fact that we could not make any massive changes to the core model that would result in something innovative. As a result, we focused this research on the capabilities of the base Stable Diffusion models with proper training only by tweaking the learning parameters of the U-NET as well as the text encoder. Lastly, the lack of computational resources yet again created challenges in active monitoring during training periods. It is quite common to test or check Generative AI models by saving checkpoints after ever n steps. After testing out the outputs we can identify whether the hyperparameters needs to be changed or not. Furthermore, lots of outputs needs to be generated to calculate metrics like FID score that are used to evaluate the model. Both of these are not possible due to the lack of proper computational resources as our free Colab runtime expires only after 4-5 generations or after one proper whole training. We shall still discuss about the immense variational amounts of future fields of research and project based on our very own fine tuned stable diffusion model.

Chapter 6

Results and Analysis

6.1 Model Training and Generation

6.1.1 Model Selection & Parameter Tuning

We chose Stable Diffusion v1.5 as our base model. Stable Diffusion is a completely open source model unlike other proprietary and closed-source diffusion models. The 1.5 version of Stable Diffusion model allows in depth fine-tuning on custom datasets without huge computational costs. Furthermore, the v1.5 model's utilization of VAE, transformers and samplers makes it very much viable to modify according to our needs. Later versions like v2 and SD_XL were introduced at the duration of our working progress, but they required more computational capabilities that were unavailable to us. Additionally, it is widely believed in the generative AI community that SD v2.0 lacks the accuracy in generating real human faces. On the other hand, SD_XL is currently commonly accepted as the model to achieve the most variance and vibrance in styles of image generations. It would have been our optimal choice for achieving our goals. But as we have already mentioned, it requires high amounts of computational resources to fine-tune on and modify according to our needs. Altogether, v1.5 as a base model provides us a well capable and versatile platform for realistic face generation.

The base model went through iterative exploration of different training parameters both for the model and the text encoder to optimize its performance. Different parameters were adjusted and tested to achieve a balance between image quality, perceived image realism and training efficiency while maintaining optimal computational costs. This iterative process of trying different combinations of parameter values helped us to refine the real face generation capabilities of our base model.

6.2 Image Generation and Evaluation

6.2.1 Models under Evaluation

To evaluate the performance of the generated images using our proposed model, we also generated images using various models, including DALL.E 3, SD_v2, SD_XL, and SD_v1.5. Each model represented their unique approach, styles and results of image generation, offering us a set of diverse images to do a comprehensive analysis of our results. 10 Pictures were generated for each model to quantitatively and qualitatively analyze the quality and realism of our generated images.

6.2.2 Quantitative Metrics

Several Quantitative Metrics, such as, Kernel Inception Distance (KID), Learned Perceptual Image Patch Similarity (LPIPS), and Structural Similarity Index (SSIM), were considered to bring objectivity to our analysis to assess the realism, accurate facial feature generation and quality of our generated images.

- **Kernel Inception Distance (KID):**
  KID is a commonly used metric which is used to assess the quality and realism of generated images. By utilizing an Inception Network to extract features from both real and generated images, it calculates how similar the feature distributions are to one another.

  $$KID = MMD(f_{real}, f_{fake})^2 \tag{6.1}$$

  Here, MMD is the maximum mean discrepancy and f_real, f_fake are the features that are extracted from real and generated images respectively. Furthermore,

  $$k(x, y) = (\gamma * x^T y + coef)^{degree} \tag{6.2}$$

  Here, a polynomial kernel of function K is used to calculate the MMD. The default feature extraction methods were used for KID (using pre-trained Inception V3 Network) and mini-batches of 3-channel RGB images of the form (3xHxW) were used as inputs. Images were also resized from 512 x 512 pixels to a size of 299 x 299 pixels to match the training data of the inception network. Features extraction of real images were done using our whole dataset of real images and 10 generated images of each diffusion model. Then, the KID scores were computed using Gram matrices. Higher similarity of distributions between real and generated images were reflected by lower KID scores. [2] [3]

- **Fréchet Inception Distance (FID):**
  Although FID is a similar metric to KID that is used to quantify the quality and realism of generated images and also a more common metric than KID, the latter was chosen as the better metric for our use case because FID is more computationally expensive than KID. Furthermore, FID assumes the distributions in images can be approximated as multivariate Gaussian, which might not be the case for generated images using latent diffusion models, such as our base model. This may indicate lower quality of images using diffusion models,

which is not always true. On the other hand, the distributional discrepancy in feature space between generated and real samples is directly measured by KID. Furthermore, Compared to FID, KID is frequently less sensitive to dataset properties like dataset size. When working with a small dataset like our generated images, this direct measurement would be helpful. Considering these and our hardware limitations, KID was a better fit to evaluate our model. [14] [19]

- **Inception Score (IS):**
  Another metric we considered but discarded was the Inception Score (IS). It is used to evaluate the quality and variety of generated images. But It has been criticized in generative AI literature for its lack of detailed interpretability regarding specific aspects of image quality, and is susceptible to the Inception model's unique architecture. [8]

- **Learned Perceptual Image Patch Similarity (LPIPS):**
  It is a metric that is used to calculate the perceptual similarity between 2 images as seen or perceived by real humans. For a pre-trained network, LPIPS basically calculates how similar two image patches' activations are to one another. Evidence suggests that this measure closely resembles human perception. In the fields of computer vision and machine learning, LPIPS is frequently utilized, particularly to evaluate the quality of images produced by models. LPIPS is a quantitative tool that researchers use to assess how well-generated images resemble the perceptual properties of real images. Furthermore, Researchers and developers can use the LPIPS metric because it is developed as an open-source library. It makes it simple for users to incorporate the assessment of perceptual similarity into their projects, such as ours. It also provides many different pre-trained models based on various convolutional neural network architectures. We chose AlexNet as it is the most widely used pre-trained model. A lower score when comparing two images using LPIPS suggests a higher degree of perceptual resemblance. We took a subset of 10 real images from our real face images dataset, and used their facial descriptions as prompts to generate 10 images using all the mentioned models and compared the LPIPS score for each corresponding pair. Then we calculated the average score for all the models. [26]

- **Structural Similarity Index (SSIM):**
  It is another popular metric that is widely accepted for assessing the quality of generated images, particularly those produced by models such as stable diffusion models. The structural similarity between the reference image and the produced image is measured using SSIM. Similarly to the utilization of LPIPS, we took the same subset of real images and generated their corresponding images and calculated the average SSI score of each model from their respective image pair scores.[1]

6.2.3 Quantitative Results

Model	**KID**(Lower Better)	**LPIPS**(Lower Better)	**SSIM**(Higher Better)
DALL.E_3	0.013569	**0.6938**	0.2626
SD_V2	0.0099369	0.7635	0.2100
Custom	**0.0097819**	**0.7089**	**0.2816**
SD_XL	0.010382	0.7753	**0.3283**
SD_V1.5	0.011132	0.7746	0.2557

Table 6.1: Score of different models

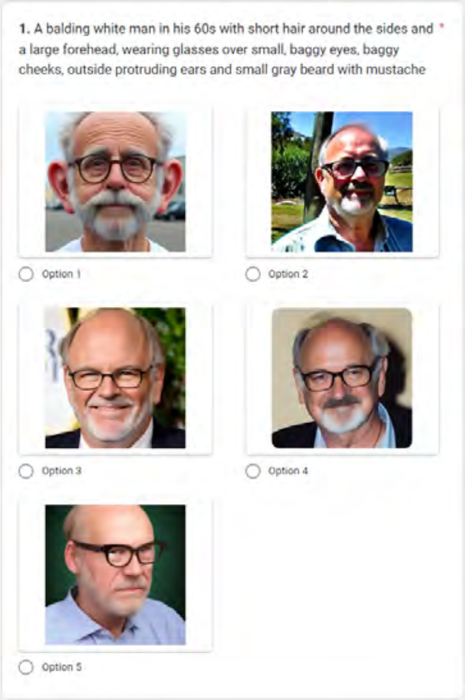
Here, it is clear that our proposed custom model performs really well in terms of both KID and LPIPS score. It has the lowest value in KID scores, and 2nd lowest value in LPIPS score. DALL.E_3 beat our custom model by a very slight margin in LPIPS score, while the other 3 models had way higher scores in this regard. Although there is more variance in the SSIM scores, our custom model achieves the spot for the 2nd best performer among all the models.

So, we can infer that our custom model consistently performs well across all metrics, showcasing its effectiveness in generating realistic faces. While "SD_V2" is a strong contender, particularly in KID and LPIPS metrics, and "SD_XL" excels in SSIM, suggesting superior preservation of structural details, we can further validate our quantitative metric results using qualitative analysis.

6.2.4 Qualitative Analysis

In addition to the quantitative evaluation, a survey with about 180 respondents was carried out. The task required of the participants was to assess pictures produced by various models. The poll results indicated a notable inclination for the visuals produced by the customized model. These qualitative findings provide insightful user opinions on the generated faces' perceived realism and aesthetic attractiveness.

The survey participants were asked to choose the best picture that matches with the given facial description while also being the most realistic.



1. A balding white man in his 60s with short hair around the sides and * a large forehead, wearing glasses over small, baggy eyes, baggy cheeks, outside protruding ears and small gray beard with mustache

○ Option 1

○ Option 2

○ Option 3

○ Option 4

○ Option 5

Figure 6.1: Sample Survey

The collected data from the survey are represented in a table below:

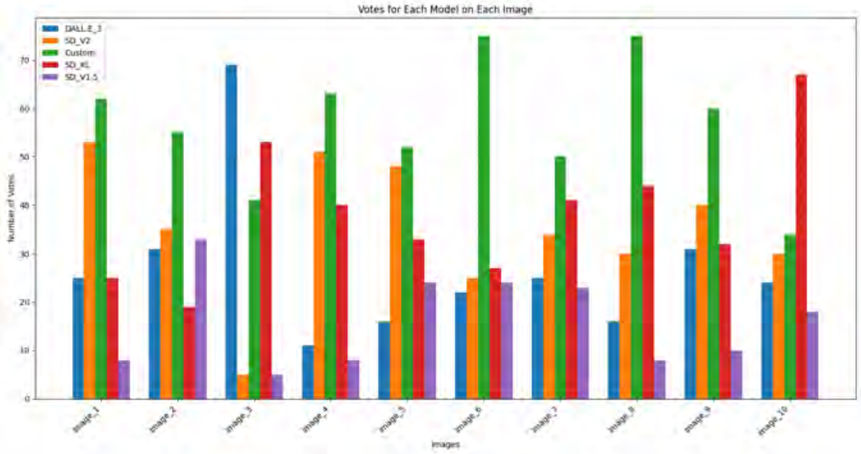


Figure 6.2: Votes for each model on each image

Here, we have visualized the data collected through our conducted survey. We can see the significant preference for images generated by our custom model. This signifies that our model fairs way better in real human perception than other stable diffusion models.

From Figure 6.2, 8 out of the 10 images our survey was conducted upon, the custom model received the highest number of votes. In images no. 4, 6, 8 and 9, the custom model wins in human perception by a long way. For image 3, Dall.E 3 and for image 10, SD_XL won the votes by a huge margin. This can be correlated with the highest SSIM score for SD_XL and lowest LPIPS score for DALL.E_3. With more generated images this can be validated further in the future.

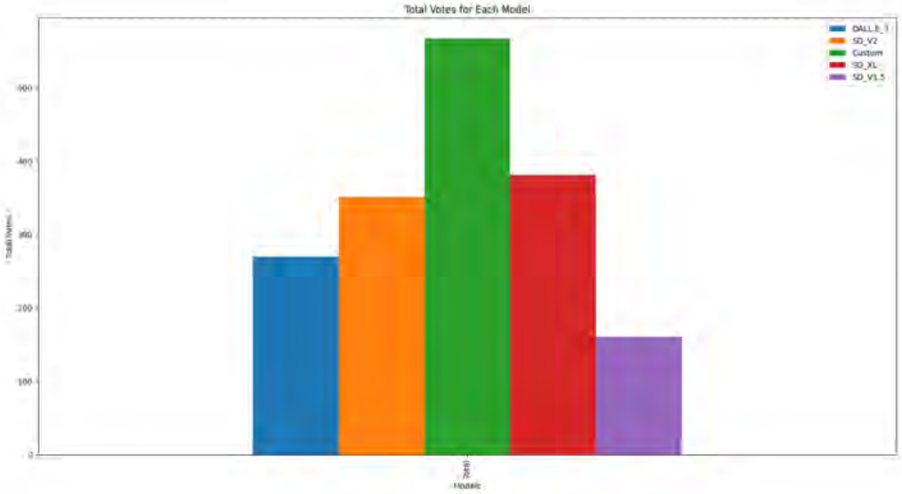


Figure 6.3: Total votes for each model

From Figure 6.3, we can see that our custom model garnered the highest total number of votes from the real human perception survey. It is important to note that SD_v1.5 had the lowest number of votes by a huge margin compared to all the other models. And, our custom model was based on this very model and got the best results.

Chapter 7

Conclusion

7.1 Conclusion

To sum everything up, a stable diffusion model is so far the best approach for face illustration which beats every other model on its way. We can provide adequate services by training our model to be efficient at face portrayal, memorization and facial recognition. One of the most challenging parts of this extrapolate was data collection. More extensive and diverse data collection can help to improve the representativeness and generalizability of the model, which will provide more accurate results and reduce potential bias. So far we were able to extract relevant data from our datasets and modify them to fit according to our own needs. In Spite of many rough patches, in terms of comparison, we gained remarkable results in the field of text to image generation which leads us to an immense opportunity in the upcoming near future.

7.2 Future Scope

While our custom trained stable diffusion model exhibits condescending performance compared to other existing models, it is essential to acknowledge certain inherent limitations and biases deriving from the training dataset. Notably, our own model exhibits stereotyping biases, especially evident when identifying certain attributes. For example, the term “American” is likely to conjure up images of white individuals, or the term “Asian” tends to evoke images of people with thin noses and lips, highlighting the dataset’s lack of diversity. Furthermore, the dataset’s restraints pose issues in handling particular scenarios. For instance, our model really struggles to perform accurate representations of people wearing sunglasses on their head rather than over their eyes. This issue arises from an indefinite imbalance in the training dataset where images of people wearing sunglasses on the head are enormously outnumbered by those with sunglasses over the eyes. Besides these, another important issue we faced while using OpenAI’s CLIP which is a neural network that effectively learns visual concepts from natural language supervision. [25] The maximum word length of CLIP’s is 77 though some of the new models support up to a higher number. Also, CLIP really struggles with more systematic or abstract works like counting the number of objects in an image. To address this problem, we propose to explore alternatives such as Visual-Language Transformer (ViLT) which might offer better capabilities beyond CLIP. Lastly, with a goal to refine our existing model for better performance than present, we can initiate an avenue for new research involving the reverse process. Which means, currently, our model generates accurate images based on the given facial attributes as input, a complementary model could be introduced to analyze facial images as input and accurately extract facial attributes as output. This bidirectional approach would contribute to a more holistic approach for understanding of facial synthesis. In a nutshell, though our custom model excels in many aspects, there are so many options for advancement in the field of realistic face generation.

References

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. “Image quality assessment: From error visibility to structural similarity.” (Apr. 2004), [Online]. Available: https://ieeexplore.ieee.org/document/1284395.

[2] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *arXiv (Cornell University)*, vol. 30, pp. 6626–6637, 2017. [Online]. Available: https://arxiv.org/pdf/1706.08500.

[3] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying mmd gans,” *arXiv (Cornell University)*, 2018. [Online]. Available: https://doi.org/10.48550/arxiv.1801.01401.

[4] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[5] D. P. Kingma, T. Salimans, B. Poole, and J. Ho, “Variational diffusion models,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 696–21 707, 2021.

[6] A. Nichol, P. Dhariwal, A. Ramesh, *et al.*, “Glide: Towards photorealistic image generation and editing with text-guided diffusion models,” *arXiv preprint arXiv:2112.10741*, 2021.

[7] A. Ramesh, M. Pavlov, G. Goh, *et al.*, “Zero-shot text-to-image generation,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 8821–8831.

[8] E. Betzalel, C. Penso, A. Navon, and E. Fetaya, “A study on the evaluation of generative models,” Jun. 2022. arXiv: arXiv:2206.10935v1 `[cs.LG]`. [Online]. Available: https://arxiv.org/abs/2206.10935v1.

[9] S. Karagiannakos and N. Adaloglou. “How diffusion models work: The math from scratch.” (Sep. 2022), [Online]. Available: https://theaisummer.com/diffusion-models/.

[10] R. O’Connor. “Stable diffusion 1 vs 2 - what you need to know.” Developer Educator at AssemblyAI. (Dec. 2022), [Online]. Available: https://www.assemblyai.com/blog/stable-diffusion-1-vs-2-what-you-need-to-know/?fbclid=IwAR0mjH8zqWuuOM0jqHEaxLl9ZLN2nTnO5lffTTyjW8u4w5vOWkPSw6UmflE.

[11] C. Saharia, W. Chan, S. Saxena, *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[12] A. Borji, "Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney, and dall-e 2," *arXiv preprint arXiv:2210.00586*, 2023.

[13] B. Kawar, S. Zada, O. Lang, *et al.*, "Imagic: Text-based real image editing with diffusion models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2023, pp. 6007–6017. DOI: 10.1109/CVPR52729.2023.00582. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00582.

[14] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, "The role of imagenet classes in fréchet inception distance," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Published as a conference paper at ICLR 2023, 2023. [Online]. Available: https://arxiv.org/abs/2203.06026v3.

[15] Z. Li, Q. Zhou, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Guiding text-to-image diffusion model towards grounded generation," *arXiv preprint arXiv:2301.05221*, 2023.

[16] D. McSwine, "Diffusion models: A comprehensive survey of methods and applications," *ACM Digital Library*, Nov. 2023, Retrieved January 9, 2024, from https://dl.acm.org/doi/abs/10.1145/3626235. DOI: 10.1145/3626235.

[17] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine-tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22 500–22 510.

[18] A. K. Saha, N. M. K. Arnob, N. N. Rahman, M. Haque, S. M. R. Al Masud, and R. Rahman, "Mukh-oboyob: Stable diffusion and banglabert enhanced bangla text-to-face synthesis," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 11, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01411142.

[19] C. J. Wang and P. Golland, "Interpolating between images with diffusion models," *arXiv (Cornell University)*, 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2307.12560.

[20] Wikipedia contributors. "DreamBooth." (2023), [Online]. Available: https://en.wikipedia.org/wiki/DreamBooth.

[21] J. Zhu, J. Jin, Z. Yang, X. Wu, and X. Wang, "Learning clip guided visual-text fusion transformer for video-based pedestrian attribute recognition," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 2626–2629. DOI: 10.1109/CVPRW59228.2023.00261.

[22] Z. Deng *et al.*, "Text to image generation with conformer-gan," in *Neural Information Processing. ICONIP 2023.*, B. Luo, L. Cheng, Z. Wu, H. Li, and C. Li, Eds., ser. Lecture Notes in Computer Science, vol. 14451, Springer, Singapore, 2024. DOI: 10.1007/978-981-99-8073-4_1.

[23] N. Klingler. "Clip: Contrastive language-image pre-training (2024)." (2024), [Online]. Available: https://viso.ai/deep-learning/clip-machine-learning/.

[24] Wikipedia contributors. “OpenAI.” (2024), [Online]. Available: https://en.wikipedia.org/wiki/OpenAI.

[25] OpenAI, *CLIP: Connecting text and images*, https://openai.com/research/clip, 2021, January 5.

[26] PyTorch-Metrics. “Learned perceptual image patch similarity (lpips) — pytorch-metrics 1.2.1 documentation.” (n.d.), [Online]. Available: https://lightning.ai/docs/torchmetrics/stable/image/learned_perceptual_image_patch_similarity.html#torchmetrics.image.lpip.LearnedPerceptualImagePatchSimilarity.