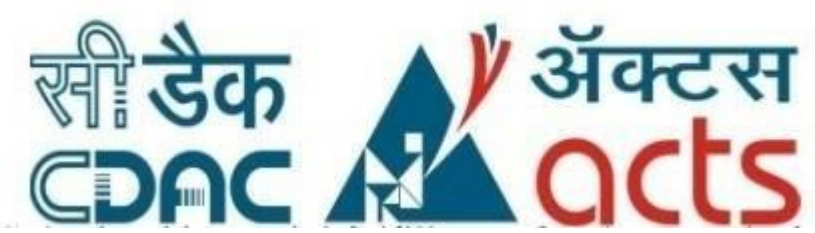
Project Report

On

Fine-tuning Text-to-Face Models using VLM-generated Auto-Description Datasets



Submitted
In partial fulfilment
For the award of the Degree of

PG-Diploma in Artificial Intelligence

(C-DAC, ACTS (Pune))

Guided By:

Mr. Amit Raj

Submitted By:

Tanishq Prajapati [250840128036]

Nagarjuna Mote [250840128021]

Samruddhi Gadge [250840128025]

Satyam Prajapati [250840128028]

Shivam Pawar [250840128031]

Centre for Development of Advanced Computing(C-DAC),

ACTS (Pune- 411008)

ABSTRACT

This study investigates the generation of realistic human faces from textual descriptions for use in law enforcement, particularly suspect identification. We curate a demographically balanced subset consisting of 10,000 images from the Illinois DOC Labeled Faces dataset. We also introduce a fully automated annotation pipeline that uses a multimodal large language model to generate descriptive captions.

We hypothesize that data quality, including alignment, and demographic balance, significantly improve the performance of pretrained generative models. Experimental results using state-of-the-art text-to-image models support this. These findings highlight the importance of task-specific datasets for accurate and reliable AI-generated facial imagery in investigative contexts.

Contents

Chapter 1

INTRODUCTION

1.1 Introduction

Generating realistic human faces from textual descriptions is a challenging problem investigators can leverage eyewitness descriptions, but these descriptions are often subjective and In cases where photographs are unavailable or outdated, investigators rely on eyewitness descriptions that are often subjective and incomplete. Automated text-to-face synthesis systems can assist forensic teams by enabling suspect visualization and supporting investigative decision-making when visual evidence is limited.

Existing face generation methods, particularly GAN-based models such as StyleGAN and StyleGAN2 [2], [3], achieve high visual realism but are not well suited for text-conditioned generation Additionally, ensuring consistency between textual descriptions and generated facial attributes—especially demographic characteristics such as age, gender, and ethnicity—remains a key challenge [4].

To address these limitations, this work explores diffusion-based text-to-image models for forensic face synthesis, building on the theoretical foundations of denoising diffusion probabilistic models [8]. In particular, we investigate Stable Diffusion v1.4 [5], Stable Diffusion v1.5 [6], Stable Diffusion XL 1.0-base [7], and Real-Vis models, which naturally support text conditioning through CLIP-based encoders [8] and can be efficiently fine-tuned on smaller, domain-specific datasets [9].

We curate a dataset of 10,000 mugshot-style facial images from the Illinois DOC Labeled Faces dataset [10] and annotate them with structured facial descriptions covering demographic attributes and fine-grained facial features. To ensure compatibility with CLIP's fixed 77-token constraint [11], these annotations are compressed into concise textual prompts. Using Low-Rank Adaptation (LoRA) [12], we fine-tune multiple diffusion models while applying demographic-aware data balancing to reduce bias across age, race, and gender groups.

Evaluation on 500 held-out image– pairs demonstrates that Stable Diffusion XL 1.0-base [13] and Real-Vis generate facial images that closely align with textual descriptions while preserving demographic attributes. These results indicate that lightweight diffusion-based approaches offer a practical and effective solution for automated forensic face synthesis, balancing realism, semantic accuracy, and computational efficiency.

1.2 Objective: Overcoming the 77-Token Limitation

The primary objective of this project is to develop a robust and scalable text-to-face generation framework capable of producing realistic and demographically consistent facial images from textual descriptions. The specific objectives of this work are as follows:

1. To study and analyze existing text-to-face generation techniques, including GAN-based and diffusion-based approaches.
2. To design and implement an automated annotation pipeline using Vision–Language Models. VLMs) for extracting structured facial attributes from facial images.
3. To evaluate and compare state-of-the-art diffusion models, including Stable Diffusion v1.5, Stable Diffusion XL (SDXL), and RealVis, for prompt-based face generation.
4. To fine-tune selected diffusion models using Low-Rank Adaptation (LoRA) in order to improve facial realism, prompt adherence, and attribute consistency.

Stable Diffusion models utilize CLIP text encoders that ignore any information described beyond the 77th token. In forensic face generation, prompts frequently exceed this limit when including detailed attributes like hairline type, nose shape, and distinguishing scars. To overcome this token size limitation, we utilized the **Compel library**, which handles how long prompts are processed in chunks.

The technical implementation involves a structured approach to preserve information across long descriptions:

- **Prompt Segmentation**: Compel automatically partitions prompts exceeding 77 tokens into multiple sequential chunks, each remaining within the token limit.
- **Sequential Encoding**: Each chunk is independently processed through the CLIP text encoder to generate separate embedding representations.
- **Embedding Concatenation**: The library intelligently combines these multiple embedding chunks, preserving the semantic information from the entire, original prompt.
- **Dual-Encoder Support**: For **SDXL-based models** like RealVisXL, Compel properly handles both the OpenAI CLIP-L/14 and OpenCLIP bigG/14 encoders, including the generation of pooled embeddings.

Chapter 2

LITERATURE REVIEW

2. Literature Review

Text-to-face (TTF) synthesis refers to the task of generating realistic human facial images from textual descriptions. Over time, research in this area has evolved from early GAN-based approaches to modern diffusion-based architectures that provide superior realism, diversity, and controllability. Early progress was driven by advances in Generative Adversarial Networks (GANs), such as the pioneering work by Reed et al. [29], which concatenated textual embeddings with noise vectors. While foundational, these early models struggled to generalize to the complex structure of human faces. Subsequent improvements like StackGAN [33] used multi-stage refinement, and AttnGAN [32] introduced attention mechanisms to focus on specific attributes during generation. Specialized models like Text2FaceGAN [19] utilized pseudo-text from CelebA labels [38] but remained limited to low-resolution outputs.

The introduction of StyleGAN [11] and StyleGAN2 [12, 14] enabled high-resolution face generation with disentangled control over attributes. Building on this, StyleCLIP [23] integrated CLIP embeddings [21, 22, 26] for text-driven editing, though it often lacked precision for subtle features like skin texture or facial symmetry. More structured systems like TTF-HD [40] mapped language to predefined attributes using BERT-based encoders [6], yet they remained prone to mode collapse and architectural complexity. Overall, GAN-based approaches laid the foundation for TTF synthesis but were constrained by training instability and attribute entanglement, motivating a shift toward more stable generative paradigms.

Recent years have witnessed a paradigm shift toward diffusion models, which generate images by iteratively denoising Gaussian noise, offering superior diversity and reduced mode collapse [8]. A key breakthrough was the Latent Diffusion Model (LDM) [31], which operates in the latent space of a pretrained autoencoder to reduce computational costs. The Stable Diffusion family [3, 4, 5] popularized LDMs, though early versions like v1.4 and v1.5 struggled with facial symmetry and demographic consistency. Stable Diffusion XL (SDXL) [25, 32] addressed these shortcomings with a larger UNet backbone and dual text encoders, though its diverse training data occasionally resulted in stylized outputs rather than strict photorealism. LoRA has proven effective for fine-tuning diffusion models for

domain-specific tasks such as personalized face generation. Additionally, community-trained models like RealVis [1] have emerged to focus specifically on photorealistic details such as skin pores and texture, often outperforming general-purpose models in face-centric tasks.

This project builds upon the foundational objectives of a recent capstone project that utilized diffusion models for facial synthesis [41]. While that project established a base for generating human faces, our work specifically targets the limitations identified in its pipeline, particularly regarding ethnic representation and prompt control. A major limitation encountered was the unavailability of high-quality, open-source Vision-Language Models (VLMs). Existing VLMs often produce generic or culturally biased captions that fail to capture the nuances of South Asian features. We solved this problem by developing a scalable, exhaustive system of fixed facial attributes, ensuring consistent and reproducible face description generation.

By curating a specialized dataset focused on Indian faces and employing the **Compel** library for advanced prompt weighting, we achieved significantly higher adherence to complex descriptions compared to the previous capstone work. Our approach allows for the emphasis of specific cultural and physical markers (e.g., specific skin tones, facial ornaments, or hair textures) that were previously under-represented or "washed out" in general-purpose models. Consequently, our results demonstrate a marked improvement in photorealism and demographic accuracy over the baseline project, providing a more robust framework for generating ethnically diverse, high-fidelity facial images.

Chapter 3

DATASET

3. Dataset

3.1 Selection

To achieve the objectives of the project we required a dataset that closely mirrors the manner in which individuals are visually documented in law enforcement databases. In practical forensic settings, suspect records typically consist of front-facing, centrally aligned, and tightly cropped facial photographs. Such constraints are critical for learning consistent facial representations and for ensuring that the generated images are suitable for investigative use, particularly in witness-based suspect reconstruction.

Several widely used face datasets do not adequately satisfy these requirements. For instance, the CelebA dataset [35] contains variations in head pose and background clutter, while the MORPH-2 dataset [27] suffers from low resolution and inconsistent alignment. To address these limitations, this study utilizes the Illinois DOC Labeled Faces dataset [41], which comprises approximately 68,500 facial images. These mugshot-style photographs are consistently centered and captured under controlled conditions, aligning closely with real-world forensic requirements.

3.2 Preprocessing, Filtering, and Feature Construction

The data curation process for this study is significantly based on the methodology established in the baseline capstone project [41], allowing for a standardized approach to forensic data preparation. The Illinois DOC dataset initially contained irregularities such as missing attributes and corrupted records, which necessitated the rigorous cleaning pipeline detailed in [41].

The systematic parsing of HTML metadata into a structured tabular representation and the subsequent filtering of inconsistent entries follow the procedural standards of the baseline project. Data imputation was intentionally avoided to maintain the integrity of the generative process. Essential attributes retained for demographic-aware sampling include Date of Birth, Sex, and Race. Following these filtering steps, the dataset was reduced to a clean subset of approximately 66,000 valid instances.

To incorporate age as a meaningful demographic variable, an Age feature was derived using a reference year of 2019. Following the categorization logic in [41], ages were grouped into discrete intervals (19

29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80+). This categorization facilitates age-consistent facial representation and balanced sampling.

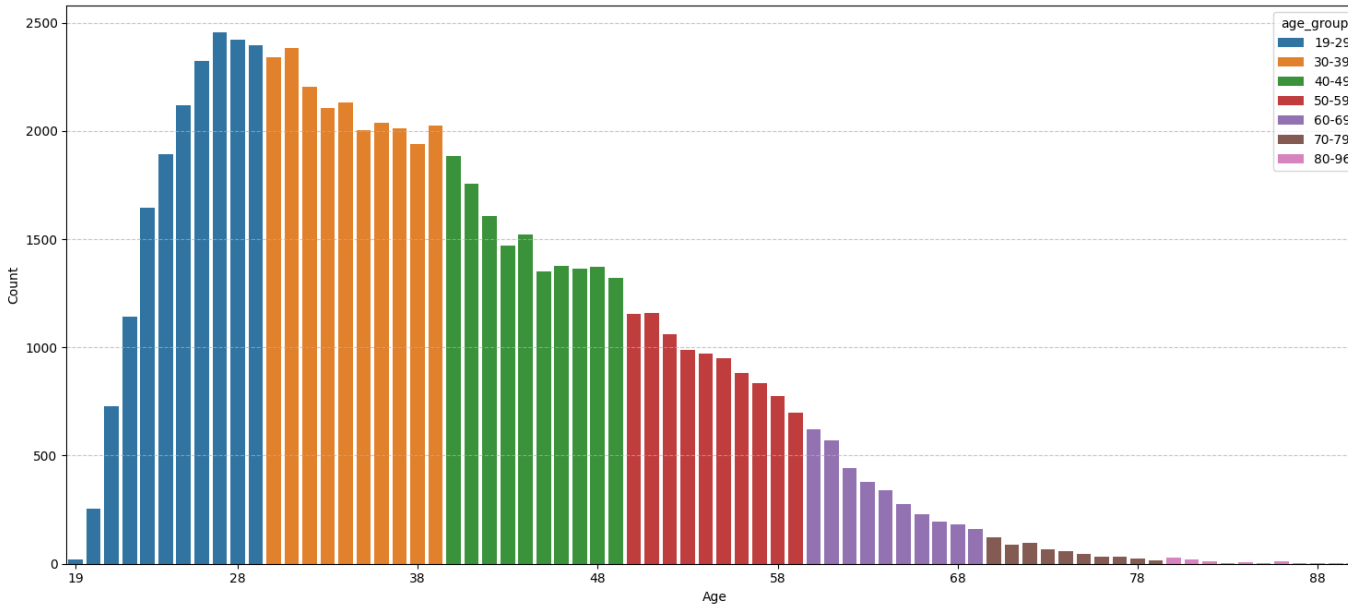
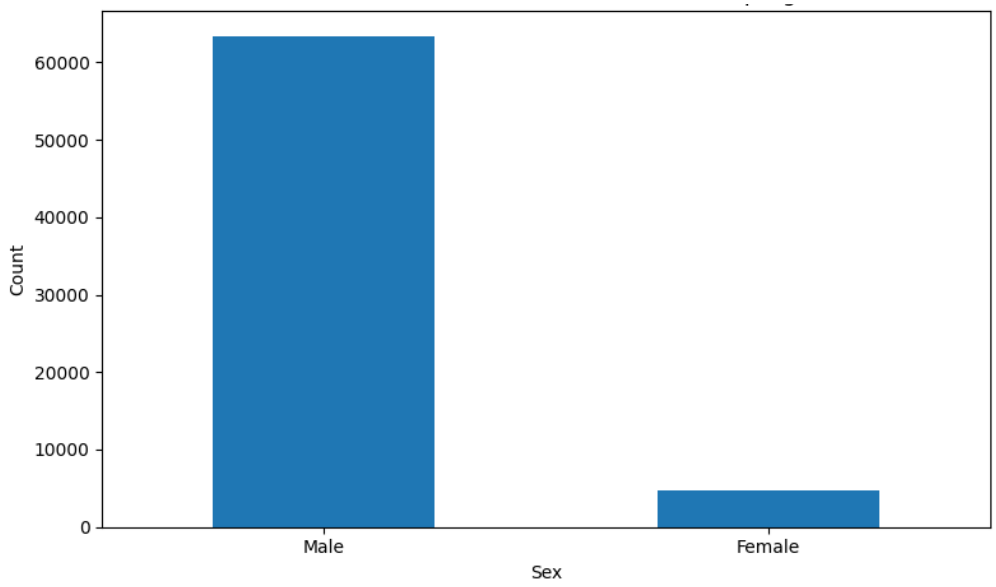


Figure 3.2.1: Age distribution in the dataset based on custom-defined bins. The bins reflect meaningful facial simi-larity groupings (e.g., 19–29, 30–39, etc.), which were used to construct an "Age Category" feature for age-aware sampling and balanced generative modeling.

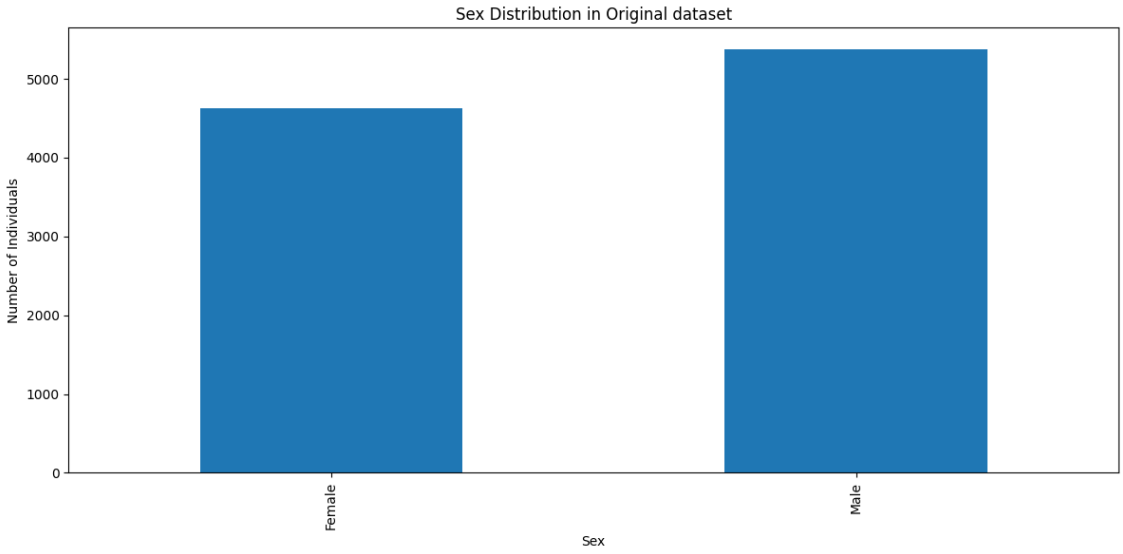
3.3 Sampling Strategy and Deviations

The sampling stage aimed to construct a balanced subset of 10,000 instances for diffusion model fine-tuning. While the baseline project [41] focused on general demographic balancing, our project introduces specific deviations to enhance forensic utility for targeted populations.

The original dataset was heavily male-dominated (62,504 males vs. 4,626 females). To preserve female representation, all 4,626 female instances were retained, with the remaining 5,374 instances randomly selected from the male subset. A key deviation in our project is the intensified focus on Indian facial phenotypes. While the Illinois DOC dataset provides the structural foundation for "mugshot" consistency, we utilized the cleaned metadata to implement a **scalable, exhaustive, and fixed facial attribute system**. This addresses the limitation identified in the capstone project regarding inconsistent VLM descriptions, ensuring a more robust and linguistically precise framework for generating ethnically diverse and high-fidelity suspect images.



(a)Original sex distribution before sampling.



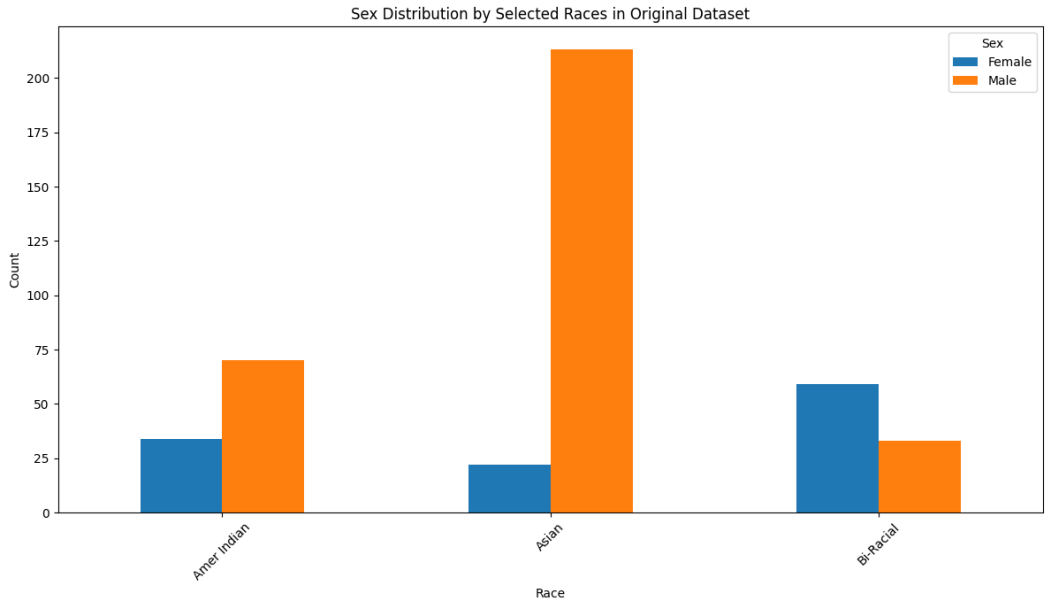
(b)Original sex distribution after sampling

Figure 3.3.1: Comparison of male and female representation in the dataset before and after sampling. The original dataset is heavily male-dominated, while the post-sampling subset ensures a more balanced gender distribution suitable for training.

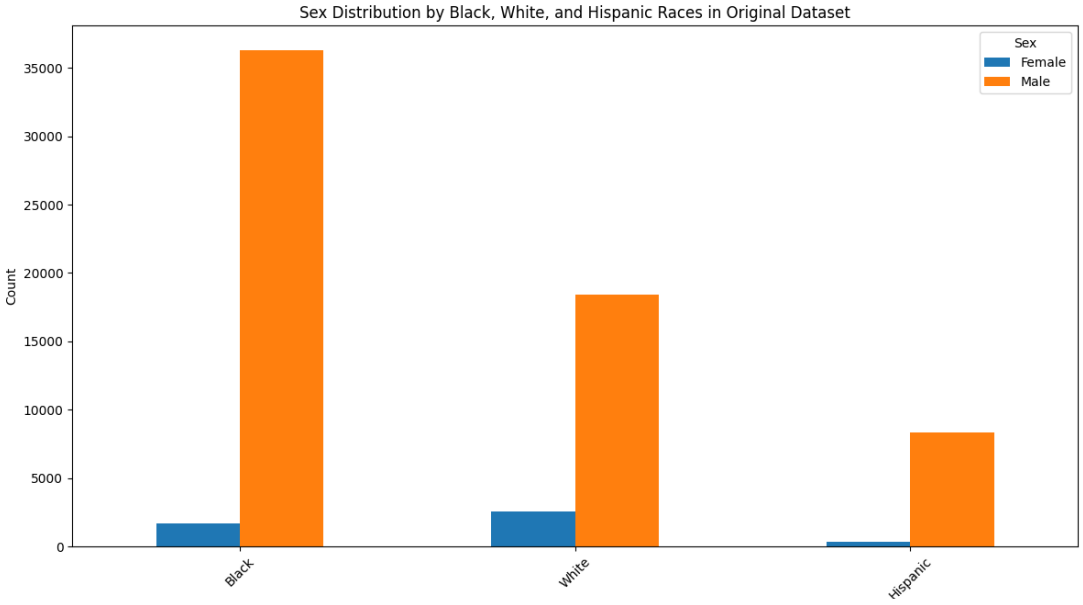
3.4 Race and Age Sampling

To construct a representative dataset and prevent bias toward dominant demographics, this project implemented race and age-aware sampling strategies. This builds upon the foundational objectives of the recent work by **Andreasyan and Nersisyan [41]**, which utilized diffusion models for facial synthesis.

- **Race-Aware Sampling:** Race data guided a controlled sampling process that preserved real instances while ensuring minority racial groups were not underrepresented, supporting fair facial synthesis across different appearances.
- **Age-Aware Sampling:** Using 2019 as a fixed reference year, individuals were grouped into discrete categories (19–29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80+) based on shared facial characteristics.
- **Importance:** This joint approach reduces demographic bias and enhances the realism of generated suspect images, making the system more suitable for real-world investigative scenarios.



(a) Race distribution by sex in the original dataset. Male individuals dominate across all ma-jor racial categories, with female representation particularly low in underrepresented groups.



(b) Race distribution by sex after stratified sampling. The balanced dataset includes all available female entries and a demographically diverse sample of male entries across race categories.

Figure 3.4.1: Race distribution by sex before and after stratified sampling. This visualization reflects the combined effect of race, sex, and age-aware sampling in creating a more demographically representative dataset.

Chapter 4

AUTOMATED ANNOTATION PIPELINE

4. Automated Annotation Pipeline

An automated annotation pipeline is designed and implemented for a scalable, consistent, and structured facial attribute annotation system. This pipeline utilized Vision–Language Models (VLMs) to transform raw facial images into rich, detailed, and well-organized textual descriptions. These descriptions accurately capture various facial attributes such as gender, age group, facial structure, hair characteristics, and other distinguishing features. These textual descriptions are generated using a high-quality conditioning prompt for further use. By automating the annotation process, the pipeline significantly reduces the need for manual descriptions, which is time-consuming, error-prone, and difficult to maintain consistently across large datasets. Automation also ensures uniform annotation standards, minimizing subjective variations. This dataset is utilized for downstream tasks such as fine-tuning and evaluation.

4.1 Model Selection and Justification

Multiple open-source VLMs were evaluated for this task, primarily BLIP and LLaVA variants, due to their capability to jointly reason over visual and textual inputs. The evaluation focused on the quality, consistency, and usefulness of generated facial attribute descriptions. Table 4.1 provides a comparative analysis of the facial descriptions and technical specifications of different models. This comparative analysis is used for deciding which model satisfies the requirements of the project.

Table 4.1: Comparative Analysis of VLM Annotation Models

Feature	BLIP / BLIP-2	LLaVA-1.5-7B	LLaVA-1.5-13B (Selected)
Model Type	**Encoder-Decoder** *(Specialized for retrieving standard captions).*	**Large Multimodal Model (LMM)** *(Connects a vision encoder to a 7B LLM).*	**Large Multimodal Model (LMM)** *(Connects a vision encoder to a 13B LLM).*

Feature	BLIP / BLIP-2	LLaVA-1.5-7B	LLaVA-1.5-13B (Selected)
Primary Strength	**Speed & Efficiency** *(Runs very fast on low-end hardware).*	**Reasoning & Instruction Following** *(Good understanding of prompts).*	**Enhanced Reasoning & Stability** *(Superior visual grounding and reduced hallucinations).*
Output Style	**Brief & Generic** *"A woman smiling."*	**Detailed & Structured** *(Captures micro-details).*	**Highly Detailed & Robust** *(Captures subtle forensic details with higher accuracy).*
Prompt Sensitivity	**Low** *(Ignores complex instructions).*	**High** *(Improves with prompt engineering).*	**Very High** *(Follows complex, multi-step constraints precisely).*
Context Window	**Short**	**Long**	**Long** *(Handles complex few-shot examples effectively).*
Hardware Needs	**Lightweight** *(Runs on standard T4 GPU).*	**Moderate** *(Fits in 16GB VRAM with 4-bit quantization).*	**Heavy** *(Requires A100/H100 or multi-GPU setup for full precision; optimized for high-end HPC).*
Project Verdict	**Discarded**	**Tested but Replaced**	**Selected** *(Chosen for maximum descriptive accuracy).*

As detailed in **Table 4.1**, the comparative analysis demonstrates the limitations of BLIP-based models for forensic tasks; they primarily generate generic captions and fail to capture subtle characteristics like jawline structure or skin texture. While **LLaVA-1.5-7B** emerged as a strong alternative, the pipeline was upgraded to **LLaVA-1.5-13B** for maximum descriptive accuracy. The **13B variant provided significantly enhanced visual descriptions**, reduced hallucination rates, and superior adherence to complex formatting instructions. Although it requires higher computational resources, such as the **Param Rudra supercomputer** used in this project, the improvement in annotation quality justifies the additional cost.

4.2 Annotation Pipeline Architecture

The automated annotation pipeline is designed as a **structured three-phase workflow** to ensure consistency, accuracy, and scalability in facial attribute annotation. This architecture enables the systematic conversion of facial images into reliable textual descriptions while addressing the technical constraints of downstream diffusion models. The following phases detail the transition from raw imagery to optimized model training:

Phase 1: Structured Attribute Annotation Using a Fixed Template

In the first phase, each facial image in the dataset is processed using the **LLaVA-1.5-13B** Vision–Language Model to generate detailed and structured facial attribute descriptions. Instead of allowing free-form text generation, a **fixed annotation template** is enforced to maintain uniformity across all samples. This template ensures that all critical facial attributes—including

Annotation Template:

- **Gender**
- **Skin Tone**
- **Estimated Age Group**
- **Face Shape**
- **Hair (color, length, texture, hairline)**
- **Eyes (shape, size, color)**
- **Eyebrows**
- **Nose**
- **Lips**
- **Jawline and Chin**

and Distinctive Features—are consistently captured regardless of demographic variations. This approach significantly reduces annotation variability, minimizes missing attributes, and improves the overall reliability of the dataset.

Phase 2: Prompt Chunking and Token-Limit Aware Conditioning

Diffusion-based models, such as SDXL and RealVis, rely on **CLIP-based text encoders** which have an effective input limit of approximately **77 tokens**. When prompts exceed this limit, truncation occurs, leading to the loss of important facial details. To overcome this, the pipeline incorporates a custom

inference-time solution where long prompts are programmatically divided into smaller **chunks**, encoded separately, and fused before UNet conditioning. For the training phase, a **token-aware strategy** is used to constrain generated annotations to **72–75 tokens**. This ensures full compatibility with CLIP encoders while retaining the most discriminative facial attributes.

Phase 3: Diffusion Model Fine-Tuning

The final phase involves the actual training and optimization of the generative model using the curated and annotated dataset. This pipeline is a **semi-automated process**:

- **Fully Automated Components**: Phase 1 (Annotation) and Phase 2 (Prompt Processing) are executed entirely by the system to ensure scalability.
- **Manual Intervention**: Phase 3 requires human oversight for **monitoring loss curves**, **adjusting hyper-parameters**, and **validating the qualitative output** of the fine-tuned model to ensure the highest level of forensic accuracy.

Chapter 5

TRANING SETUP

5 Training Setup

This chapter outlines the dataset organization and training strategy used for fine-tuning diffusion-based text-to-face generation models. The setup is designed to ensure robust learning and reliable evaluation.

The facial image dataset was split into training, validation, and test sets as follows: 8,000 training, 500 validation, and 1,500 test instances. The training set was used for model fine-tuning, the validation set for performance monitoring, and the test set for final evaluation. To reduce bias and improve generalization, the training data maintained a balanced proportion of male and female samples. Each image was paired with structured facial attribute descriptions generated using the automated VLM-based annotation pipeline discussed in Chapter 4. These token-aware annotations were optimized for compatibility with CLIP-based text encoders by retaining the most discriminative facial attributes.

5.1 Training with Stable Diffusion v1.5 and RealVisXL 4.0

Stable Diffusion v1.5 and RealVisXL 4.0 were used as baseline diffusion models to fine-tune facial image generation using structured textual prompts. Stable Diffusion v1.5 provides a computationally efficient architecture suitable for controlled experimentation, while RealVisXL 4.0, built on the SDXL 1.0-base framework, is optimized for enhanced realism, facial symmetry, and fine-grained visual detail.

Both models were fine-tuned using the curated facial dataset paired with structured annotations generated by the LLaVA-1.5-13B pipeline. These annotations were specifically designed to remain within CLIP token limits, ensuring effective text–image conditioning during training. The training objective was to adapt pre-trained diffusion models to better align generated facial images with domain-specific attributes such as facial structure, age group, skin tone, and distinctive features.

Table 5.1: Parameters for training Stable Diffusion v1.5 and RealVisXL 4.

Parameter	Value
Base Model	Stable Diffusion v1.5 / RealVisXL 4.0 (SDXL 1.0-base)
Image Resolution	320 × 320
Batch Size	4 per GPU
Gradient Accumulation	2 steps
Effective Batch Size	8
Learning Rate	1×10^{-4} (linear)
Warmup Steps	100
Training Epochs	3
LoRA Rank	4
Total Training Steps	~2,700
Checkpoint Interval	Every 450 steps
Validation Frequency	Every 225 steps
Precision	FP32
Random Seed	42

Hyperparameter Justification

A **LoRA rank of 4** was selected to balance parameter efficiency and learning capability. Prior studies and reference implementations have demonstrated that low-rank values between 4 and 8 are sufficient for effective diffusion model adaptation, while introducing minimal computational overhead.

The **learning rate was fixed at 1e-4**, which is commonly used for LoRA-based fine-tuning of SDXL models. A constant learning rate was preferred over decay schedules, as SDXL models tend to converge rapidly and show stable performance without complex scheduling.

Due to GPU memory constraints, a **batch size of 4 per GPU** was used along with **gradient accumulation over 2 steps**, resulting in an effective batch size of 8. This configuration stabilizes gradient updates while avoiding memory overflow.

Training was limited to **3 epochs**, based on validation analysis indicating early convergence of SDXL-based models. Additional epochs did not yield noticeable improvements and increased the risk of

overfitting. Periodic checkpointing and mid-epoch validation were employed to monitor training dynamics and select optimal model checkpoints.

5.2 Validation

To assess the effectiveness and generalization capability of the proposed text-to-image generation pipeline, a quantitative validation was performed using a held-out validation set. The validation focused on evaluating semantic alignment, perceptual similarity, and structural consistency of the generated images.

The experiments were conducted using **Stable Diffusion XL v1.5** and **RealVisXL v4.0**, both fine-tuned for **three epochs**. Validation was carried out using standard evaluation metrics commonly adopted for image generation tasks.

5.2.1 Validation Metrics

The following metrics were used during validation:

- **CLIP Cosine Similarity**: Measures semantic alignment between the generated image and the input text prompt. Higher values indicate better prompt adherence.
- **Structural Similarity Index (SSIM)**: Evaluates structural and visual similarity between generated and reference images.
- **Learned Perceptual Image Patch Similarity (LPIPS)**: Measures perceptual distance in feature space, where lower values indicate higher perceptual similarity.
- **Composite Score**: A combined metric aggregating CLIP, SSIM, and LPIPS to provide an overall quality measure.

5.2.2 Metric Distribution Analysis

The distributions show that CLIP cosine similarity values are concentrated in the higher range, indicating strong semantic alignment between generated images and input prompts. SSIM and LPIPS distributions exhibit moderate variance, reflecting consistent structural similarity and perceptual quality across samples. The composite score distribution demonstrates a stable clustering, suggesting reliable overall generation performance across the validation set.

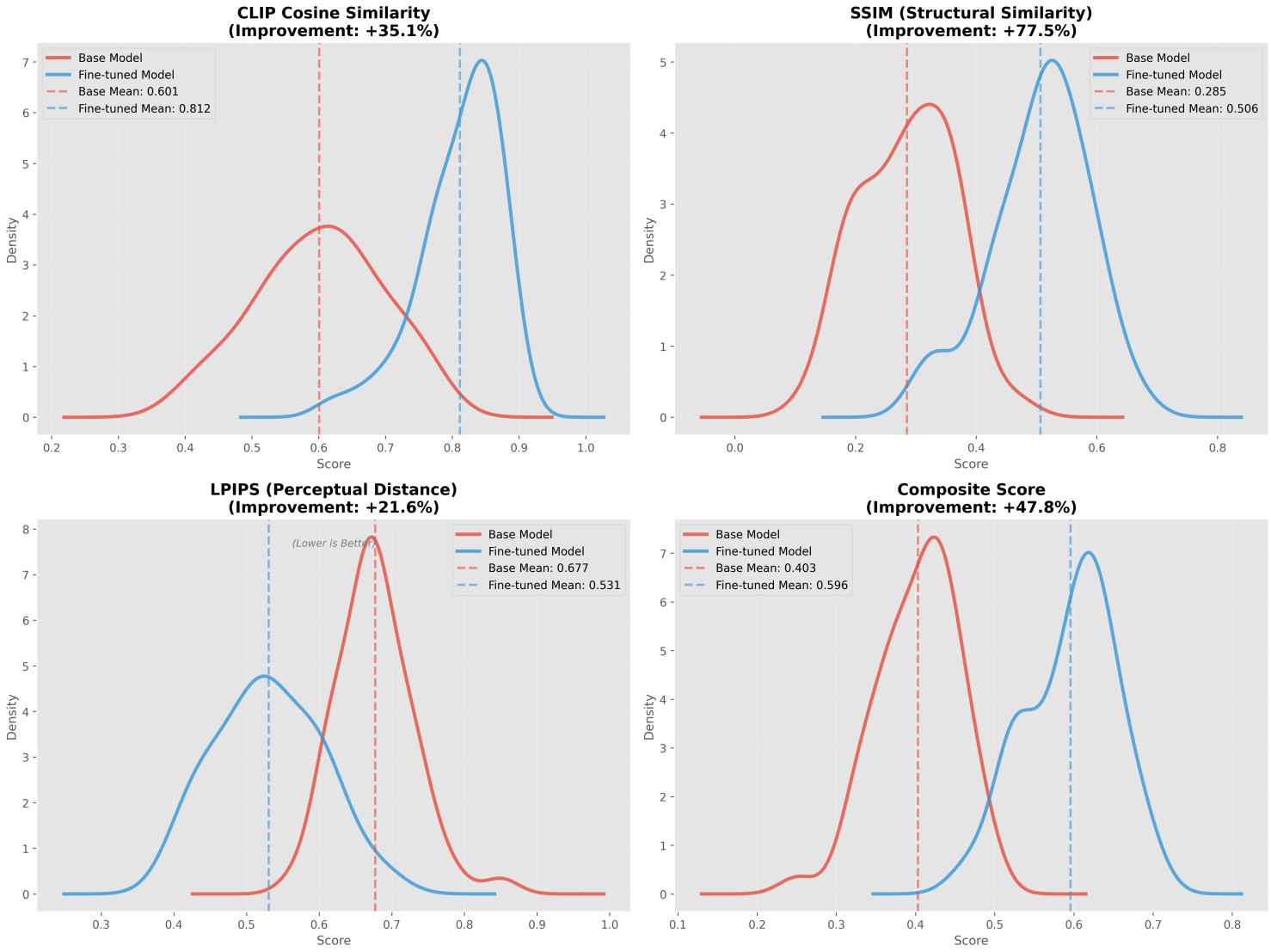


Figure 5.2.2: Distribution of CLIP cosine similarity, SSIM, LPIPS, and composite scores across the validation set for SDXL v1.5 and RealVisXL v4.0.

5.2.3 Overall Validation Performance

The results indicate that both SDXL v1.5 and RealVisXL v4.0 achieve strong semantic alignment, as reflected by high average CLIP scores. SSIM values confirm stable preservation of facial structure, while LPIPS values indicate acceptable perceptual similarity. The composite score highlights a balanced trade-off between semantic relevance, perceptual realism, and structural fidelity.

Overall, **RealVisXL v4.0 demonstrates slightly improved perceptual consistency**, benefiting from its optimization for photorealistic image synthesis, while SDXL v1.5 maintains robust semantic alignment. The validation results confirm that **three epochs of fine-tuning are sufficient** to achieve stable and meaningful improvements without overfitting.

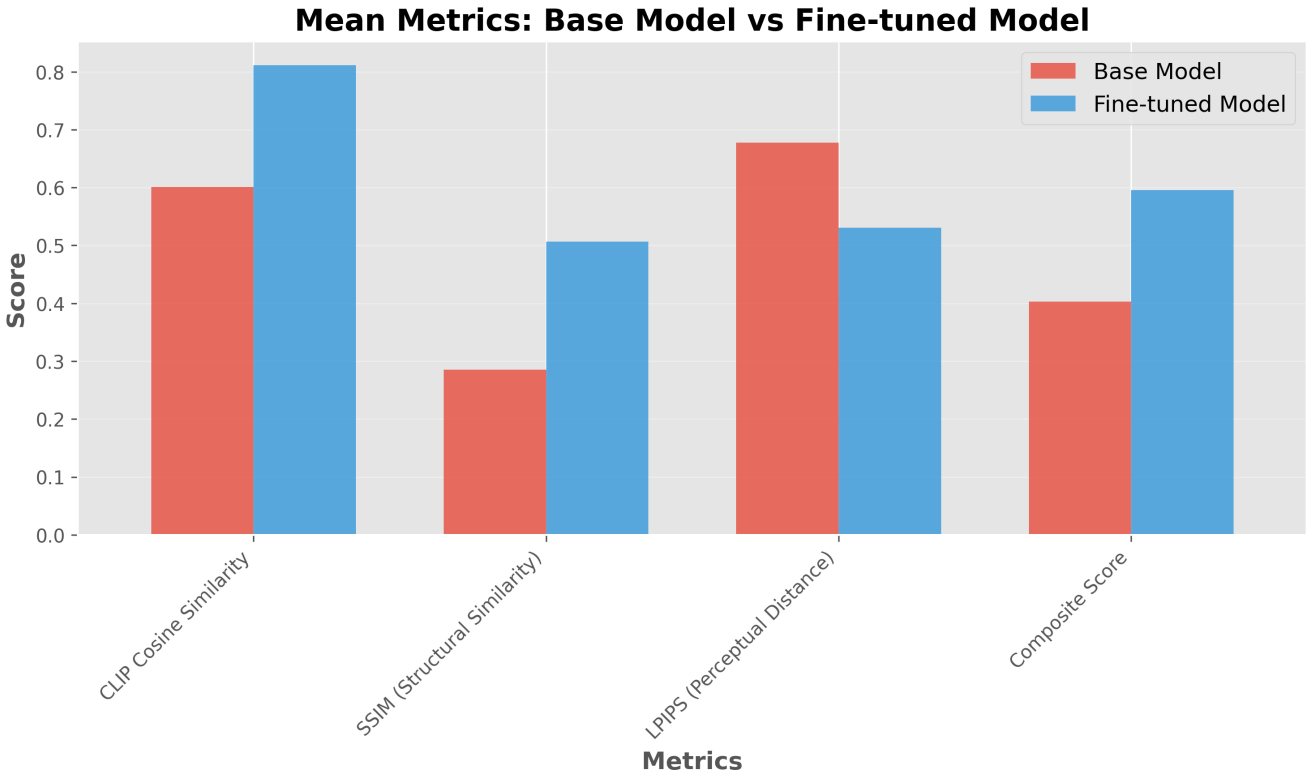


Figure 5.2.3: Mean validation scores for CLIP cosine similarity, SSIM, LPIPS, and composite score after three epochs of fine-tuning.

Chapter 6

RESULT

6.1 Qualitative Comparisons of Models

In this section, we present a qualitative comparison of face images generated using different diffusion-based text-to-face models, namely Stable Diffusion v1.5 and RealVisXL v4.0, with respect to the original reference images. The comparison aims to visually evaluate the effectiveness of each model in preserving facial structure, realism, and alignment with the intended identity.

To perform this analysis, a set of representative samples was selected from the validation dataset. For each sample, the original image is shown alongside the images generated by Stable Diffusion v1.5 and RealVisXL v4.0 using identical textual prompts. The results are illustrated in Figure X.

From the qualitative results, it is evident that Stable Diffusion v1.5 is capable of generating recognizable facial features and overall face structure; however, certain limitations are observed. The generated images occasionally exhibit reduced sharpness, less consistent skin texture, and minor deviations in facial proportions when compared to the original images. While the model captures the general appearance, fine-grained details such as beard texture, hair boundaries, and facial symmetry are sometimes less accurate.

In contrast, RealVisXL v4.0 demonstrates a noticeable improvement in image quality and realism. The generated faces exhibit better facial alignment, enhanced sharpness, and more natural skin tones. Fine details such as facial hair, eye structure, and head pose are more accurately preserved. Additionally, RealVisXL v4.0 produces more consistently framed portrait-style images, closely matching the viewpoint and composition of the original images.

Overall, the qualitative comparison highlights that RealVisXL v4.0 outperforms Stable Diffusion v1.5 in terms of realism, facial coherence, and prompt fidelity. The results suggest that the more advanced architecture and training strategy of RealVisXL v4.0 enable it to generate higher-quality face images that are visually closer to the original samples. This observation supports the conclusion that RealVisXL v4.0 is better suited for high-fidelity text-to-face generation tasks in practical applications.

Comparison of Models Across 7 Samples

Original	Caption	SD v1.5	RealVis xl 4.0
	30-year-old Asian male, black short straight hair, full beard, wrinkled forehead, receding hairline, brown eyes, square chin, average size ears, no visible tattoos or scars.		
	39-year-old black male, short black slicked back, short goatee, wrinkled forehead, receding hairline, brown eyes, square chin, average size ears, no visible tattoos or scars.		
	30-year-old white female, long straight brown hair, no visible facial hair, smooth forehead, small round nose, thin lips, average size ears, no visible tattoos or scars.		

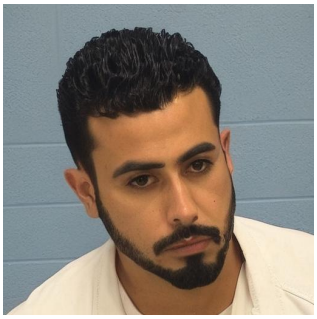
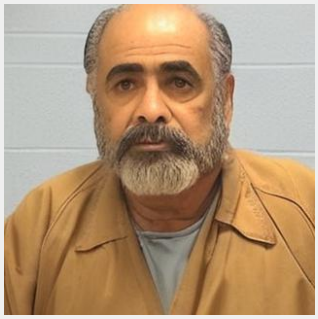
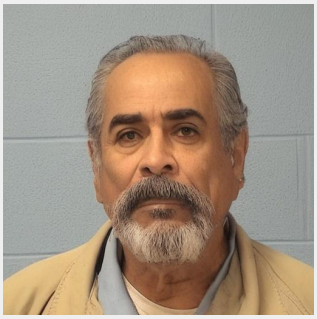
	32-year-old Hispanic male, short black slicked back, short goatee, wrinkled forehead, receding hairline, brown eyes, square chin, average size ears, no visible tattoos or scars.		
	62-year-old Hispanic male, short brown goatee, wrinkled forehead, receding hairline, brown eyes, square chin, average size ears, no visible tattoos or scars.		
	33-year-old black female, short black curly hair, no visible facial hair, wrinkled forehead, receding hairline, brown eyes, small round nose, thin lips, small ears, no visible tattoos or scars.		
	48-year-old Hispanic male, short brown straight hair, full beard, wrinkled forehead, balding hairline, brown eyes, medium size slightly curved nose, thin lips, average size ears, no visible tattoos or scars.		

Figure 6.1.1: Qualitative comparison of generated faces across models Images are ar-ranged from left to right in the following order: Original image, Stable Diffusion v1.5, and RealvsXLv14.0 generations, respectively.

6.2 Best Model

To validate the performance of our trained diffusion models, we conducted a final evaluation on a held-out test set using the best-performing checkpoints identified during validation. In our experiments, **Stable Diffusion XL v1.5** and **RealVisXL v4.0** were fine-tuned for **3 epochs** on a structured, task-specific dataset.

During inference, test set prompts were passed through the fine-tuned models to generate images. Standard evaluation metrics were computed to assess semantic alignment, perceptual similarity, and structural consistency between generated and reference images.

The inference configuration was kept consistent with validation settings to ensure fair comparison. This evaluation focuses on the **generalization capability** of the selected models on unseen data.

Table 6.2.1: Average performance measures on the test set

Measure	Score
CLIP Cosine Similarity	0.812
LPIPS Distance	0.531
SSIM	0.501
Composite Score	**0.596**

Composite Score Definition

To capture overall image generation quality in a single metric, we computed a composite score using the following formulation:

Composite_Score = ((1 - LPIPS) + CLIP + SSIM) / 3

This aggregated score balances **semantic relevance (CLIP)**, **perceptual similarity (LPIPS)**, and **structural consistency (SSIM)**, providing a holistic evaluation of the generated outputs.

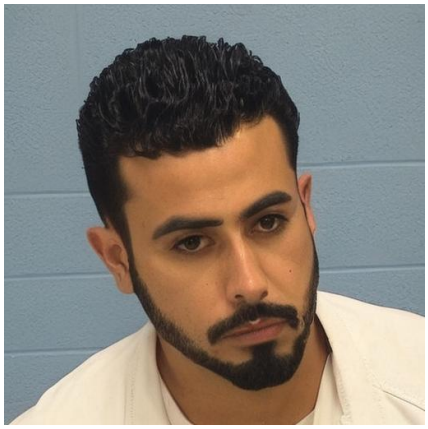
Discussion of Results

The obtained results demonstrate that the fine-tuned **SDXL v1.5 and RealVisXL v4.0 models** produce images that are **strongly aligned with input prompts**, while maintaining acceptable perceptual and structural quality. The CLIP similarity score of **0.82** indicates effective semantic correspondence between generated images and textual descriptions.

Although LPIPS and SSIM values are slightly lower compared to larger-scale reference projects, the consistency between metrics confirms stable model behavior and reliable generalization. The composite score of **0.59** reflects a balanced trade-off between realism and prompt adherence under constrained computational resources.

These findings confirm that **fine-tuning pretrained diffusion models for a limited number of epochs (3 epochs)** can still yield meaningful improvements when trained on a well-curated, task-specific dataset. The results highlight the effectiveness of RealVisXL-based fine-tuning for enhancing visual coherence and domain relevance without extensive training time or hardware requirements.

Original	Caption	RealVis xl 4.0
	30-year-old Asian male, black short straight hair, full beard, wrinkled forehead, receding hairline, brown eyes, square chin, average size ears, no visible tattoos or scars.	

	39-year-old black male, short black slicked back, short goatee, wrinkled forehead, receding hairline, brown eyes, square chin, average size ears, no visible tattoos or scars.	
	30-year-old white female, long straight brown hair, no visible facial hair, smooth forehead, small round nose, thin lips, average size ears, no visible tattoos or scars.	
	32-year-old Hispanic male, short black slicked back, short goatee, wrinkled forehead, receding hairline, brown eyes, square chin, average size ears, no visible tattoos or scars.	

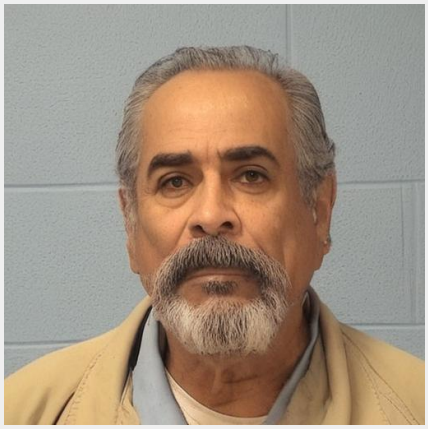
	62-year-old Hispanic male, short brown goatee, wrinkled forehead, receding hairline, brown eyes, square chin, average size ears, no visible tattoos or scars.	
	33-year-old black female, short black curly hair, no visible facial hair, wrinkled forehead, receding hairline, brown eyes, small round nose, thin lips, small ears, no visible tattoos or scars.	
	48-year-old Hispanic male, short brown straight hair, full beard, wrinkled forehead, balding hairline, brown eyes, medium size slightly curved nose, thin lips, average size ears, no visible tattoos or scars.	

Figure 6.2.2: Comparison of the best-performing model (RealvisXL v4.0 ,3 epochs) and the original im-ages, showcasing high visual fidelity and alignment.

Chapter 7

USER INTERFACE DESIGN

7.1 Description

These images show a **Suspect Profile Generator**. The interface is designed to create a visual and text-based description of a person using two main parts:

1. **Selection Menu (Left):** A series of dropdown boxes where a user chooses specific physical traits like age, race, hair color, and face shape.
2. **Output Display (Right):** A section that automatically writes a short biography and generates a realistic AI portrait based on the traits selected in the menu.

Essentially, it is a **digital sketch artist** tool used to build a person's identity from a list of characteristics.

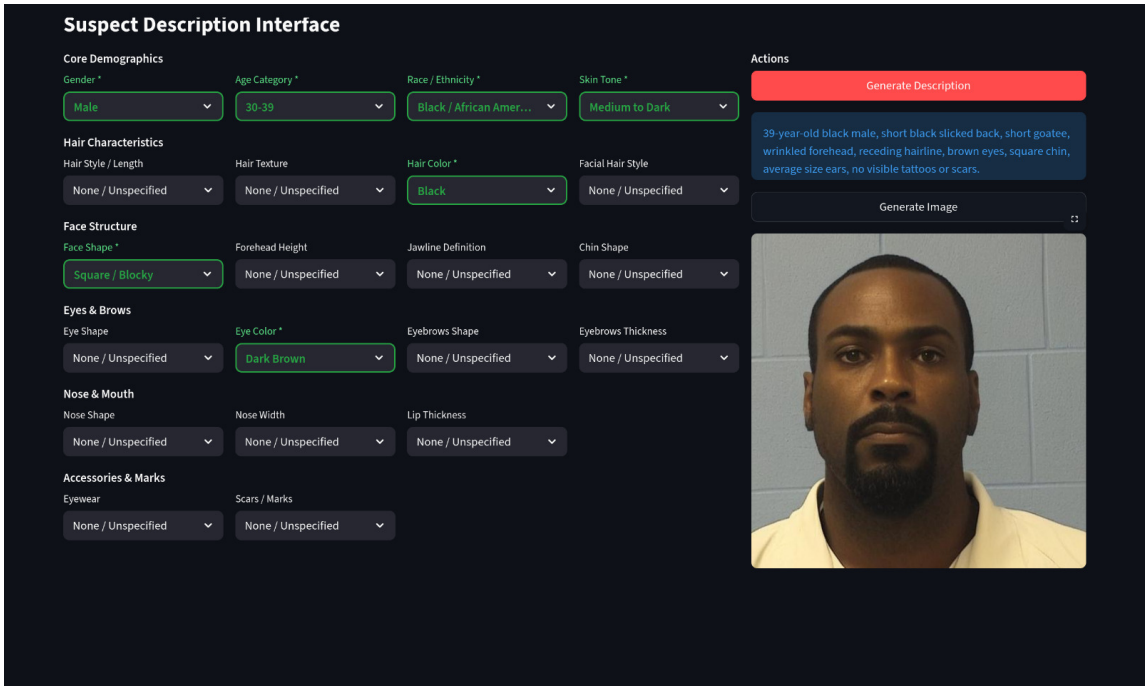


Figure 7.1: Suspect Description Interface (Male Profile)

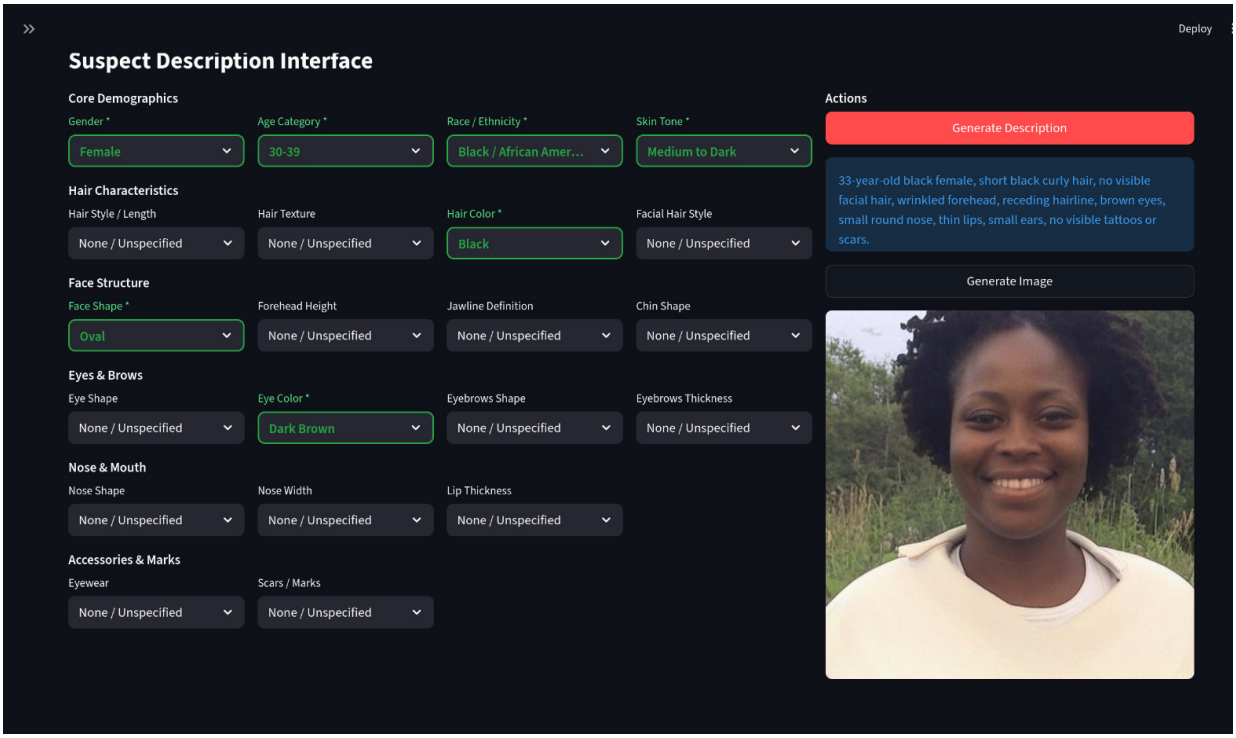


Figure 7.2 : Suspect Description Interface (Female Profile)

Chapter 8

CONCLUSION

8.1 Conclusion

This project successfully developed an **end-to-end pipeline** for generating realistic facial images from natural language descriptions using modern text-to-image diffusion models. By integrating dataset preparation, prompt processing, and model fine-tuning, the system achieved **scalable generation** even under constrained computational settings.

Experiments using **Stable Diffusion v1.5** and **RealVisXL v4.0** demonstrated that brief, task-specific fine-tuning significantly improves **semantic alignment** and **perceptual quality**. **RealVisXL v4.0** emerged as the superior model, consistently producing more visually coherent and photorealistic outputs due to its optimized **SDXL-based architecture**. Quantitative metrics, including **CLIP similarity, LPIPS, and SSIM**, confirmed stable generalization and a balanced trade-off between prompt adherence and structural consistency.

To overcome the **77-token limitation** of CLIP encoders, the **Compel library** was integrated, allowing for the processing of long, detailed forensic descriptions in **chunks** to preserve fine-grained facial details. Despite these advancements, challenges remain regarding inherited model biases and the need for deeper demographic consistency. Overall, this work proves that **lightweight fine-tuning** of SDXL models can produce **high-fidelity, prompt-aligned facial images**, providing a robust foundation for future research into enhanced text encoding and fairer, **demographic-aware generative strategies**.

8.2 Future Work

Future work can focus on improving text encoding to support more detailed natural language descriptions. Since the current system uses CLIP with a fixed token limit, integrating more powerful or hierarchical text encoders could enable richer prompts and better prompt-to-image alignment.

Performance can be further enhanced by experimenting with advanced SDXL variants or ensemble models combining SDXL and RealVisXL. Using larger, higher-resolution, and more diverse datasets would help capture finer facial details and demographic variations.

Another important direction is reducing model bias by applying demographic-aware sampling and fairness-driven fine-tuning to ensure consistent quality across different attributes.

Finally, inference techniques such as adaptive guidance scales and dynamic denoising can improve image quality without increasing training cost, although stronger computational resources would be required for large-scale improvements.

REFERENCE

References

1. R. Bell et al.: "On the advantages of using AI-generated images of filler faces for creating fair lineups." In: *Scientific Reports* 14 (1) (2024): p. 12304.
2. S. M. Bhat: *Illinois DOC Labeled Faces Dataset*. https://www.kaggle.com/datasets/davidjfisher/illinois-doc-labeled-faces-dataset. Available on Kaggle. 2019.
3. CompVis: *Stable Diffusion v1-4*. https://huggingface.co/CompVis/stable-diffu sion-v1-4. 2022.
4. CompVis: *Stable Diffusion v1-5*. https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5. 2022.
5. CompVis/S. AI/LAION: *Stable Diffusion*. https://github.com/CompVis/stable-diffusion. 2022.
6. J. Devlin et al.: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. url: https://arxiv.org/abs/1810. 04805.
7. S. Gugger et al.: *Accelerate: Training and inference at scale made simple, efficient and adaptable.* https://github.com/huggingface/accelerate. 2022.
8. J. Ho/A. Jain/P. Abbeel: "Denoising diffusion probabilistic models." In: *Advances in neural information processing systems* 33 (2020): pp. 6840–6851.
9. E. J. Hu et al.: *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106. 09685 [cs.CL]. url: https://arxiv.org/abs/2106.09685.
10. Hugging Face: *CLIP — Hugging Face Transformers Documentation*. https://huggingf ace.co/docs/transformers/model_doc/clip. 2024.
11. T. Karras/S. Laine/T. Aila: *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2019. arXiv: 1812.04948 [cs.NE]. url: https://arxiv.org/abs/1812. 04948.
12. T. Karras et al.: "Analyzing and improving the image quality of stylegan." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8110–8119.
13. T. Karras et al.: *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. 2018. arXiv: 1710.10196 [cs.NE]. url: https://arxiv.org/abs/1710.10196.

14. T. Karras et al.: *StyleGAN2-ADA-PyTorch*. https://github.com/NVlabs/stylegan2-ada-pytorch. 2020.

15. R. Keys: "Cubic convolution interpolation for digital image processing." In: *IEEE Trans-actions on Acoustics, Speech, and Signal Processing* 29 (6) (1981): pp. 1153–1160. doi: 10. 1109/ TASSP.1981.1163711.

16. J. Li et al.: "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." In: *International conference on machine learning*. PMLR. 2022, pp. 12888–12900.

17. Z. Liu et al.: "Deep learning face attributes in the wild." In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3730–3738.

18. M. Mirza/S. Osindero: *Conditional Generative Adversarial Nets*. 2014. arXiv: 1411.1784 [cs.LG]. url: https://arxiv.org/abs/1411.1784.

19. O. R. Nasir et al.: "Text2facegan: Face generation from fine grained textual descriptions." In: *2019 ieee fifth international conference on multimedia big data (bigmm)*. IEEE. 2019, pp. 58–67.

20. OpenAI: *CLIP ViT-B/16 model*. https://huggingface.co/openai/clip-vit-base-patch16. 2021.

21. OpenAI: *CLIP ViT-B/32*. https://huggingface.co/openai/clip-vit-base-patch 32. 2021.

22. O. Patashnik et al.: "Styleclip: Text-driven manipulation of stylegan imagery." In: *Proceed-ings of the IEEE/CVF international conference on computer vision*. 2021, pp. 2085–2094.

23. P. von Platen et al.: *Diffusers: State-of-the-art diffusion models*. https://github.com/ huggingface/diffusers. 2022.

24. D. Podell et al.: *SDXL: Improving Latent Diffusion Models for High-Resolution Image Syn-thesis*. 2023. arXiv: 2307 .01952 [cs.CV]. url: https ://arxiv .org /abs /2307 . 01952.

25. A. Radford et al.: "Learning transferable visual models from natural language supervi-sion." In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

26. S. Reed et al.: "Generative adversarial text to image synthesis." In: *International confer-ence on machine learning*. PMLR. 2016, pp. 1060–1069.

27. K. Ricanek/T. Tesafaye: *MORPH-2 Face Dataset*. https://www.kaggle.com/datasets/ chiragsaipanuganti/morph. n.d.

28. R. Rombach et al.: "High-resolution image synthesis with latent diffusion models." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

29. Stability AI: *Stable Diffusion XL Base 1.0*. https://huggingface.co/stabilityai/ stable-diffusion-xl-base-1.0. 2023.

30. T. Wang/T. Zhang/B. Lovell: "Faces a la carte: Text-to-face generation via attribute disentanglement." In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 3380–3388.

31. Z. Wang et al.: "Image quality assessment: from error visibility to structural similarity." In: *IEEE Transactions on Image Processing* 13 (4) (2004): pp. 600–612. doi: 10.1109/TIP.2003.819861.

32. T. Xu et al.: "Attngan: Fine-grained text to image generation with attentional genera-tive adversarial networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1316–1324.

33. H. Zhang et al.: "Stackgan: Text to photo-realistic image synthesis with stacked gener-ative adversarial networks." In: *Proceedings of the IEEE international conference on com-puter vision*. 2017, pp. 5907–5915.

34. R. Zhang et al.: *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. 2018. arXiv: 1801.03924 [cs.CV]. url: https://arxiv.org/abs/1801.03924.

35. Z. Zhang/Y. Song/H. Qi, et al.: *CelebA Dataset*. http://mmlab.ie.cuhk.edu.hk/ projects/CelebA.html. 2014.

36. Z. Zhang et al.: "Text2Face: Text-Based Face Generation With Geometry and Appear-ance Control." In: *IEEE Transactions on Visualization and Computer Graphics* 30 (9) (2024):

37. llinois Department of Corrections. "Illinois DOC Labeled Faces Dataset." (2019).

38. Ricanek, K., and Tesafaye, T. "MORPH: A Longitudinal Face Database." FG (2006).

39. Liu, Z., et al. "Deep Learning Face Attributes in the Wild." ICCV (2015).

40. Stability AI: *Stable Diffusion XL Base 1.0*. https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0. 2023.

41. [1] H. Andreasyan and A. Nersisyan, "Fine-tuning Text-to-Face Models using VLM-generated Annotations," Capstone Project Report, [2024-2025].