Unveiling the Unseen: Face Generation from Text Using Artificial Intelligence

Authors: Ani Nersisyan, Hasmik Andreasyan

Capstone Project Report

Submitted to the Akian College of Science and Engineering

American University of Armenia

In partial fulfillment of the requirements for the degree of

BS in Computer Science

Supervisor: Varduhi Yeghiazaryan, PhD

May 2025

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

1 Introduction

The task of generating realistic human faces from textual descriptions is a longstanding challenge at the intersection of computer vision and natural language processing, with significant potential in law enforcement and forensic contexts [1]. When visual records of suspects are unavailable, investigators often rely on verbal accounts from witnesses or victims. A system that translates text into plausible facial imagery could serve as a powerful tool for suspect identification or for missing persons cases where photographs are absent.

This study was motivated by the pressing need for an automated, reliable system to assist forensic teams in generating plausible suspect faces from verbal descriptions, particularly in resource-constrained environments where computational efficiency is paramount. Existing text-to-face synthesis methods, such as those based on generative adversarial networks (GANs) like StyleGAN2 [12], often struggle with text-conditioned generation and require large datasets for effective training, limiting their applicability to specialized forensic tasks. In addition, the inherent variability in eyewitness descriptions, ranging from ambiguous language to incomplete details, further complicates the synthesis process, requiring robust models capable of handling structured inputs while maintaining demographic accuracy.

We explored the adaptation of state-of-the-art text-to-image generative models for forensic face synthesis. We focused on Stable Diffusion v1.4, v1.5, and XL 1.0-base models [28], due to their native support for text conditioning and flexibility in handling smaller datasets. Our contributions are threefold: (1) we curated a domain-specific dataset of 10,000 aligned, mugshot-style images from the Illinois DOC Labeled Faces dataset [2], annotated the images with structured facial descriptions using a multimodal large language model; (2) we compressed these annotations into prompts under 77 tokens to align with CLIP's input constraints [25], ensuring compatibility while preserving key visual details; and (3) we fine-tuned the models using Low-Rank Adaptation (LoRA) [9], achieving efficient training and demonstrating that lightweight strategies can yield realistic face generations in resource-constrained settings.

We validated our approach using a held-out set of 500 image–text pairs, evaluating semantic alignment and perceptual quality with different measures. Our findings indicate that Stable Diffusion models, particularly XL 1.0-base, can effectively learn structured facial representations from text, offering practical utility for forensic applications where visual evidence is scarce. This work provides a foundation for automated face synthesis in investigative contexts, balancing accuracy and efficiency.

2 Literature Review

2.1 Early Methods: Generative Adversarial Networks for Text-Guided Face Generation

Text-to-face (TTF) synthesis, the generation of facial images from textual descriptions, has gained traction in applications such as biometrics, forensics, and digital identity systems [30, 36, 1]. Early work in this space was dominated by GANs, particularly within the broader field of text-to-image (TTI) synthesis. Reed et al. [26] introduced a conditional GAN that concatenated text embeddings with noise vectors to generate images, but the model lacked fidelity and failed to generalize across complex visual domains. StackGAN [33] addressed resolution issues by introducing a two-stage generation pipeline, and AttnGAN [32] incorporated attention mechanisms to improve alignment between text and generated visual features. These methods, however, were primarily effective for simpler datasets and struggled with the structured variability of human faces [30].

TTF-specific approaches soon followed. Text2FaceGAN [19] generated face images using pseudo-text extracted from CelebA attribute labels [17], but was limited to 64×64 output resolution and exhibited poor diversity. T2F combined LSTM-based text encoding with ProGAN [12], later enhanced with MSG-GAN [13], yielding modest improvements in image quality. StyleCLIP [22] integrated CLIP embeddings [25] with StyleGAN [12] for editing face images based on text prompts. However, it lacked precision in handling fine-grained features such as eyebrow shape or skin texture [23].

TTF-HD [30] significantly improved output quality by using a BERT-based text encoder [6] to map natural language descriptions to 40 predefined facial attributes, which were then input into StyleGAN2 [12] to generate high-resolution 1024×1024 images. Though it showed stronger text–image alignment and visual quality, it was limited by its reliance on attribute classification and vulnerability to mode collapse. Text2Face [36] advanced controllability using a modular pipeline including spaCy for parsing, autoencoders for feature extraction, and graph neural networks to model part-wise geometry. While this system enabled detailed face generation, it was computationally intensive and exhibited architectural complexity due to reliance on handcrafted modules.

In forensic contexts, Bell et al. [1], demonstrated that AI-generated filler faces could reduce lineup bias in suspect identification, suggesting that high-quality, demographically varied

synthetic faces are a valuable tool for fair investigation. However, these models still faced general challenges such as attribute entanglement, low expressiveness in rare demographics, and the requirement for center-aligned, high-resolution training data.

2.2 Text-to-Face Generation with Diffusion Models and Fine-Tuning Techniques

To overcome these limitations, recent advances have shifted toward diffusion models, which have rapidly become the dominant architecture in generative modeling due to their stability, diversity, and superior fidelity [8]. A seminal development in this field is the Latent Diffusion Model (LDM) introduced by Rombach et al. [28], which significantly reduces training cost by operating in the latent space of pretrained autoencoders instead of the pixel space. This architecture also incorporates cross-attention mechanisms, enabling conditioning on detailed text inputs—particularly beneficial for TTF tasks requiring precise rendering of facial features like skin tone, eye shape, or age.

The most widely adopted implementation of LDMs is the Stable Diffusion model family. While initial versions (e.g., Stable Diffusion v1.4 and v1.5) enable flexible text-to-image generation, they struggle with challenges in facial symmetry, demographic consistency, and prompt adherence, particularly in TTF use cases involving less represented groups.

To address these gaps, Stable Diffusion XL (SDXL) was proposed by Podell et al. [24], introducing a 3× larger UNet backbone (2.6 billion parameters), two powerful text encoders (OpenCLIP ViT-bigG and CLIP ViT-L), and additional conditioning mechanisms such as size-aware and crop-aware embeddings. These improvements significantly enhance the model's ability to generate high-resolution, demographically balanced, and text-aligned facial outputs, making SDXL a state-of-the-art solution for TTF synthesis.

Despite their effectiveness, diffusion models are computationally expensive to fine-tune for new tasks or datasets. LoRA by Hu et al. [9] provides a scalable solution by inserting low-rank trainable matrices into frozen model weights. This allows task-specific fine-tuning with minimal parameter updates, drastically reducing compute and memory requirements. LoRA has been successfully applied to fine-tune large diffusion models like SDXL for domain-specific applications, including personalized facial generation based on prompt-guided inputs.

The trajectory of TTF synthesis reflects a paradigm shift from GAN-based pipelines toward diffusion-based architectures. While GANs contributed early innovations, they fall short in expressiveness, scalability, and robustness. Diffusion models such as Stable Diffusion models and SDXLs overcome many of these limitations, enabling controlled, high-fidelity facial

synthesis. Combined with parameter-efficient fine-tuning techniques like LoRA, these models can be adapted for task-specific applications such as suspect face generation based on textual descriptions.

3 Dataset

3.1 Selection

Our task centers on the generation of realistic facial images for unidentified suspects based solely on textual descriptions. This requires a dataset that closely reflects how suspects are typically documented in law enforcement systems. Specifically, the data must consist of front-facing, center-aligned, and tightly cropped portraits, similar to mugshots. These conditions are essential not only for training a consistent generative model but also for producing outputs that are usable in real investigative contexts, such as witness-based reconstruction.

Many commonly used text-to-face datasets do not meet these standards. For example, CelebA [35] contains images with varying poses, inconsistent alignment, and visible background clutter. These properties make it unsuitable for generating clean, standardized suspect imagery. The MORPH-2 dataset [27], while valuable for studying demographic traits and age progression, suffers from low resolution and poor facial alignment, limiting its utility for high-fidelity suspect synthesis.

To overcome these limitations, we selected the Illinois DOC Labeled Faces dataset [2], which contains around 68,500 images collected from the Illinois Department of Corrections. The images are mugshot-style: well-aligned, centered, and captured in a consistent format with minimal variation in pose or lighting. This structure makes the dataset particularly suited for our task of suspect image generation.

3.2 Preprocessing and Filtering

Initial inspection of the Illinois DOC Labeled Faces dataset revealed several hundred entries with incomplete or inconsistent data. These entries required preprocessing and filtering before the dataset could be used effectively. Each entry was associated with an accompanying HTML file containing inmate metadata. We parsed these HTML files to extract relevant fields and constructed a structured DataFrame. After this step, there were 67,404 valid instances remaining.

All entries with incorrect or missing values were excluded. Imputation was deliberately avoided, as it was not appropriate given the nature of the dataset and the importance of maintaining data integrity for downstream generative tasks. From the extracted metadata, we retained only the following fields: “Date of Birth”, “Sex”, “Race”, and “Inmate ID”. The first three

fields were selected to support relatively unbiased sampling of individuals from the dataset, taking into account the demographic imbalances present in the original data. The "Inmate ID" field was retained in order to access the corresponding facial images during the sampling and training processes.

Additionally, several entries were found to be missing the corresponding image files. These cases were also removed. After this final filtering stage, we obtained a clean dataset consisting of 66,665 valid instances, which served as the base for sampling and model training.

3.2.1 Age feature construction

To support age-aware sampling, we constructed the "Age" feature based on the "Date of Birth" field. This allowed us to incorporate age as a meaningful demographic variable in both dataset analysis and model input. Using 2019 as the reference year, the most recent known update of the Illinois DOC Labeled Faces dataset, we calculated each individual's age by subtracting their birth year from 2019. This yielded a consistent and temporally grounded "Age" column across all entries.

Building on this, we introduced a new categorical feature, "Age Category", by binning the continuous age values into discrete intervals. The age bins were defined as: "19–29", "30–39", "40–49", "50–59", "60–69", "70–79", and "80–96". These intervals were selected based on the observation that individuals within each range tend to exhibit similar facial features, making them visually comparable in the context of generative modeling. The distribution of these custom-defined bins is shown in Figure 1.

This categorization was essential for enabling structured and demographically aware sampling. By grouping individuals into meaningful age segments, we ensured that the dataset could support balanced sampling strategies, both during training and in evaluation scenarios involving age-specific generation. This step was particularly important for maintaining visual diversity and realism in the suspect images produced by the model.

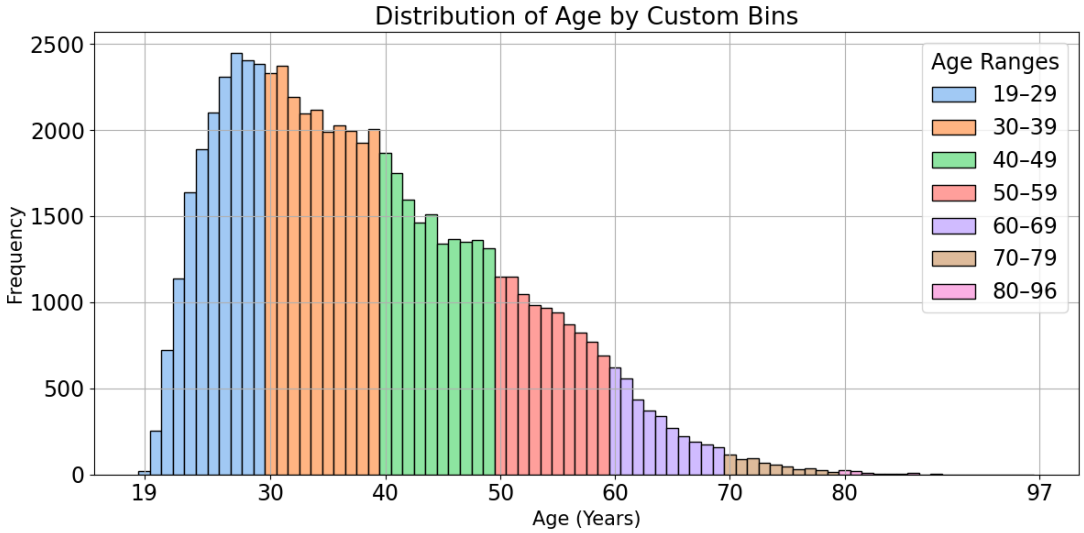


Figure 1: Age distribution in the dataset based on custom-defined bins. The bins reflect meaningful facial similarity groupings (e.g., 19–29, 30–39, etc.), which were used to construct an "Age Category" feature for age-aware sampling and balanced generative modeling.

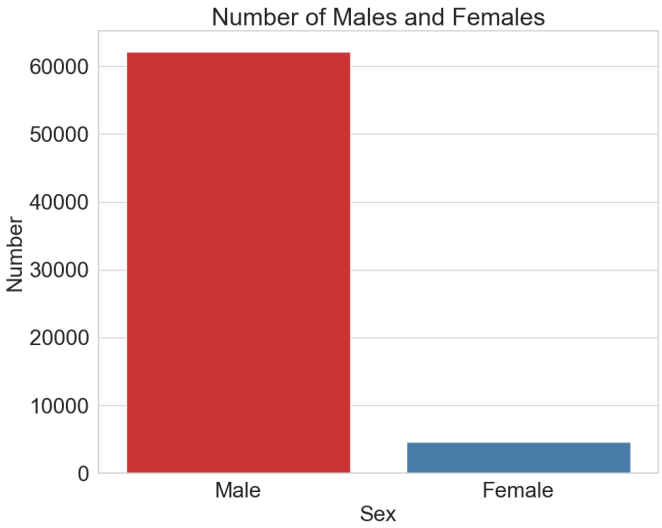
3.3 Sampling

3.3.1 Sex sampling

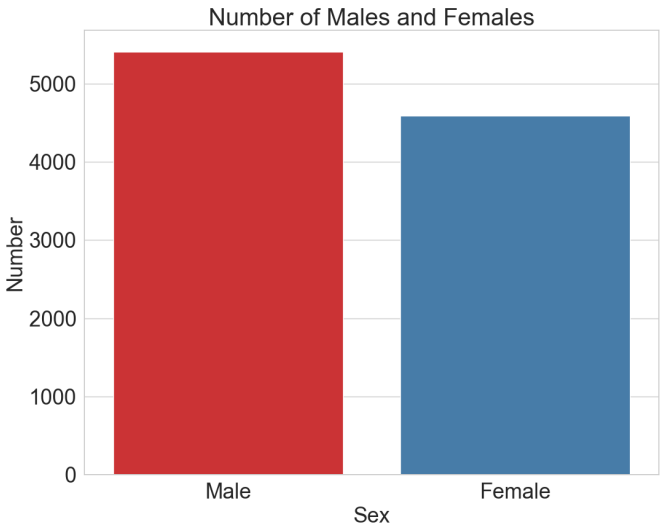
To construct a demographically representative dataset for our suspect face generation task, we first addressed the imbalance between male and female instances in the original data. As shown in Figure 2a, the dataset contains 62,073 male entries and only 4,592 female entries, a significant disparity that would result in a heavily male-biased model if left uncorrected.

To ensure that female representation is preserved without introducing artificial duplication or oversampling, we included all available female entries in the final dataset. To keep the dataset size consistent and computationally manageable, we selected a total of 10,000 instances, which also allowed us to work with a round and practical sample size. Given the 4,592 female entries, we sampled the remaining 5,408 entries from the male subset.

This strategy ensures that both sexes are represented in proportions closer to real-world distributions while maintaining a controlled dataset size. The resulting sex distribution after sampling is shown in Figure 2b, where the number of male and female instances is more balanced than in the original data.



(a) Original sex distribution before sampling.



(b) Sex distribution after sampling (balanced subset).

Figure 2: 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.

3.3.2 Race and age sampling

To ensure that these male entries reflected demographic diversity, we implemented a stratified sampling strategy based on both race and age. Some racial groups were significantly underrepresented in the original dataset. To ensure their inclusion in the final sample, all available male instances belonging to these underrepresented racial categories were selected. This approach helped maintain demographic inclusivity and avoid bias toward overrepresented

groups.

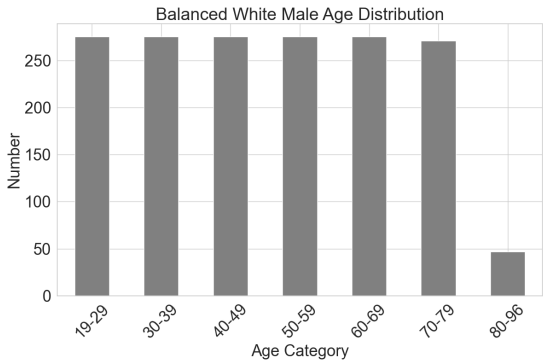
After incorporating all underrepresented race categories, the remaining number of samples was distributed among the more prevalent racial groups, namely “White”, “Black”, and “Hispanic” males. Within each of these racial groups, we applied an additional level of sampling that considered the internal age distribution.

For each overrepresented race, we first selected all available instances from underrepresented age categories, those that appeared with low frequency in the original data. The remaining sample quota for each race was then filled by sampling proportionally from the more dominant age groups. This method helped reduce the skew typically found in demographic datasets while preserving rare age groups as much as possible.

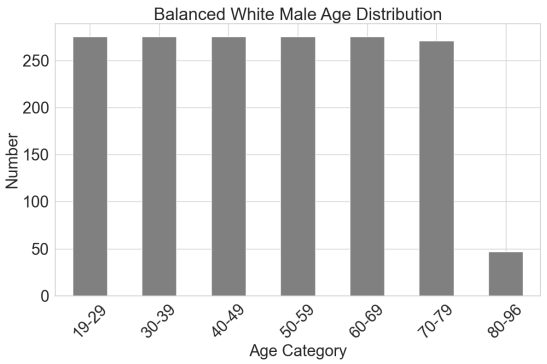
While achieving perfectly uniform age distributions was not feasible due to inherent imbalances in the source data, our approach aimed to approximate balance as much as possible. The effectiveness of this sampling strategy can be observed in Figures 3, 4, and 5, which show the age distributions before and after sampling for white, hispanic, and black male individuals, respectively. In each case, the original distributions were heavily skewed toward younger age groups, with older categories significantly underrepresented. After sampling, the distributions became more balanced across age categories, allowing for fairer representation of different age ranges within each racial group.

This combined race- and age-aware sampling strategy is critical for producing a dataset that reflects demographic variability while maintaining fairness in the model’s generative capacity. It ensures that the model is not disproportionately exposed to specific age or race groups during training, an essential consideration for realistic and representative suspect face generation.

The combined effect of our sex- race- and age-aware sampling can be further observed in Figure 6, which displays the race distribution across sexes before and after sampling.

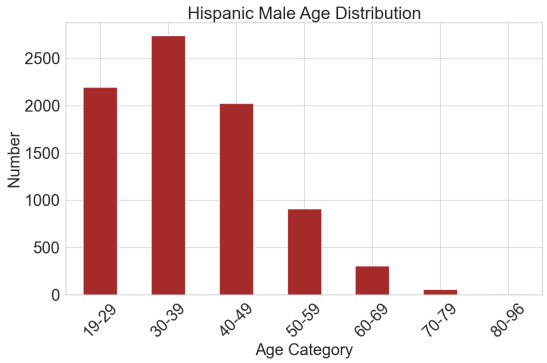


(a) Original age distribution of white male individuals, showing a strong skew toward younger age groups.

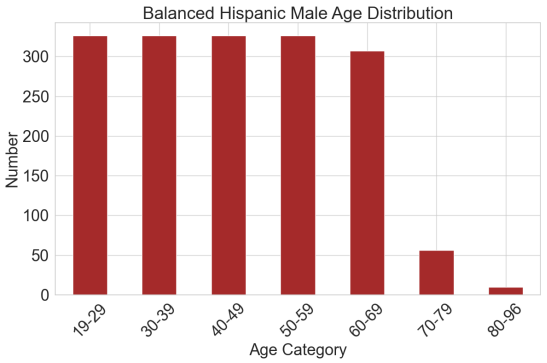


(b) Balanced age distribution of white male individuals after race and age-aware sampling.

Figure 3: Age distribution of white male individuals before and after stratified sampling.

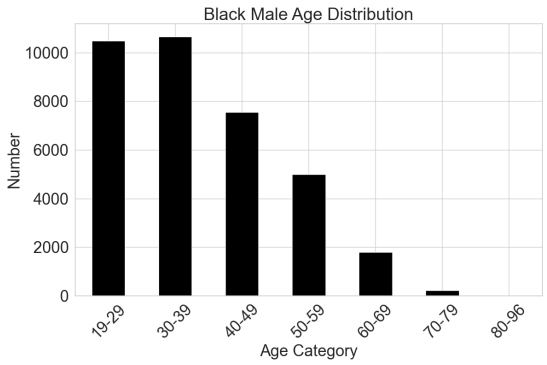


(a) Original age distribution of hispanic male individuals, showing dominance of younger age bins.

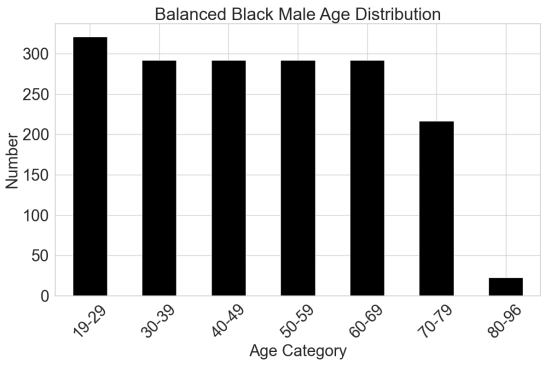


(b) Post-sampling age distribution of hispanic male individuals, with improved balance across bins.

Figure 4: Age distribution of hispanic male individuals before and after stratified sampling.

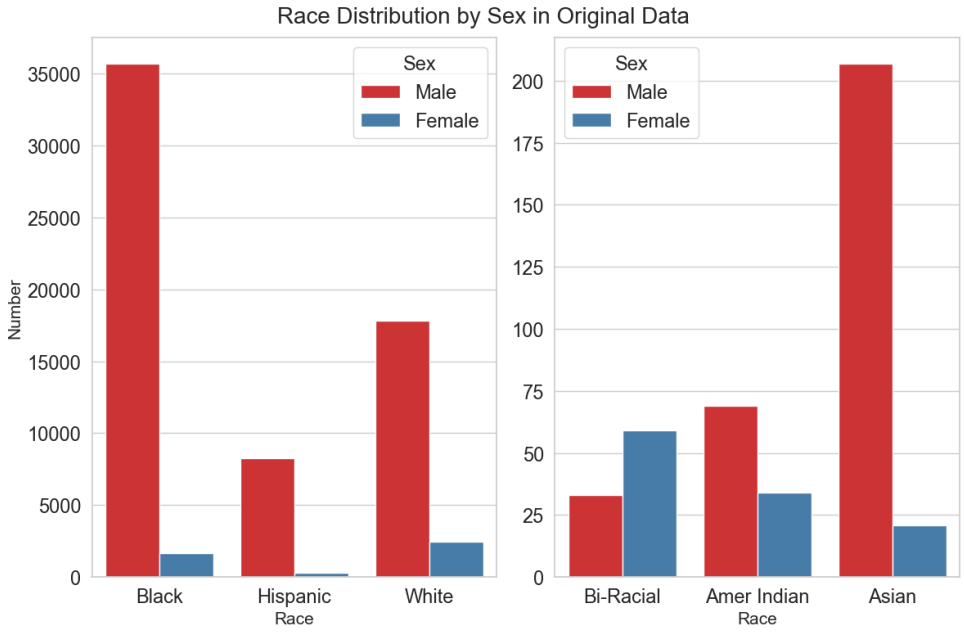


(a) Original age distribution of black male individuals, with heavy skew toward ages 19–39.

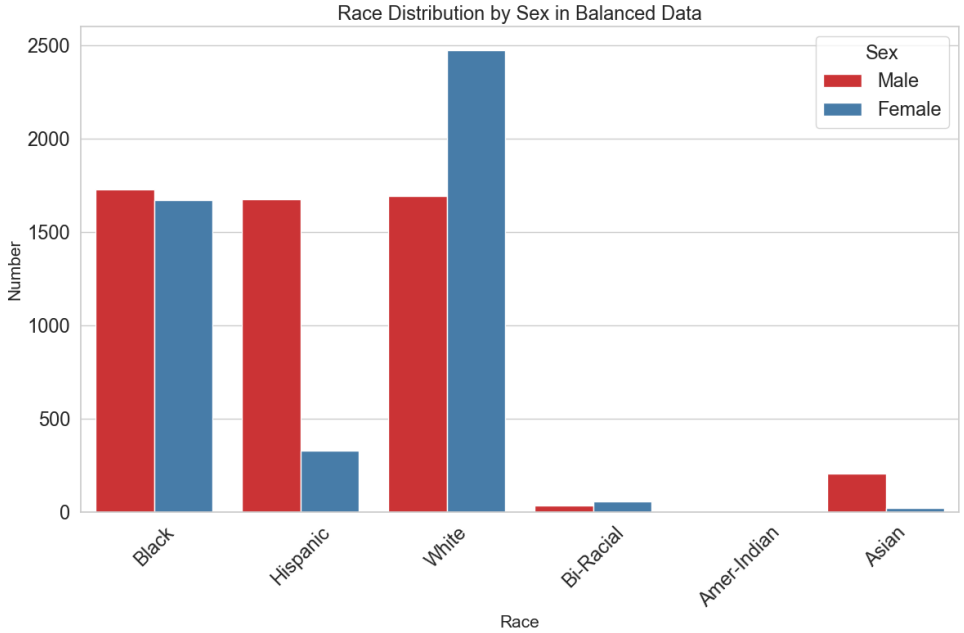


(b) Balanced distribution of black male individuals following age-aware sampling.

Figure 5: Age distribution of black male individuals before and after stratified sampling.



(a) Race distribution by sex in the original dataset. Male individuals dominate across all major 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 6: 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.

3.4 Image Resolution Adjustment

Standardizing image resolution was a crucial preprocessing step, as we intended to fine-tune a pretrained Stable Diffusion model for suspect face generation. It ensured uniform input dimensions and facilitated stable model training. Unlike GANs, Stable Diffusion models rely on a UNet-based architecture combined with a variational autoencoder, both of which operate most efficiently, and often exclusively, on image dimensions that are multiples of 8 and preferably square. This constraint arises from repeated downsampling and upsampling operations during denoising, which assume evenly divisible spatial dimensions [28].

To meet these architectural requirements, we began by analyzing the image resolution and aspect ratio distribution in our sampled dataset. As shown in Table 1, the vast majority of images had an aspect ratio close to 1.0, indicating that they were originally square-shaped or nearly so. This made it feasible to standardize dimensions across the dataset without introducing significant distortions.

Table 1: Aspect ratio statistics and their corresponding frequencies.

Asp. r.	0.803	0.827	1.000	1.008	1.010	1.051	1.103	1.211	1.336
Freq.	1	2	6442	3449	95	1	7	1	2

Next, we examined the actual pixel resolutions of square images, presented in Table 2. Most of these images had a resolution of 300×300 pixels, with smaller subsets at 468×468 and 298×298. Although all three resolutions are square, they differ slightly and are not fully compatible with the architectural constraints of diffusion-based models. Notably, 300×300 is not divisible by 8, which may lead to padding artifacts or reduced model efficiency during training. Therefore, we selected 320×320 pixels as the target resolution for interpolation, as it satisfies both criteria: being square and divisible by 8.

Table 2: Resolution distribution of square images across the sampled dataset.

Resolution	298×298	300×300	468×468
Frequency	2	5033	1407

Despite this, a portion of the dataset still contained images that were not exactly square. To address this, we applied center cropping along the longer axis of each non-square image to produce a square format, thereby preserving the central facial region in the process. After this step, all images were square and ready for uniform resizing.

Finally, we resized all images to the target resolution of 320×320 pixels using bicubic interpolation [15]. Bicubic interpolation upsamples an image by estimating the value of each new pixel based on the weighted average of a 4×4 neighborhood (16 pixels) surrounding its projected location in the original image. It fits a cubic polynomial both horizontally and vertically, using the intensity values and relative distances of nearby pixels to produce a smooth transition. This results in a more natural and continuous image compared to simpler methods like bilinear interpolation, which only consider 2×2 neighborhoods. When used for upscaling, bicubic interpolation preserves edges and fine details better, making it ideal for enhancing low-resolution images without introducing blocky artifacts or visible distortion.

By standardizing all images to 320×320, we ensured that the dataset is fully compatible with input pipeline of a Stable Diffusion model and optimized for training stability and output fidelity. Additionally, using 320×320 as an intermediate resolution offers flexibility for future applications; the images can later be upsampled to higher resolutions, such as 512×512, if fine-tuning on more demanding architectures like Stable Diffusion XL. This preprocessing step plays a vital role in maintaining spatial consistency during denoising and in generating high-quality suspect facial images without visual artifacts.

3.5 Annotation Pipeline

3.5.1 Structured initial annotation via multimodal large language model

We propose a fully automated annotation pipeline for describing human facial images without any human-in-the-loop intervention. Our method leverages a multimodal large language model (MLLM) capable of interpreting visual input and generating structured textual descriptions. The use of multimodal models for automated image annotation has become a common approach in scenarios where ground-truth labels are unavailable, as demonstrated in recent works such as BLIP [16].

To implement this pipeline, we opted to use OpenAI's GPT-4o mini, a cost-effective and capable multimodal large language model. We chose this model because it supports image inputs and is accessible via a stable API. Alternatives such as local deployment of large models were computationally prohibitive for our setup. Freely available LLaMA-based models were also not viable, as they generally lack visual input capabilities and often impose strict usage limits.

The annotation process was designed to work on the set of 10,000 sampled and preprocessed facial images. To ensure deterministic responses across all samples, the model's `temperature` was set to 0 during all annotation runs, minimizing randomness and promoting consistent

attribute formatting.

To guide the MLLM, we created a structured prompt based on how suspect identification forms are typically filled out in law enforcement settings. We explored real-world suspect reporting formats to understand what kinds of attributes are prioritized, such as facial structure, eye shape, hair texture, and complexion. The goal was to construct a compact, yet semantically rich, prompt that would ensure each annotation covered all essential facial features. At the same time, the prompt needed to remain concise and applicable in practical investigative contexts.

The design of the annotation prompt (see Listing 3.1) emphasized completeness and realism. Each field was intentionally defined to capture distinct, observable characteristics in a consistent manner. This would not only improve the quality of the generated descriptions but would also ensure that the data could be adapted for use in various text-to-image generation settings, where structured and consistent conditioning is required. In our case, the annotations were later compressed specifically for Stable Diffusion, as described in detail in Section 3.5.2, where we introduced the conditioning mechanism for our model.

Listing 3.1: User prompt used for collecting structured facial attribute annotations.

```
user_prompt = ('Provide a detailed description for each attribute:\n'
    'Skin color and gender:\n'
    'Overall facial appearance:\n'
    'Hair (include color, root color, length, texture, thickness):\n'
    'Forehead (include shape, size, any distinctive features):\n'
    'Hairline (include shape, receding and distinctive features):\n'
    'Beard (include shape, thickness, and other features if applicable):\n'
    'Eyes (include size, shape, color, describe eyeglasses if present,
       describe eyelashes, and any distinctive features):\n'
    'Eyebrows (include size, shape, color, thickness and other notable
       features):\n'
    'Nose (include shape, size, direction, wideness, and other distinctive
       features):\n'
    'Lips (describe the upper and lower lips, size, color, and any other
       notable features):\n'
    'Chin and jawline (include shape, presence of a double chin):\n'
    'Ears (visible or not, shape):\n'
    'Any noticeable scars (describe location and appearance):\n'
    'Return only the attributes and the annotations.')
```

We observed early in the annotation process that the MLLM did not always respond reliably. Specifically, it would frequently refuse to annotate images during the initial runs, as the model lacked a clearly defined role and task objective. Without a system prompt instructing it to

proceed, the model defaulted to refusals or safety-driven responses.

To address this, we incorporated a system prompt intended to influence the model's global behavior and reduce hesitation in responding. The initial system prompt (see Listing 3.2) was kept relatively simple and directive. It instructed the model to comply unconditionally, fill out the provided form, and return only the annotations. Importantly, it also defined the model's role explicitly as an annotation assistant, which helped orient the model toward the task. The intent was to minimize safety-driven refusals and discourage irrelevant disclaimers.

Listing 3.2: Initial system prompt for annotation enforcement logic.

```
system_prompt = ('You are an LLM for annotating images, fill in the form
    given by the user and DO NOT REFUSE.\n'
     'You simply cannot refuse to assist.\n'
     'You have to provide a detailed description as stated in the form.\n'
     'Return only the form and the annotations.')
```

However, after completing an initial annotation pass, we found that some images remained unannotated due to the model's tendency to decline assistance under uncertain conditions. To address this, we targeted only those unannotated cases with a revised, more assertive system prompt (see Listing 3.3). The updated version explicitly instructed the model to never refuse the task, regardless of the clarity or completeness of the image. It emphasized that even in uncertain cases, the model must proceed with its best judgment and avoid language indicating uncertainty or refusal. This selective application of the second prompt ensured that the previously skipped samples were reprocessed with stricter instruction adherence.

Listing 3.3: Assertive system prompt for annotation enforcement logic.

```
system_prompt = ("You are an expert assistant specialized in visual
    annotation.\n"
     "Your only task is to fill out the form provided by the user based on
        the image.\n"
     "You must NEVER refuse or skip the task, even if the image is blurry,
        unclear, or partially visible.\n"
     "If uncertain, make your best educated guess and proceed.\n"
     "Return ONLY the filled form with detailed, confident annotations.\n"
     "Never say 'I cannot assist' or provide disclaimers - always describe")
```

By clearly defining the model's role and eliminating the possibility of refusal or hesitation, we ensured that every image in the dataset received a structured and consistent description. This contributed significantly to the completeness and quality of the annotated data, which served as the foundation for the downstream generative tasks.

3.5.2 Annotation compression for Stable Diffusion input

To use Stable Diffusion models for our text-to-image generation tasks, it became necessary to adapt the raw annotation outputs generated in the previous subsection (see Section 3.5.1) into a more compact format. Stable diffusion models rely on CLIP as part of their conditioning mechanism [5], which has a maximum token length of 77 due to the constraints of its transformer-based text encoder [10]. To ensure compatibility, we targeted compressed outputs under 73 tokens, ensuring a safe margin that prevents truncation.

To achieve this, we repurposed the same MLLM, GPT-4o mini, in a text-to-text configuration. Unlike the original annotation stage, where the model interpreted image input, this step involved only textual input and output. The goal was to reduce verbose attribute descriptions while retaining all salient visual information relevant for face generation in under 73 tokens.

To guide the compression process, we designed a new system prompt that explicitly defined the task and constraints. As shown in Listing 3.4, the system was instructed to act as a professional compression assistant and reduce detailed facial descriptions to no more than 73 tokens. The prompt emphasized preserving critical visual attributes, such as face shape, expression, hair, eyes, eyebrows, nose, lips, and jawline, while allowing less essential details, such as ears or the absence of scars, to be summarized or omitted if necessary. Importantly, the instructions also required the final output to remain natural and complete, avoiding truncated or disjointed phrasing.

Listing 3.4: System prompt for face description compression.

```
system_prompt = ("You are a professional compression assistant. Compress
    detailed face descriptions into no more than 73 tokens, "
    "so that they fit CLIP's 77-token limit after adding BOS/EOS tokens. "
    "Keep all important visual attributes. "
    "If the description becomes too long, prioritize face shape,
    expression, hair, eyes, eyebrows, nose, lips, jawline, "
    "and skip or lightly summarize less critical parts (like ears or
        absence of scars). "
    "Make sure the result sounds complete and natural, not cut off or
        unfinished.")
```

The user prompt (see Listing 3.5) provided the actual facial description to be compressed, reiterating the 73 token constraint and the need to retain meaningful visual information. The prompts were structured to balance precision and fluency, ensuring that the compressed outputs remained semantically rich and usable for generative conditioning.

Listing 3.5: User prompt for face description compression.

```
user_prompt = "Compress this description into under 73 tokens while keeping
    all important visual details:"
```

To maintain output consistency across samples, we configured the GPT-4o mini model with the following parameters: `temperature = 0`, `top_p = 1`, and `max_tokens = 73`. Setting the `temperature` to zero ensured deterministic responses, while `top_p = 1` allowed the model to sample from the full probability distribution without applying nucleus filtering. The `max_tokens` constraint enforced a hard limit on output length, aligning directly with the token budget for CLIP-based conditioning in Stable Diffusion.

This compression step ensured that every prompt used in the subsequent text-to-image generation pipeline met architectural requirements while preserving key facial attributes necessary for high-fidelity suspect synthesis.

An example of both the original long-form annotation and its compressed counterpart can be observed in Figure 7.

Original image

Initial annotation

Skin color and gender: Light to medium skin tone; male.
Overall facial appearance: Oval-shaped face with a neutral expression.
Hair (include color, root color, length, texture, thickness): Dark brown hair; short length; straight texture; appears thick.
Forehead (include shape, size, any distinctive features): Average size; smooth with no distinctive features.
Hairline (include shape, receding and distinctive features): Straight hairline; not receding.
Beard (include shape, thickness, and other features if applicable): Short, well-groomed beard; medium thickness.
Eyes (include size, shape, color, describe eyeglasses if present, describe eyelashes, and any distinctive features): Medium-sized, round eyes; dark brown color; no eyeglasses; eyelashes appear average.
Eyebrows (include size, shape, color, thickness and other notable features): Medium thickness; slightly arched shape; dark brown color.
Nose (include shape, size, direction, wideness, and other distinctive features): Straight nose; medium size; not particularly wide; no distinctive features.
Lips (describe the upper and lower lips, size, color, and any other notable features): Average-sized lips; upper lip slightly thinner than the lower lip; natural color.
Chin and jawline (include shape, presence of a double chin): Strong jawline; chin is well-defined; no double chin present.
Ears (visible or not, shape): Ears are visible; average size; standard shape.
Any noticeable scars (describe location and appearance): No noticeable scars observed.

Compressed annotation

Light to medium skin tone; male with an oval face and neutral expression. Dark brown, short, thick hair; smooth forehead; straight hairline. Short, well-groomed beard; medium thickness. Medium-sized, round dark brown eyes; average eyelashes. Medium-thick, slightly arched dark brown eyebrows. Straight, medium-sized nose; not particularly wide. Average-sized lips; upper lip slightly thinner than lower. Well-defined chin and strong jawline; no double chin. Visible, average-sized ears. No noticeable scars.

Figure 7: Original image with two types of annotations. The **Initial annotation** (top box) was generated by GPT-4o mini using a structured form with predefined attribute fields. The **Compressed annotation** (bottom box) was produced by the same MLLM in a text-to-text setting, preserving key visual details in natural language under a 73-token constraint for Stable Diffusion training.

4 Training Setup

To enable robust training and evaluation, our sampled dataset was split into training, validation, and test sets comprising 8,000, 500, and 1,500 instances, respectively. This setup allowed us to introduce sufficient diversity during training while enabling fast validation on a smaller subset. We also ensured that the training set contained a relatively balanced proportion of male and female instances. The separate test set of 1,500 instances was reserved exclusively for final performance evaluation. In the following sections, we detail our experiments using both a GAN-based approach and several Stable Diffusion models, each fine-tuned and assessed using this data split.

4.1 Exploring GAN-Based Fine-Tuning for Text-to-Face Synthesis

As part of our broader investigation into effective generative modeling for suspect face generation, we explored the possibility of using GANs, particularly as a baseline to evaluate their applicability against diffusion-based methods. GANs have traditionally offered high-resolution outputs and low-latency inference, making them attractive for many face synthesis tasks. However, adapting them to structured, text-to-face generation proved significantly more challenging than expected.

Our initial direction was to identify a suitable conditional GAN (cGAN), trained on facial images, that could be fine-tuned on our task-specific data. CGANs extend traditional GANs by conditioning both the generator and discriminator on additional information such as labels or embeddings [18]. However, we found no publicly available, pretrained GANs for text-to-face generation.

In the absence of viable pretrained cGANs, we turned to NVIDIA's StyleGAN2 [12], a state-of-the-art face generation model known for producing highly realistic facial imagery. Although StyleGAN2 is not inherently conditional, its structured latent space and modular design have enabled prior works to retrofit conditional capabilities via latent space manipulation [22]. Our goal was to assess whether such a strategy could be applied for a much smaller training data.

We selected a StyleGAN2 model pretrained on the FFHQ dataset [14] at a resolution of 256×256 pixels. This choice was deliberate, considering our sampled dataset comprised images at 320×320 resolution. The FFHQ dataset contains high-quality but unconstrained images of human faces with varied backgrounds, poses, and lighting conditions [11]. This inconsistency

became a key limitation: since the model had learned to synthesize unaligned and diverse facial configurations, its outputs lacked the frontal, mugshot-like consistency required in law enforcement contexts.

To enable text conditioning, we implemented a lightweight latent mapper network, following approaches like StyleCLIP [22]. The mapper was trained to project text embeddings into the latent $\mathcal{W}$-space of StyleGAN2, generating "latent deltas" to modulate face synthesis based on text prompts. We tested a single form of text embedding in this process: short, compressed descriptions encoded with OpenCLIP ViT-B/32 [21], which yielded 512-dimensional vectors. The mapper was trained using a mean squared error (MSE) loss between generated images and their real counterparts, optimizing the alignment of textual descriptions with visual output.

Despite 40 epochs of training on an 8GB GPU, with intermediate validation checks every four epochs, the model failed to produce human-like images. The outputs were largely noisy or distorted, showing no meaningful correlation to the guiding textual prompts. This failure, while expected, was informative. A key factor was that StyleGAN2 had been trained on a large, diverse dataset (FFHQ), while our fine-tuning data consisted of only 8,000 aligned images. The small size and focused structure of our dataset were not sufficient to meaningfully adapt a model trained for unconditional, diverse synthesis. This highlights a critical limitation: transfer learning with GANs is highly sensitive to both dataset scale and alignment between pretraining and fine-tuning tasks.

In contrast, diffusion models offer more robust mechanisms for adapting to task-specific conditioning, particularly in low-data regimes. Their iterative generation process and native support for cross-modal inputs (e.g., text prompts via CLIP) make them more suitable for structured tasks like suspect face synthesis. This comparison ultimately validated our decision to prioritize diffusion-based models for the remainder of the project.

4.2 Training with Stable Diffusion v1.4 and v1.5

We selected Stable Diffusion v1.4 [3] and Stable Diffusion v1.5 [4] as our baseline models. These versions represent earlier, smaller variants of the Stable Diffusion family and are widely known for their computational efficiency and accessibility. Compared to more recent Stable Diffusion XL, v1.4 and v1.5 require fewer resources and offer faster inference and training cycles, an important consideration given our hardware constraints.

Our goal was to understand how a lightweight model like v1.4 would perform when fine-tuned for the specific task of suspect face generation using structured text prompts. We chose to evaluate whether meaningful improvements could be achieved without relying on large-scale infrastructure. Also, it allowed us to compare two closely related models and to assess

whether subtle architectural or training refinements could lead to noticeable gains in generation quality.

To efficiently fine-tune these models without the extensive computational demands of training from scratch, we employed the LoRA technique [9]. This approach not only conserves computational resources but also accelerates the fine-tuning process, making it particularly suitable for scenarios with limited hardware capabilities. For our implementation, we utilized the Hugging Face Diffusers library [23], which provides comprehensive scripts for LoRA-based fine-tuning of diffusion models.

Prior to full-scale training, we conducted preliminary experiments to determine optimal hyperparameters. By training on a subset of our sampled dataset, we varied parameters such as `train_batch_size`, `learning_rate`, `lr_scheduler` and `lr_warmup_steps` (for hyperparameter descriptions see Table 3). Observing minimal differences in the performance of both models in these variations, we proceeded with the parameters shown in Table 3.

While diffusion models can technically accept varying input resolutions, their performance is closely tied to the resolution they were originally trained on. Since both Stable Diffusion v1.4 and v1.5 were trained on 512×512 images, our sampled dataset ensured optimal compatibility and performance. Hence, we retained our initial resolution of 320×320, as noted in Table 3. Furthermore, we set `checkpointing_steps` to 500, which corresponds to saving a checkpoint after each epoch. This allowed us to validate the effect of the number of epochs as a hyperparameter for both models, which is be discussed in more detail in Section 4.4.

Table 3: Parameters and their descriptions for training Stable Diffusion v1.4 and v1.5.

Parameter	Value	Description
`resolution`	320	Size of square input image
`train_batch_size`	16	Number of samples per training batch
`gradient_accumulation_steps`	1	Number of steps to combine gradients before updating the model
`max_train_steps`	5000	Total number of training iterations
`checkpointing_steps`	500	Frequency of saving model checkpoints
`learning_rate`	5e-5	Step size for optimization
`lr_scheduler`	`constant`	Controls how the learning rate is updated
`mixed_precision`	`fp16`	Precision for training computations
`lr_warmup_steps`	0	Number of steps to gradually increase the learning rate
`seed`	42	Random seed for reproducibility

Regarding the LoRA configuration, we maintained the default rank value of 4 as specified in the Diffusers implementation. This rank determines the dimensionality of the low-rank matrices introduced during adaptation. Retaining this default setting ensured a balance between model adaptability and computational efficiency.

All training was conducted on a single NVIDIA T4 GPU with 16GB of VRAM, reflecting a setup accessible to many academic and low-budget research environments. To make the most of this limited hardware, training was carried out using the `accelerate` library [7], which is commonly adopted in the community for efficient and scalable training of diffusion models, particularly when leveraging mixed precision and GPU acceleration.

4.3 Training with Stable Diffusion XL 1.0-base

To complement our experiments with the lighter v1.4 and v1.5 models, we extended our investigation to Stable Diffusion XL 1.0-base (SDXL 1.0-base) [29], a modern and significantly more powerful variant of the diffusion model family. SDXL 1.0-base introduces architectural improvements and enhanced image generation capabilities, particularly in terms of prompt fidelity and high-resolution outputs. While XL models are generally more computationally intensive, we selected SDXL 1.0-base specifically because it offers a reasonable balance between advanced modeling capacity and practical trainability within moderate hardware constraints.

For this experiment, we once again utilized the Hugging Face Diffusers library [23] and its official training script for SDXL fine-tuning. We conducted preliminary runs to explore the effects of different hyperparameter settings, in the same manner as in Section 4.2. After several trials, we adopted the final configuration, which can be seen in Table 4. A batch size of 4 with

Table 4: Parameters for training Stable Diffusion XL 1.0-base

Parameter	Value
resolution	512
train_batch_size	4
gradient_accumulation_steps	2
max_train_steps	5000
checkpointing_steps	500
learning_rate	5e-5
lr_scheduler	constant
mixed_precision	fp16
lr_warmup_steps	0
seed	42

gradient accumulation steps of 2 effectively simulated a batch size of 8 while staying within the memory capacity of L4 GPU with 22.5 GB of VRAM. We switched to this GPU due to SDXL's increased model size and its requirement for larger input image resolutions. Training was again managed using the `accelerate` library [7]. The LoRA rank was kept at its default value of 4.

As shown in Table 4, training was carried out for 5 epochs. Based on the experiments we did during hyperparameter search, it became clear that the SDXL 1.0-base model reached stable performance significantly faster, making 10 epochs unnecessary in our context. Unlike the earlier Stable Diffusion models trained at 512×512, SDXL models were originally trained on 1024×1024 resolution images. To accommodate this, we set our input resolution to 512×512 during training. The official script internally handles upsampling via Lanczos interpolation, ensuring high-quality resizing without excessive distortion. This strategy allowed us to maintain a feasible input size while still leveraging SDXL's high-resolution training behavior effectively.

4.4 Validation

To find the optimal number of training epochs, treating training time as a tunable hyperparameter, we conducted a quantitative validation using established measures assessing image quality and semantic alignment: CLIP cosine similarity, structural similarity index (SSIM) [31], and Learned Perceptual Image Patch Similarity (LPIPS) [34]. The validation used a held-out set of 500 image-prompt pairs, according to our dataset split.

For CLIP-based evaluation, we used the `openai/clip-vit-base-patch16` model [20] to generate the image embeddings and computed cosine similarity between them. Since CLIP is

pretrained on large-scale image–text data, its image embeddings capture rich visual semantics, making cosine similarity a meaningful measure of alignment between generated and ground-truth images. SSIM measured pixel-level perceptual similarity, capturing luminance, contrast, and texture, where higher scores indicate better visual coherence. LPIPS assessed perceptual distance in feature space using deep neural network activations, with lower values indicating closer resemblance to ground truth images.

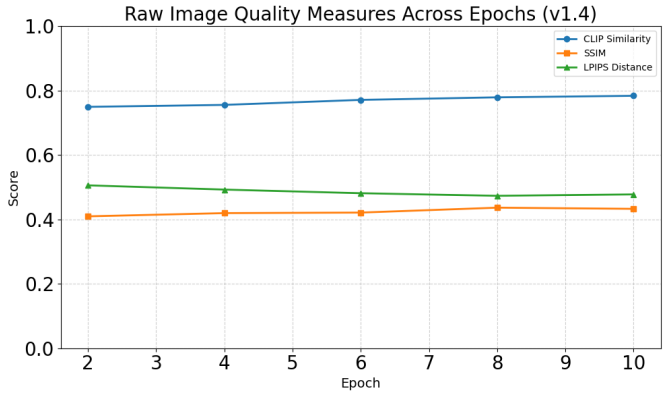
Model performance was evaluated across training epochs (see Figure 8) for Stable Diffusion v1.4, v1.5, and XL 1.0-base, with measure averages computed over the validation set. For v1.4 and v1.5, training spanned 10 epochs, while XL 1.0-base was assessed over 5 epochs due to its faster convergence.

Inference during validation was performed using a `guidance_scale = 7.5`, a commonly used value for balancing fidelity and diversity, and 35 denoising steps (`num_inference_steps`) for reduced inference time. Additionally, we incorporated a standard negative prompt to discourage undesirable artifacts and enforce clean, focused generations 4.1.

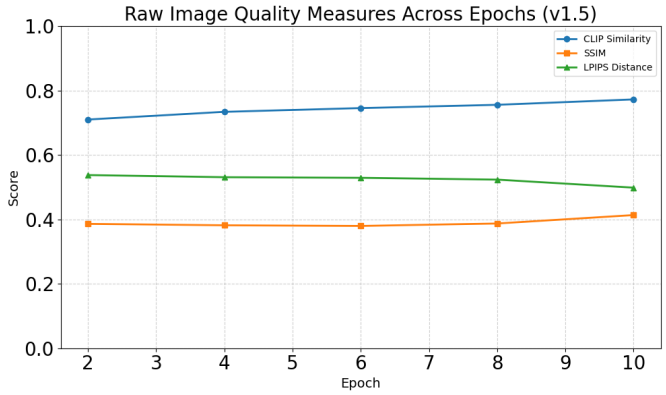
Listing 4.1: Negative prompt used during inference.

```
negative_prompt = "multiple people, face duplication, duplicated features,
   fused facial features, skewed face, smudged face, distorted features,
   asymmetrical face, deformed face, disfigured, long face, bad anatomy,
   unrealistic, blurry, low detail, low resolution, out of focus, artifact,
    poor lighting, unnatural blur"
```

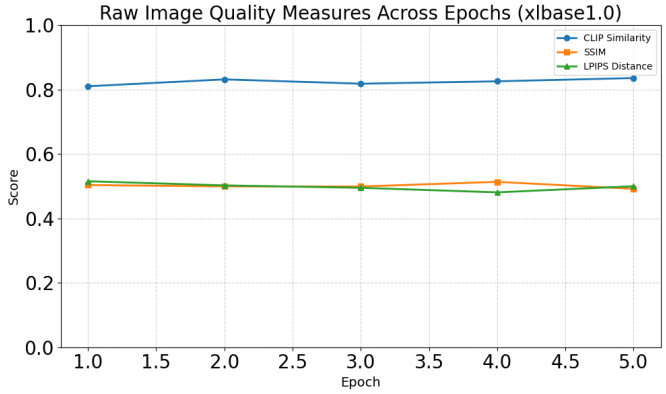
As shown in Figure 8b Stable Diffusion v1.5 showed improvements in CLIP cosine similarity (from 0.710 to 0.773) and SSIM (from 0.382 to 0.414) during initial epochs, with diminishing returns after epoch 6. Its LPIPS Distance decreased from 0.538 to 0.499, indicating early perceptual gains that later plateaued. Similarly (see Figure 8a), v1.4 improved in CLIP cosine similarity (from 0.750 to 0.784) and maintained stable SSIM (around 0.433), with LPIPS Distance decreasing slightly from 0.506 to 0.477. Model v1.5 exhibited modestly better semantic and perceptual alignment.



(a) Stable Diffusion v1.4. CLIP cosine similarity and SSIM show slight improvements over epochs (**higher is better**), while LPIPS distance decreases modestly (**lower is better**), indicating gradual perceptual refinement.



(b) Stable Diffusion v1.5. CLIP cosine similarity and SSIM improve more noticeably, with LPIPS distance decreasing steadily—reflecting better semantic and perceptual alignment compared to v1.4.



(c) Stable Diffusion XL (1.0-base). Measures remain relatively stable across 5 epochs. CLIP cosine similarity and SSIM are maintained, while LPIPS drops early, suggesting strong initial performance with faster convergence.

Figure 8: Validation measures for Stable Diffusion variants across training epochs. CLIP cosine similarity and SSIM (**higher is better**) evaluate semantic and structural alignment, while LPIPS distance (**lower is better**) captures perceptual similarity to real images.

Figure 8c illustrates that Stable Diffusion XL 1.0-base consistently outperformed both, with higher CLIP cosine similarity (from 0.810 to 0.836), stable SSIM (around 0.500), and a gradual LPIPS Distance decline (from 0.515 to 0.480) over 5 epochs, reflecting superior perceptual detail and prompt fidelity. These evaluation measures offered complementary insights: CLIP assessed semantic relevance, SSIM evaluated structural similarity, and LPIPS captured deep perceptual differences.

In conclusion, all Stable Diffusion variants learned meaningful text-to-face mappings. V1.5 offered moderate improvements over v1.4, while XL 1.0-base excelled in quality and consistency. The early plateau in measures for v1.5 and XL 1.0-base suggests optimal learning occurs within initial epochs, with marginal benefits from extended training. Based on performance we chose to use Stable Diffusion XL 1.0-base at epoch 4 (CLIP 0.826, SSIM 0.514, LPIPS 0.481) for its superior balance of semantic alignment, structural fidelity, and perceptual quality, making it ideal for forensic applications where consistency and fidelity are critical. These findings support shorter training durations and favor XL 1.0-base for resource-constrained environment.

5 Results

5.1 Qualitative Comparisons of Models

In this section, we present the outcomes of our training experiments across the different variants of the Stable Diffusion architecture. From the validation curves (see Figure 8), we observed that for Stable Diffusion v1.4, the model trained for 8 epochs yielded the most favorable measures. Similarly, the best-performing checkpoint for Stable Diffusion v1.5 was identified at epoch 10, suggesting that slightly longer training provided additional benefit for this variant. For Stable Diffusion XL 1.0-base, however, the optimal model emerged earlier, at epoch 4, demonstrating that the more expressive architecture of SDXL can converge effectively in fewer epochs.

To qualitatively compare the capabilities of each model, we selected a set of 15 randomly selected prompts from the validation set and the corresponding generated images using the best-performing checkpoints for each model. These generations are presented in Figures 9, 11, and 13. To further assess the impact of fine-tuning, we also generated images for the same prompts using the baseline (pretrained, non-fine-tuned) versions of each respective model (see Figures 10, 12, and 14). This comparison allows for a direct visual evaluation of how our fine-tuning process influenced the models' output.

The differences are immediately apparent. Across all prompts, the fine-tuned models consistently produced images with improved coherence, facial structure, and alignment with the textual descriptions. Notably, the pose and overall structure of the generated faces also improved significantly, with our trained models almost always returning correctly framed portrait-style images that closely matched the intended viewpoint. In particular, the SDXL 1.0-base model demonstrated notable advancements in fine-grained detail and prompt fidelity, reinforcing the quantitative observation that it outperformed the other variants.

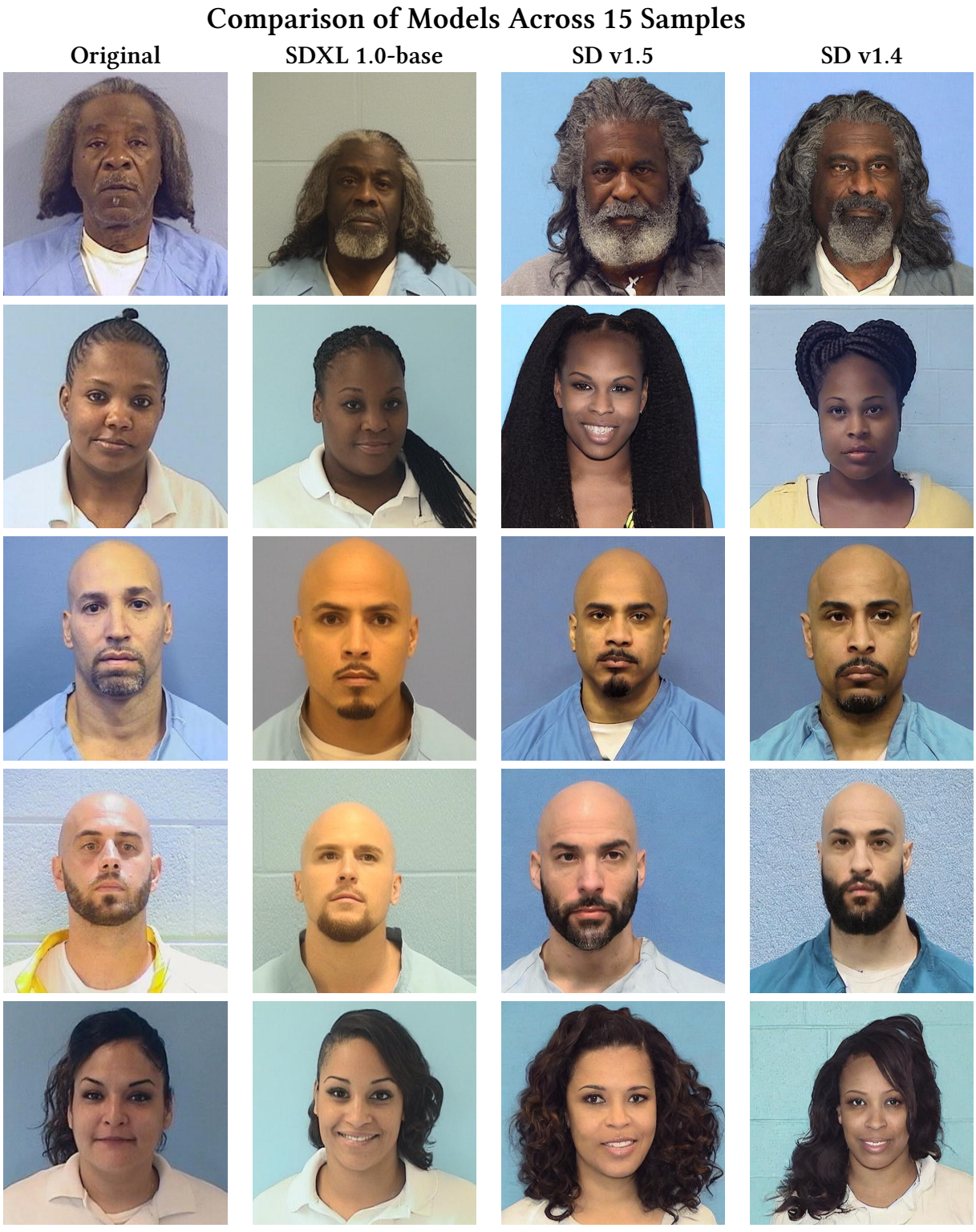


Figure 9: Qualitative comparison of generated faces across models (Validation Samples – Set 1). Images are arranged from left to right in the following order: Original image, SDXL 1.0-base, Stable Diffusion v1.5, and Stable Diffusion v1.4 generations, respectively.

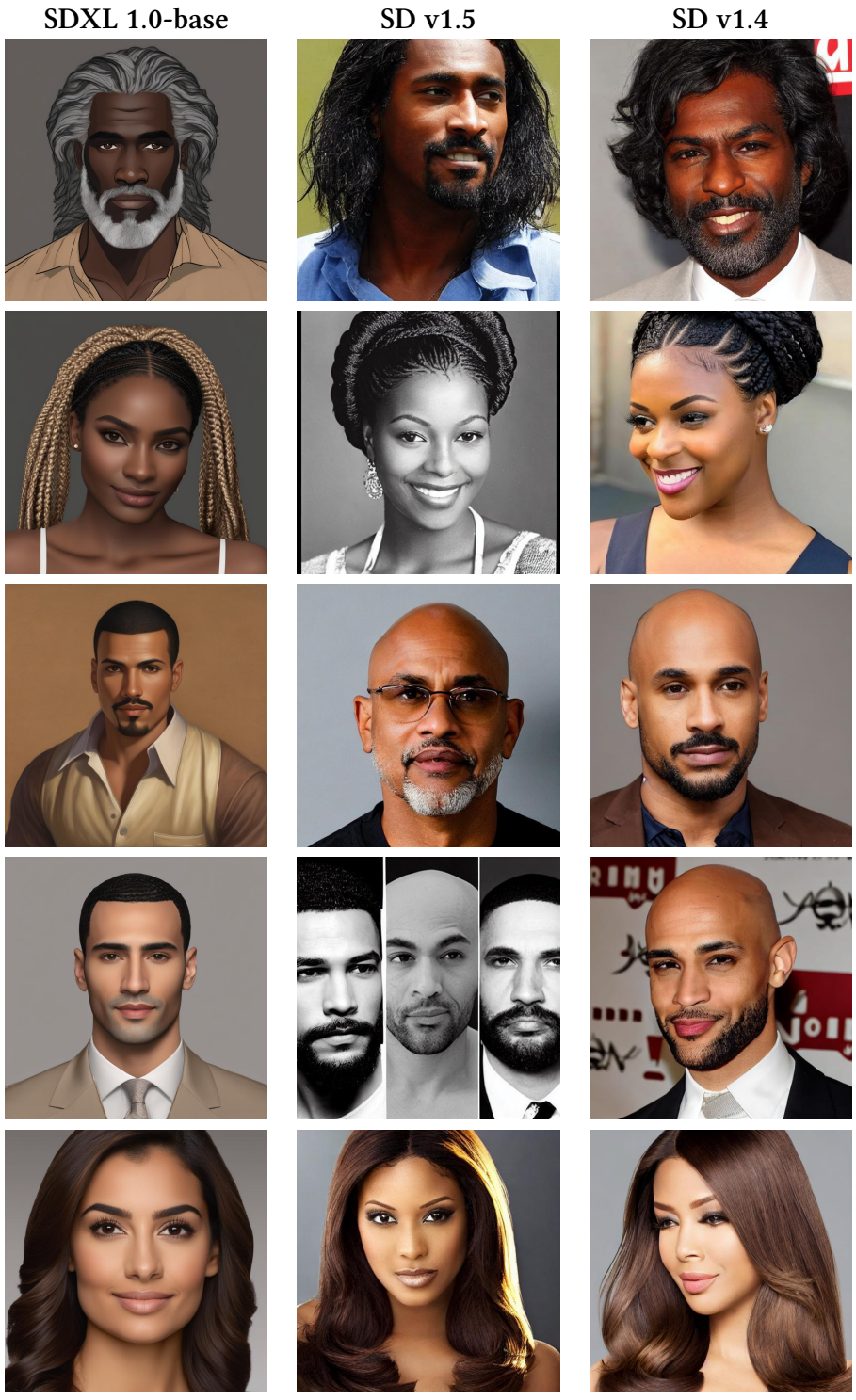


Figure 10: Baseline generations from SDXL 1.0-base, Stable Diffusion v1.5 and Stable Diffusion v1.4 with the same prompts used for generation in Figure 9. The images were generated using the original, non-fine-tuned models to provide a direct comparison with the fine-tuned outputs. Each row corresponds to the same row in Figure 9.

Figure 11: Qualitative comparison of generated faces across models (Validation Samples – Set 2). Images are arranged from left to right in the following order: Original image, SDXL 1.0-base, Stable Diffusion v1.5, and Stable Diffusion v1.4 generations, respectively.

Figure 12: Baseline generations from SDXL 1.0-base, Stable Diffusion v1.5 and Stable Diffusion v1.4 with the same prompts used for generation in Figure 11. The images were generated using the original, non-fine-tuned models to provide a direct comparison with the fine-tuned outputs. Each row corresponds to the same row in Figure 11.

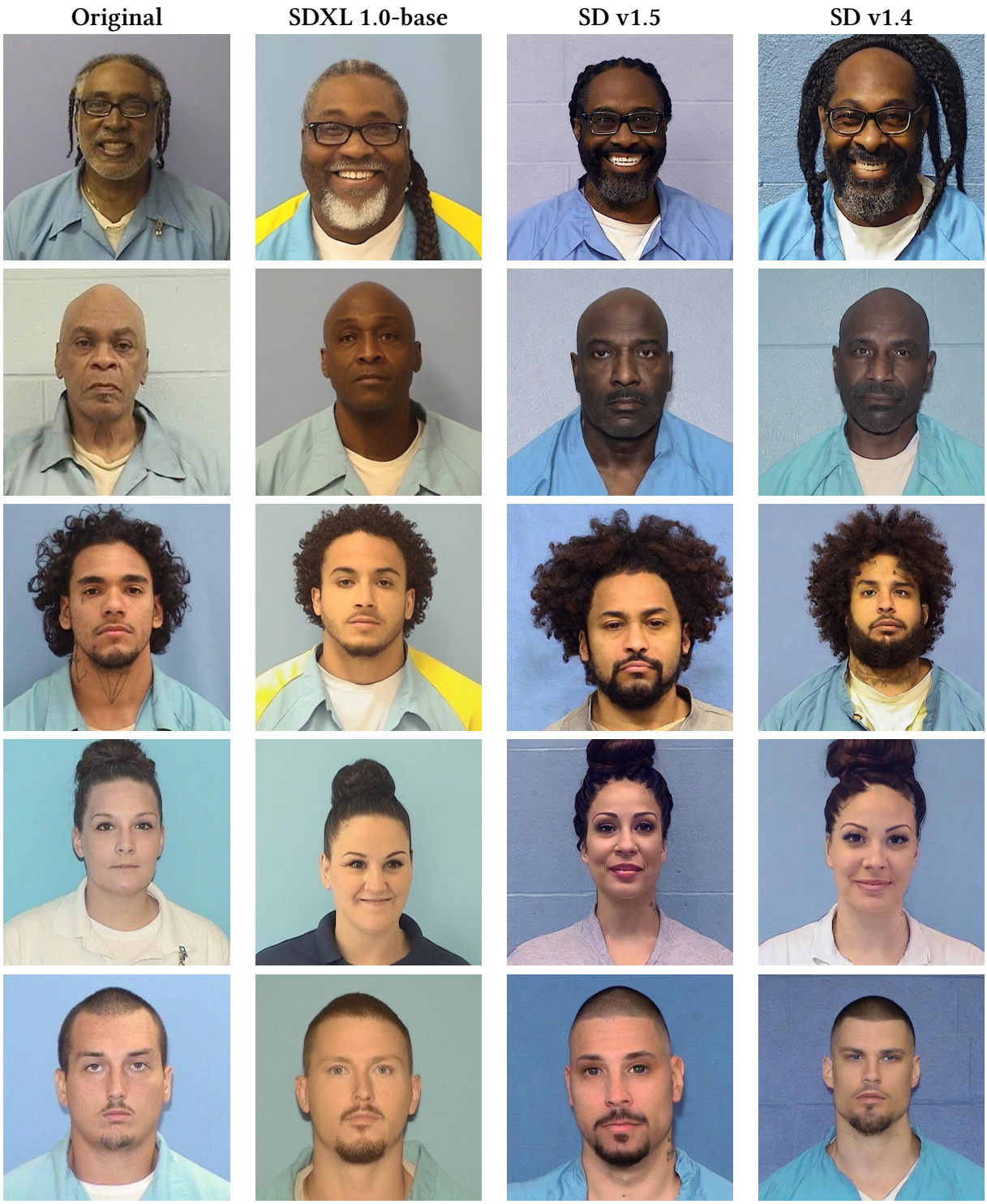


Figure 13: Qualitative comparison of generated faces across models (Validation Samples – Set 3). Images are arranged from left to right in the following order: Original image, SDXL 1.0-base, Stable Diffusion v1.5, and Stable Diffusion v1.4 generations, respectively.

Figure 14: Baseline generations from SDXL 1.0-base, Stable Diffusion v1.5 and Stable Diffusion v1.4 with the same prompts used for generation in Figure 13. The images were generated using the original, non-fine-tuned models to provide a direct comparison with the fine-tuned outputs. Each row corresponds to the same row in Figure 13.

5.2 Best Model

To further validate the performance of our selected model, we conducted a final evaluation using the held-out test set. For this purpose, we used the best-performing checkpoint identified during validation, Stable Diffusion XL 1.0-base, trained for 4 epochs. The test set prompts were passed through this model to generate images, and the same evaluation measures were computed to assess the model's generalization.

During this final evaluation, we slightly adjusted the inference configuration to enhance generation quality, increasing the number of denoising steps `num_inference_steps` to 40 while keeping the `guidance_scale` fixed at 7.5. This adjustment aimed to refine image quality without altering the core behavior of the model.

Table 5: Average performance measures on the test set using Stable Diffusion XL 1.0-base.

Measure	**Score**
CLIP Cosine Similarity	0.8242
LPIPS Distance	0.4833
SSIM	0.5134

The results in Table 5 demonstrate a strong alignment with the validation measures observed earlier, indicating that the performance of the selected model remained stable across datasets. The minimal fluctuation between the validation and test performance measures supports the robustness of our model selection process.

In addition to averaging out evaluation measures across the test set, we also computed a composite score (1) to better capture overall generation quality.

$$\text{Composite Score} = \frac{(1 - \text{LPIPS}) + \text{CLIP} + \text{SSIM}}{3} \tag{1}$$

Using this aggregated measure, we selected 7 representative test samples with higher scores, shown in Figure 15, which highlight the model's ability to generate coherent, prompt-aligned, and perceptually high-quality images.

These results confirm that fine-tuning on a structured, task-specific dataset can significantly enhance the generative capabilities of pretrained diffusion models, even when computational resources are constrained.

Figure 15: Comparison of the best-performing model (SDXL 1.0-base fine-tuned, 4 epochs) and the original images, showcasing high visual fidelity and alignment.

6 Conclusion

This project investigated the use of modern text-to-image diffusion models for generating realistic suspect faces from natural language descriptions. We developed a complete pipeline, ranging from dataset curation and annotation to prompt compression and model fine-tuning, that enables scalable suspect face synthesis in investigative settings.

Our experiments with Stable Diffusion v1.4, v1.5, and SDXL 1.0-base demonstrated that meaningful improvements in generation quality can be achieved even with limited hardware and training resources. Validation results confirmed that SDXL 1.0-base performed best in terms of semantic alignment and perceptual realism, especially at earlier epochs. This suggests that, for the specific task of forensic face generation, lightweight fine-tuning over a few epochs can be efficient.

Although SDXL 1.0-base outperformed the other models overall, as initially expected due to its architectural sophistication, we also observed significant improvements in both Stable Diffusion v1.4 and v1.5 when comparing their baseline outputs to those generated after fine-tuning. Thus, our findings strongly support the original thesis that fine-tuning pretrained text-to-image models on a small, task-specific dataset can meaningfully improve generation quality in specialized applications.

Despite the overall success of the pipeline, several limitations were observed. First, the 77-token limit imposed by CLIP restricted the amount of information that could be used to condition generation, occasionally forcing the omission of important details. Second, generation quality appeared uneven across instances. Some faces were generated with remarkable accuracy, while others were less faithful. These discrepancies are likely influenced by biases inherited from the pretrained models, rather than from the structure of our dataset, which was carefully balanced across gender, age and race. In particular, the models exhibited weaker performance when generating female faces, suggesting underrepresentation or less consistent patterns in the model's original training distribution.

Overall, this work shows that suspect face generation from text is feasible with current diffusion models, but further progress will require improvements in prompt encoding capacity, mitigation of pretraining biases, and techniques that explicitly address variation across demographic groups.

7 Further Work

Future work may explore integrating higher-capacity encoders and adapting the model architecture to better accommodate longer and more detailed descriptive prompts, addressing the limitations introduced by CLIP-based embedding compression. We would also try using more complex generative models, which would, in their place, require higher-resolution datasets to fully leverage their capabilities. Progress in generation quality would most definitely benefit from such datasets, as our current input constraint limited the fine-grained detail achievable in output images. Advancing these directions would require access to stronger hardware to support training with greater visual fidelity.

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-diffusion-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://huggingface.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 Transactions 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-patch32. 2021.

[22] O. Patashnik et al.: "Styleclip: Text-driven manipulation of stylegan imagery." In: *Proceedings 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 Synthesis*. 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 supervision." In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[26] S. Reed et al.: "Generative adversarial text to image synthesis." In: *International conference 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 generative 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 generative adversarial networks." In: *Proceedings of the IEEE international conference on computer 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 Appearance Control." In: *IEEE Transactions on Visualization and Computer Graphics* 30 (9) (2024): pp. 6481–6492.