Addressing the Representation Gap: A Study on Synthetic Face Generation for Underrepresented Global South Demographics

Anonymous authors
Paper under double-blind review

Abstract

Existing benchmark face image datasets are primarily scraped from the web and exhibit limited representation of the Global South. Text-to-Image (T2I) diffusion models are increasingly used to augment these datasets with synthetic data. Due to the inherent subjectivity in human-designed prompts and poor model alignment, it is challenging to generate realistic, non-stereotypical, high-fidelity images; consequently, these images often have limited utility in practice, particularly for auditing or debiasing face recognition systems. In this work, we introduce a novel framework – SemCLoRA, that first aligns VLMs across multiple curated face image datasets to obtain more relevant attribute-rich prompts, followed by a novel "trajectory-sensitive" contrastive learning approach on text-to-image diffusion models to ensure greater prompt-image alignment. We evaluate the approach on two curated seed datasets representing different forms of visibility and underrepresentation: sportspersons from around the world (CheSS) and individuals from rural India (PARI). On both datasets, we observe lower FID and face-matching accuracy, indicating the efficacy of our framework in generating diverse, realistic images that are difficult to match to the training seed set. We further assess the utility of the synthetic images in an auditing & debiasing setting by fine-tuning a face recognition system for binary gender classification. We observe improved accuracy across eight datasets compared to a zero-shot scenario, with gains of up to $\sim 22\%$, while simultaneously reducing the disparity in accuracy between the two binary gender groups by 51%. A user study with 57 participants reveals that humans have difficulty identifying synthetic images with an uncertainty rate of $\sim 15\%$. We conclude with a discussion of responsible-use guidelines and the policy and governance implications of using synthetic data for auditing face recognition systems.

1 Introduction

The World Wide Web has long served as the training dataset for modern data-hungry AI systems. A wide variety of deep learning-based facial analysis systems rely heavily on high-quality, realistic, and diverse facial image data primarily sourced from the web. This dependency on the web deeply embeds and propagates existing demographic, geographical, and sociopolitical inequalities into such automated systems (Scheuerman et al., 2023). Thus, a particular demographic's online visibility determines what type of faces are measurable, recognizable, and worthy of being accounted for. Such systems support multiple tasks ranging from face recognition (Zhao et al., 2003) to facial attribute analysis (Zheng et al., 2020) and affect recognition (Mellouk & Handouzi, 2020). These tasks form the foundation of multiple web-mediated applications in security (Verma et al., 2020), social media (Facebook, 2024), healthcare (Raposo, 2024) and surveillance (Imran et al., 2024).

However, the over-reliance on large-scale "in-the-wild *real*" face datasets (Kemelmacher-Shlizerman et al., 2016) has led to the creation and deployment of systems associated with documented harms (Buolamwini & Gebru, 2018; Jaiswal et al., 2022; Ghosh et al., 2024b). Specifically, current datasets suffer from– **(a)** ***representation gap***: Historically, web data has been overtly skewed toward the Global North-centric (Dados &

Connell, 2012) Western demographic, while the billions of Global South individuals have remained digitally invisible (Jaiswal et al., 2024). Thus, face recognition systems (FRSs) trained on these datasets exhibit systematic performance disparities across marginalized identities (Buolamwini & Gebru, 2018; Raji et al., 2020; Jaiswal et al., 2022), **(b)** ***presence of toxic, non-consensual and harmful content***: Existing web-scraped datasets have been found to contain very toxic and harmful content (Birhane et al., 2023) of private individuals, often scraped without their consent (Spitzley & Gesle, 2024). This impacts the outputs of generative models trained on such datasets (Ghosh et al., 2024b). More recent work has attempted to address both (a) through more representative datasets (Buolamwini & Gebru, 2018; Raji et al., 2020; Jaiswal et al., 2024) and (b) through consensual data collection (Ma et al., 2015; Lakshmi et al., 2021), but such efforts are few and far between.

While both these gaps may seem inhibitive, a potential solution involves the use of synthetically generated images. In fact, industry practitioners have highlighted the importance of synthetic data in addressing data scarcity and evaluating generative AI systems (Kapania et al., 2025). Such synthetic face images can be created using deep generative models, such as GANs (Goodfellow et al., 2014), VAEs (Kingma & Welling, 2014) and more recently, diffusion-based models (DM) (Rombach et al., 2022). The goal is to use these synthetic images to supplement or even replace real data in training and testing pipelines. While some of these approaches have led to impressive outcomes, they face many challenges– **(a)** ***Generative bias***: The training data is curated from existing benchmarks, and the resulting outputs are therefore biased (Luccioni et al., 2023), especially against vulnerable populations in the Global South (Ghosh et al., 2024b). This may lead to amplification of stereotypes or erasure of underrepresented identities. **(b)** ***Semantic drift***: Human-generated prompts are highly subjective and difficult to scale, leading to a lack of fine-grained control over the generation process. Such ill-crafted prompts can also be a source of generative bias in DMs (Skilton & Cardinal, 2024; Kuchlous et al., 2024). Advanced ***T2I*** (text-to-image) diffusion models (Rombach et al., 2022) address (b) to a large extent (Papa et al., 2023; Xiang et al., 2024), but still suffer from data leakage (Carlini et al., 2023) and, low diversity (Ghosh et al., 2024a) and large-scale biases (Luccioni et al., 2023). Hence, current approaches seldom serve more than a casual or hobbyist use case.

Thus, we identify an underexplored, albeit critical gap in existing work– *is it possible to leverage generative AI to create fair, diverse, and realistic synthetic face datasets for web-scale applications in the Global South, without scraping the entire Web?*. In this work, we propose a novel framework – SemCLoRA to bridge this gap by generating synthetic data representative of faces across the world, with a particular focus on the often-invisibilized Global South. We study whether carefully constrained synthetic data can help audit and debias platforms without indulging in harmful deployment practices already prevalent (Whitney & Norman, 2024). We now describe our main contributions.

(1) Democratizing representation in datasets: To create synthetic face images using our framework, we need a seed set of real face images. We curate two such distinct sets, at different levels of geographical granularity, representing the global richness and heterogeneity in human face images.

• *The globally visible*: We sample images of sportspersons, in particular *chess players* (dataset: **CheSS**) from 160 countries to establish the generalizability capability of our framework.

• *The Global South's invisible majority*: We sample images from the People's Archive of Rural India (**PARI**) (Sainath, 2020)[1], which is very rich in terms of geography, facial structure, ethnicities, and aesthetics and serves as a highly representative example of over 900 million Indians living in more than 270 districts of rural India.

(2) Semantic alignment via VLMs and diffusion: To address prompt subjectivity and a lack of control from human inputs, we devise a novel two-stage pipeline. First, we align a vision Language model (VLM) (Liu et al., 2023; Li et al., 2023; Zhu et al., 2024) to generate detailed, *objective* description of real face images. Second, we use these descriptions to fine-tune a Text-to-Image (T2I) diffusion model. We devise a novel *hard negative-based contrastive loss* that performs intermediate semantic supervision to generate realistic, prompt-aligned output images. Our approach prevents semantic drift and ensures high-fidelity in the output. This leads to our first research question – **RQ1. Can an aligned generative pipeline produce high-fidelity, diverse face images that accurately reflect underrepresented demographics?** Through this research question, we aim to understand whether an aligned pipeline can generate truly realistic

[1]available under a CC BY-NC-ND 4.0 license

face images compared to an unaligned one. Experiments reveal that our semantic supervision based model, SemCLoRA, generates realistic and diverse images that are more aligned to input prompts than baselines (see Fig. 3 for examples). SemCLoRA has a $\simeq 47\%$ lower FID compared to IDiffFace (Boutros et al., 2023) on the **CheSS** dataset.

(3) Fairer downstream FRSs: Once the VLM-aligned prompts have been used to generate face images from the T2I diffusion model, it is imperative to evaluate whether these synthetic face images can be used as training data for downstream applications. This brings us to our second research question–
RQ2. Can exclusively synthetic images improve the performance of biased FRSs and reduce the intersectional disparity in their performance? Through this question, we aim to evaluate the effectiveness of our generated images in downstream applications. Our experiments with fine-tuning an FRS using the synthetic images report a significant simultaneous improvement in accuracy *and* reduction in disparity on diverse benchmark datasets. This observation holds true for all seed datasets used to generate these synthetic images, thus demonstrating the generalizability of our approach. The FRS trained on the **PARI** synthetic image set reduces the disparity on benchmark datasets by as much as 51%.

2 Related work

We discuss related literature on face generation and prompt curation in generative models.

Face image generation: GANs (Kammoun et al., 2022) were the earliest architectures used for synthetic face generation, followed by VAEs (Chen & Lu, 2019). More recently, diffusion models (DM) (Ho et al., 2020) have become the de facto platform for image generation tasks. Their use was enhanced with the development of latent DMs (Rombach et al., 2022), which improved speed and operated in the latent space. Both unconditioned (Graikos et al., 2023; Dhariwal & Nichol, 2021; Boutros et al., 2023; Kim et al., 2023) and text-conditioned DMs (Sun et al., 2021; Stypulkowski et al., 2024; Rahman et al., 2024; Melzi et al., 2023; Wang et al., 2024; Baltsou et al., 2024) have been used to generate realistic face images for a multitude of tasks. More recently, text-conditioned DMs have also been made available as ready-to-use web-enabled applications, provided by corporations such as Google (Nano Banana (Google, 2025)), OpenAI (DALL-E 3 (OpenAI, 2023)) and Stability AI (Stable Diffusion (AI, 2025)). While some of these models report exceptional performance, they suffer from some major issues–(a) trained on non-diverse West-centric (Luccioni et al., 2023) and problematic datasets (Birhane et al., 2023) which has led to many generative biases, (b) loss of facial consistency (Kim et al., 2023) due to changes in lighting, expression, pose, (c) poor representation of non-Western communities (Ghosh et al., 2024a;b; Ghosh, 2024). We address concerns (a) and (c) by employing multiple culturally rich, face image datasets – chess sportspersons (**CheSS**) and individuals from rural India (**PARI**) that provide a much-needed Global South perspective to face generation research.

Prompt generation: As discussed above, text-guided diffusion has improved the ability of DMs to generate better and more realistic outputs. Existing research has shown that human-generated prompts fail to produce high-quality or semantically accurate images (Desai & Vasconcelos, 2024). Prompt engineering efforts have helped address these concerns (Hertz et al., 2023; Kim et al., 2022), but these have not focused exclusively on face generation, which requires more nuanced textual inputs. Recently, automated tools, like LLMs (Grattafiori et al., 2024), are being used to generate prompts that can be provided to DMs for realistic face generation (Sun et al., 2021; Wang et al., 2024). We differ from these two works in the following aspects –(a) we do not generate specific gazes (Wang et al., 2024), (b) we do not restrict our generation to celebrity faces (Sun et al., 2021; Simsar et al., 2025), rather focusing on the general population and, (c) we do not generate the prompts completely from LLMs, instead aligning a VLM to generate the description of an already existing face image to ensure that the produced face image is more realistic.

Comparative baselines: We evaluate our framework against two representative works in synthetic face generation: an unconditioned diffusion model, the identity-conditioned IDiffFace (Boutros et al., 2023), and one text-conditioned framework, SDFD (Baltsou et al., 2024). IDiffFace (Boutros et al., 2023) uses identity-conditioning, which enhances intra-class variations, but restricts it from generating under-represented groups not already well-represented in its training data. By relying on datasets predominantly curated from the Global North, the authors risk reinforcing biases. Furthermore, this architecture necessitates prohibitively large-scale training data. Our framework addresses these bottlenecks by utilizing descriptive

Figure 1: Sample images from the three datasets (Top: **CheSS**, Bottom: **PARI**). **PARI** is the most culturally rich dataset with strong South Asian cultural and geographic artefacts in the images.

Table 1: Summary statistics of the two face image datasets.

Dataset	**# Countries**	**# Images**	**M:F**
CheSS	160	1700	M: 1479, F: 221
PARI	1 (270 Districts)	1697	M: 1079, F: 618

textual prompting on a curated, data-efficient Global South-centric dataset, thereby decoupling generation from the biases of massive pre-training sets.
Next, examining the text-conditioned models, we observe that SDFD (Baltsou et al., 2024) only generates face images using rudimentary prompts, without leveraging the semantic details of real face images. Our work differs fundamentally in its objective: rather than using fixed, rudimentary prompts, we devise more detailed descriptions of the seed face images using an aligned VLM, which are then used to fine-tune the downstream diffusion model to generate more realistic face images.

3 Dataset

In this study, we curate two face datasets from around the world, with increasing levels of geographical granularity, to train our framework.

First, we curate a set of face images of chess sportspersons (**CheSS**) from 160 countries worldwide. We follow a similar methodology to Jaiswal et al. (2024), but expand the scope to 160 countries. This dataset, comprising 1700 images, captures diversity in terms of gender, demographics, culture, and age worldwide, enabling generalizable learning for the diffusion model. The dataset has images with a resolution of 200×256px.

Next, we curate an even more granular dataset – a subset of the People's Archive of Rural India (**PARI**) Faces image dataset (Sainath, 2020). The dataset is very rich in terms of gender, age, demographic, cultural and economic diversity, as it has images of people residing in 270+ districts of India. This dataset is one of the only publicly available resources that exhibits the rich facial diversity of a country like India. Note that these faces are also representative of the ethnicities across peninsular India, including Sri Lanka, Nepal, Bhutan, and Bangladesh. The original dataset contains 3213 images, but after filtering out children's images, we are left with 1697 final images. The images in this dataset, unlike the above, are high-resolution.

As diffusion models are known to generate explicit content (Li et al., 2024), to prevent misuse or harmful generation during fine-tuning or downstream tasks, we use a combination of manual and automatic filtering to discard images of children from our datasets. Some images from our final subset are present in Figure 1. We note some of the dataset metadata in Table 1.

4 Semantic supervision of diffusion models

In this section, we present our proposed framework – SemCLoRA that builds on a novel semantic supervision based contrastive loss. We first give a brief overview of latent diffusion models (LDMs) and then describe our novel contrastive objective that enforces semantic supervision during image generation.

Preliminaries: Text-conditioned latent diffusion models (LDMs) (Rombach et al., 2022) operate in compressed latent space instead of the representation space. Given an image $x \in \mathbb{R}^{H \times W \times 3}$, the VAE encoder $\mathcal{E}$ maps it to a latent representation $z_0 = \mathcal{E}(x)$. The forward process in the diffusion model progressively adds Gaussian noise to z_0 over T timesteps according to a fixed schedule β_t. At any given timestep t, the noisy latent z_t can be expressed as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_i)$. During the reverse process (learning step), a neural network ϵ_θ, typically a U-Net architecture, is trained to predict the added noise ϵ conditioned on the timestep t and text condition c. The standard loss is the mean squared error (MSE) between the actual and predicted noise:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,t,c,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right] \tag{2}$$

This loss function is inadequate for tasks like face generation where text prompts are more linguistically complex and the output images need high levels of semantic alignment.

The SemCLoRA objective: We introduce a novel intermediate semantic supervision to address this gap, which intervenes during the reverse process to ensure that the intermediate representations of the synthetic face align with the fine-grained attributes of the text prompt.

Since LDMs operate in latent space, standard contrastive losses, which operate on clean images, cannot be applied directly. Thus, we predict a denoised latent $\hat{z}_0$ at an arbitrary timestep t using the model's current noise prediction module $\epsilon_\theta(z_t, t, c)$. Rearranging Eq. (1), we get:

$$\hat{z}_0 = \frac{z_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(z_t, t, c)}{\sqrt{\bar{\alpha}_t}} \tag{3}$$

We then create an approximate "real" image by projecting this latent back to representation space using the VAE decoder $\mathcal{D} \rightarrow \hat{x}_0 = \mathcal{D}(\hat{z}_0)$. This allows us to apply semantic supervision in an intermediate step, instead of at the very end of generation when the image has already been created and cannot be modified.

To enforce even higher attribute alignment, we create *"hard negatives"* for each prompt using a small LLM. For example, given a positive prompt c^+ : *"Man from India. Features: Oval face, fair skin, contemplative expression* $\cdots$*"*, we perturb each facial feature to curate a hard negative prompt c^- : *"Man from India. Features: Square face, dark skin, smiling expression* $\cdots$*"*, while keeping the sentence structure identical. In addition, for each image in a batch of size bsz, the remaining $bsz - 1$ prompts associated with other samples serve as batch negatives, denoted collectively as c_{batch}.

Final objective: We use a frozen CLIP image encoder (E_I) and a text encoder (E_T) to map the predicted image $\hat{x}_0$ and the prompts to a shared embedding space. Let the image embedding $E_I(\hat{x}_0)$ be denoted as I, the positive prompt embedding $E_T(c^+)$ be denoted as C^+ and the negative prompt embeddings, i.e., $E_T(c^-) \cup E_T(c_{batch})$ be denoted as C^-. The SemCLoRA loss is an adaptation of the standard InfoNCE loss (Oord et al., 2018) and is written as follows.

$$\mathcal{L}_{\text{SemCLoRA}} = -\log \frac{\exp\left(q(I, C^+)/\tau\right)}{\exp(q(I, C^+)/\tau) + \sum_{j=1}^{n} \exp(q(I, C^-)/\tau)} \tag{4}$$

Here $q(u, v)$ is the softmax of the cosine similarity of the vectors x and y. $\mathcal{L}_{\text{SemCLoRA}}$ pulls the image trajectory towards the positive prompt c^+ and away from the hard negative c^- and other batch negatives c_{batch}. The final training objective loss combines the standard diffusion loss (Eq. 2) with the SemCLoRA loss as follows.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda \cdot \mathcal{L}_{\text{SemCLoRA}} \tag{5}$$

where λ is a hyperparameter to balance the impact of semantic alignment.

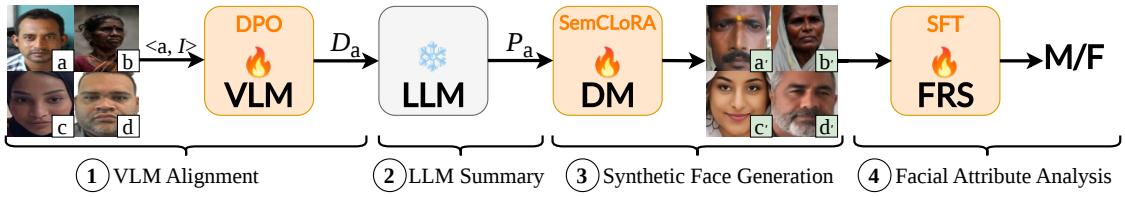

Figure 2: An overview of our proposed framework. (a) Images from a real dataset are provided to a VLM which has been ① aligned using DPO-finetuning, which generates descriptions of these images. (b) These descriptions are passed through a pre-trained frozen LLM to ② summarize and reduce the length of the generated text. (c) The summaries are provided to a T2I DM fine-tuned with the novel SemCLoRA loss to ③ generate synthetic face images. (d) The synthetic face images are used to fine-tune an FRS for ④ facial attribute analysis.

5 Framework & Methodology

Here, we describe the framework and experimental methodology that we employ to generate realistic synthetic face images.

5.1 Our framework

We now describe each component of our framework that is joined together in a chain and helps generate realistic, synthetic face images while maintaining diversity. We use three deep generative models viz. a VLM, an LLM and a T2I DM, and one CNN model as part of the pipeline, as shown in Figure 2 . We use the following architectures for each model –

1. VLM: LlaVa-1.5-7B (Liu et al., 2024) is used to generate detailed descriptions of images from the three datasets, using a common instruction (mentioned later) for all the images. Since zero-shot generation does not yield satisfactory results, we perform a DPO (Rafailov et al., 2023) fine-tuning to improve the model performance. The preferred response is calculated using GPT-4.1. This helps align the VLM to generate more descriptive and useful responses.

2. LLM: Llama-3-8B (Grattafiori et al., 2024) is used to summarize the generated description, which is then provided to the DM as a prompt. This summary ensures that – (1) the final prompt has $\leq$ *77 tokens* so as to fit within the DM's prompt length requirement (Zhang et al., 2024) and, (2) the final prompt *retains all important parts-of-speech* describing the person's face. This model is used without modification for a zero-shot generation.

3. T2I DM: Stable Diffusion-2 (Rombach et al., 2022) is used to generate the highly realistic face images. The model, by default, performs very poorly, generating unrealistic face images (Ghosh et al., 2024a). Hence we fine-tune it using a novel hard negative based contrastive loss in addition to the standard diffusion loss via LoRA adapters (Hu et al., 2022). Some example generations are shown in Figure 3.

4. CNN: DeepFace (Serengil & Ozpinar, 2021), a pre-trained VGG16-based network is used for downstream analysis using the generated face images. We perform SFT using the generated face images and test on standard benchmark datasets which show the effectiveness of using the synthetic face images for facial attribute analysis.

5.2 Methodology

We now describe the methodology of our fine-tuning, and experimentation, the baselines, and the ablations against which we compare. The four major steps– ①, ②, ③, ④ are as shown in Figure 2.

5.2.1 Fine-tuning methodology

We fine-tune the VLM, T2I DM and FRS as follows. The process is the same for both datasets, viz. **CheSS** and **PARI**.

DPO training of the VLM: A small set of images (individually for each dataset) are used for the DPO training as follows.
(a) $\mathbf{IMG}_{DPO} \subset \mathbf{IMG}$, $|\mathbf{IMG}_{DPO}| = 300$
(b) $\mathbf{IMG}_{DPO}$ is chosen for aligning the VLM, to learn the *alignment* function $\mathbf{f}_1()$ using DPO technique.
(c) For every image $a \in \mathbf{IMG}_{DPO}$, the instruction is $\mathcal{I}$, preferred response is $\mathcal{R}_a^p$ and rejected response is $\mathcal{R}_a^r$, given to the model as– $\langle \mathcal{I}, \mathcal{R}_a^p, \mathcal{R}_a^r \rangle$.
(d) $\mathcal{I}$ = *"Generate a detailed description of the subject's face. Describe the facial shape (e.g., angular, oval), skin tone and texture (e.g., smooth, weathered), key features (eyes, nose, mouth) and expression (e.g., contemplative, lively), and any distinctive details (facial hair, scars) using vivid, specific adjectives."*
(e) We use GPT-4.1 to generate the preferred responses, and two annotators manually verified 30% of the post-fine-tuning generated responses to ensure they comply with our requirements. 80% of the responses are deemed acceptable with a Cohen's-κ of 0.74.

LoRA fine-tuning of the DM: A separate subset of images is used for diffusion model fine-tuning as follows.
(a) $\mathbf{Img}_{DM} \subset \mathbf{Img}$, $|\mathbf{Img}_{LoRA}| = 800$
(b) $\mathbf{Img}_{DM} \bigcap \mathbf{Img}_{DPO} = \varnothing$
(c) Each image $a \in \mathbf{Img}_{DM}$ is passed through $\mathbf{f}_1()$ to generate a description of a.
(d) Summary of the description – P_a generated by the LLM ($\mathbf{f}_2()$) for each description in (d).
(e) For every summary description – P_a generated by $\mathbf{f}_2()$ for images $a \in \mathbf{Img}_{DM}$, the T2I DM is fine-tuned to learn $\mathbf{f}_3()$ using the total loss described in Eq. 5 above.
(f) $\mathbf{f}_3()$ to generate realistic synthetic face images.
(g) There is a 1-to-1 mapping between images in $\mathbf{Img}_{DM}$ and the synthetically generated images.

Synthetic test image generation: Once the SemCLoRA based fine-tuning is complete, we generate a fresh set of synthetic outputs (call it $\mathbf{S} - \mathbf{Img}$) for the remaining images of the $\mathbf{Img}$ dataset (call it $\mathbf{Img}_{Test}$) described as follows.
(a) $\mathbf{Img}_{Test} \subset \mathbf{Img}$, $|\mathbf{Img}_{Test}| \sim 600$
(b) $\mathbf{Img}_{Test} \bigcap \mathbf{Img}_{DM} \bigcap \mathbf{Img}_{DPO} = \varnothing$
(c) Set of synthetic images $\mathbf{S} - \mathbf{Img}$ is generated from each item of $\mathbf{Img}_{Test}$. $|\mathbf{S} - \mathbf{Img}| = |\mathbf{Img}_{Test}|$.
Fine-tuning the FRS: Finally, we shortlist a subset of images $\mathbf{S} - \mathbf{Img}_{train} \subset \mathbf{S} - \mathbf{Img}$ to fine-tune the FRS using SFT to learn $\mathbf{f}_4()$ for the task of gender classification.
(a) Each training example is an image $a' \in \mathbf{S} - \mathbf{Img}_{train}$, given as $\langle a', g \rangle$, where $g \in \{M, F\}$.
(b) We manually verify the gender of each generated image. We test the efficacy of fine-tuning on a test set $\mathbf{S} - \mathbf{Img}_{test}$. The following mathematical formula encapsulates the entire sequence of steps during test experimentation– $\mathbf{f}_4(\mathbf{f}_3(\mathbf{f}_2(\mathbf{f}_1(\langle a, \mathcal{I} \rangle))))$.

5.2.2 Baselines and metrics for comparison.

We compare our framework against two baselines – IDiffFace(Boutros et al., 2023) and SDFD(Baltsou et al., 2024) (described in Section 2) and one competitive ablation setup – **B**: DPO-aligned VLM is used to generate prompts and the DM is fine-tuned using a standard LoRA operation on the prompts and corresponding images.
Synthetic image quality (RQ1.): We evaluate the quality of the synthetic face images along three dimensions – (a) how realistic the synthetic images are, using the FID (Heusel et al., 2017) metric which compares the distribution of the generated images with the distribution of a set of real images (rather than comparing individual image pairs) in terms of the mean and covariance statistics, (b) how sincerely do the output images align with the input prompts, using the TIFA (Hu et al., 2023) metric, and (c) whether the source images ($\mathbf{Img}$) from whose descriptions the synthetic images ($\mathbf{S} - \mathbf{Img}$) are generated, can be matched with the synthetic images in terms of face-matching accuracy by comparing embeddings using cosine similarity and cryptographic hashes.
Downstream task efficacy (RQ2.): Next, we compare the accuracy and disparity in gender classification in the fine-tuned FRS against a baseline zero-shot setup on three benchmark datasets – **CelebSET** (Raji et al., 2020) and **CFD** (Ma et al., 2015). This experiment compares the effectiveness of fine-tuning the

Figure 3: Synthetic face images generated using SemCLoRA for both seed datasets – **CheSS** (cols 1-6) and **PARI** (cols 7-12).

Table 2: FID, TIFA scores, and face-matching accuracy for **CheSS** and **PARI** for the baselines, standard LoRA model, and our proposed SemCLoRA. The SemCLoRA model provides the best overall trade-off between image realism, semantic alignment, and reduced identifiability. All values are averages over 3 generated synthetic image sets.

Dataset	**Model**	**FID Score** ↓	**TIFA** ↑	**Face match %** ↓	
				Cosine	**Crypt**
None	SDFD	ϕ	0.529	ϕ	ϕ
CheSS	IDiffFace	109.84	ϕ	66	27.5
	Standard LoRA (**B**)	58.74	0.523	11.43	2.17
	SemCLoRA	**58.53**	**0.533**	**10.50**	**1.33**
PARI	IDiffFace	90.32	ϕ	53.40	13.23
	Standard LoRA (**B**)	72.60	0.415	24.97	1
	SemCLoRA	**69.15**	**0.428**	**18.57**	**0.95**

FRS on **S** − **Img** and then evaluating gender classification performance on other benchmark datasets, versus simply testing on a pre-trained model without any fine-tuning.

5.2.3 Additional manual evaluation

To evaluate the quality of the synthetic face images generated by our method, we conducted a user study using an online survey. Participants were shown *20 pairs of face images* and asked to identify which image in each pair was synthetically generated or to choose a third option if they were unsure. We curate 10 pairs each from **CheSS** and **PARI**, randomly distributed in the form. Among the 10 pairs, 8 pairs consisted of one real and one synthetic image, while 2 control pairs included either both synthetic or both real images (1 pair each). At the end of the survey, participants were asked to provide a brief explanation for their choices.

We conducted the survey using Google Forms[2] and obtained informed consent from all participants at the outset. We recruited 57 participants through snowball sampling, with the following age distribution– 19-25: 32, 26-30: 10, 31-35: 9, 35 & above: 6. Participation was voluntary and did not involve any payment. We did not collect any PII and only present aggregate results here. Participants were advised to complete the survey within 15 minutes.

6 Experiments & Observations

In this section, we discuss the experimental results and our associated observations, segregated by the RQs. presented in Section 1.

6.1 RQ1. Generating realistic images through VLM alignment

In our experimental pipeline, our main goal is to generate realistic synthetic face images. This involves aligning the VLMs to generate effective prompts, that are then used to fine-tune the T2I diffusion model

[2]See Appendix for example questions.

for image generation. Thus, we study the performance of the diffusion model from three aspects – realism (FID (Heusel et al., 2017)), prompt alignment (TIFA (Hu et al., 2023)), and face matching (Serengil & Ozpinar, 2021; Boddeti, 2018). To ensure results are not impacted by the stochasticity of the generation process, each prompt is executed 3 times and all values reported in Table 2 are averages.

(a) Realistic images: We evaluate the realism of the generated images using the Fréchet Inception Distance (FID) (Heusel et al., 2017), a widely used metric for quantifying the similarity between synthetic and real images. Specifically, we compute the FID between the real images in the test set $\mathbf{Img}_{Test}$ and the corresponding synthetic images produced by three models: (i) the baseline IDiffFace (Boutros et al., 2023) fine-tuned on our datasets, (ii) an ablation using DPO-aligned VLM prompts with a standard LoRA fine-tuned T2I diffusion model (**B**), and (iii) our proposed SemCLoRA model. The input prompts for **B** and SemCLoRA models are identical, ensuring that evaluation is limited to the diffusion model.
Table 2 (col 3) summarizes the results. We note that on both datasets, SemCLoRA outperforms both IDiffFace and **B**, with the largest improvement (51.53 FID points) on the **CheSS** dataset. Overall, our findings indicate that the SemCLoRA framework produces synthetic face images that closely follow a real-world distribution, compared to the SOTA baseline IDiffFace and the ablation. This highlights the effectiveness of combining DPO-aligned VLMs with SemCLoRA fine-tuning in producing realistic, high-quality, and diverse face images.
(b) Prompt alignment: We evaluate the text-image alignment of the generated images using TIFA (Hu et al., 2023), a metric that uses LLMs to decompose textual prompt into probes and uses a VQA model to capture more nuanced aspects of the generated image. In col 4 of Table 2 we present the TIFA results. An analysis of the TIFA metric reveals that our DPO-aligned model yields more detailed prompts, which benefit the DM in generating better outputs. Since IDiffFace does not use textual prompts, we do not evaluate it for this metric. While SDFD (Baltsou et al., 2024) has a high TIFA score, SemCLoRA, which has DPO-aligned prompts as the input, outperforms it on the **CheSS** dataset. Next, we similarly note that SemCLoRA generates more faithful images as evidenced by the higher TIFA score compared to **B** on both datasets.
(c) Face matching: We use the Deepface (Serengil & Ozpinar, 2021) and Cryptographic face matching (Boddeti, 2018) facial image processing systems to evaluate the face matching accuracy. Our aim is to achieve a low face matching score, allowing for the efficient decoupling of the generated image from its seed. The results are presented in Table 1 (col 5). SDFD (Baltsou et al., 2024) generates images without any seed prompts and hence cannot be evaluated for this task.
We observe that the identity-conditioned IDiffFace (Boutros et al., 2023) model achieves the highest face matching accuracy, which effectively means that identity of source images leaks into the generated outputs. In contrast, **B** exhibits significantly lower face-matching accuracy, indicating that fine-tuning the T2I model by decoupling it from identity improves the realism while reducing the possibility of synthetic faces being mapped back to their real counterparts. Finally, our SemCLoRA model achieves an even lower face matching accuracy on both datasets. This indicates that SemCLoRA generated faces are realistic and reduce the possibility of synthetic faces being matched to the real images.
Overall, we note that the combination of FID, TIFA and face matching results demonstrates that SemCLoRA is uniquely positioned to create diverse and realistic face images, thus answering RQ1.

6.2 RQ2. Utility of using synthetic images for FRS applications

Having generated the synthetic face images, we now test their usefulness for the real-world task of improving the performance of FRSs. We choose gender classification as our task and fine-tune the DeepFace (Serengil & Ozpinar, 2021) model (based on VGG16 architecture) to improve the accuracy as well as reduce the disparity between the male and female accuracies.

We fine-tune the model on a small subset of synthetic images and compare the change in accuracy against the pre-trained DeepFace. We conduct a similar comparison to examine the disparity. More formally, we define the two metrics as follows– $Acc = \frac{C_M+C_F}{T_M+T_F}$ where C_G: $G \in \{M \text{ (male)}, F \text{ (female)}\}$ is the number of correct predictions for each gender group, and T_G is the number of total data points for each gender group. Similarly, the disparity is formally defined as $|Acc_M - Acc_F| = \frac{C_M}{T_M}$ - $\frac{C_F}{T_F}$. The results are presented in Tables 3 and 4. Recall that the training set is a subset of the generated synthetic images- $\mathbf{S} - \mathbf{Img}_{train}$. We perform this

Table 3: Change in accuracy and disparity between the FRS fine-tuned on the $S - \mathbf{CheSS}_{train}$ dataset and the zero-shot model available off-the-shelf. Our fine-tuned setup results in superior accuracy and lower disparity (-ve sign implies disparity has reduced) for all datasets. **B** refers to the Standard LoRA setup. All values are in %.

Dataset	**Type**	**Zero-Shot**		**FT (B)**		**FT (SemCLoRA)**	
		Acc	Disp	$\Delta_{Acc} \uparrow$	$\Delta_{Disp} \downarrow$	$\Delta_{Acc} \uparrow$	$\Delta_{Disp} \downarrow$
CheSS	Real	78.00	28.00	+8.00	−37.33	+8.00	−32.00
CheSS	**B**	82.67	2.67	+2.67	−10.67	+6.67	−8.00
CheSS	SemCLoRA	76.00	16.00	+8.00	−21.33	+10.00	−20.00
PARI	Real	54.03	69.36	+8.34	−18.82	+7.53	−15.06
PARI	**B**	66.40	48.93	+15.32	−39.25	+17.74	−43.02
PARI	SemCLoRA	67.74	55.91	+19.89	−46.23	+18.28	−43.01
CelebSET	Real	84.17	22.50	+7.98	−17.36	+7.01	−15.14
CFD	Real	80.47	39.06	+6.25	−12.50	+4.17	−8.33

Table 4: Change in accuracy and disparity between the FRS fine-tuned on the $S - \mathbf{PARI}_{train}$ dataset and the zero-shot model available off-the-shelf. Our fine-tuned setup results in superior accuracy and lower disparity (-ve sign implies disparity has reduced) for all datasets. **B** refers to the Standard LoRA setup. All values are in %.

Dataset	**Type**	**Zero-Shot**		**FT (B)**		**FT (SemCLoRA)**	
		Acc	Disp	$\Delta_{Acc} \uparrow$	$\Delta_{Disp} \downarrow$	$\Delta_{Acc} \uparrow$	$\Delta_{Disp} \downarrow$
CheSS	Real	78.00	28.00	+3.33	−14.67	+6.67	−24.00
CheSS	**B**	82.67	2.67	+3.33	−4.00	+6.00	−6.67
CheSS	SemCLoRA	76.00	16.00	+6.67	−8.00	+8.00	−13.33
PARI	Real	54.03	69.36	+3.50	−8.07	+6.18	−14.52
PARI	**B**	66.40	48.93	+15.06	−33.33	+19.09	−47.85
PARI	SemCLoRA	67.74	55.91	+11.83	−29.03	+21.51	−50.53
CelebSET	Real	84.17	22.50	+3.40	−7.92	+5.97	−13.34
CFD	Real	80.47	39.06	+2.86	−5.73	+5.21	−10.41

comparison for two benchmark datasets – **CelebSET** (Raji et al., 2020) and **CFD** (Ma et al., 2015) and on the subset of images in our two datasets which have a 1-to-1 mapping with $\mathbf{S} - \mathbf{Img}_{test}$, referred to as $\mathbf{Img}_{test}^{1-to-1}$. To ensure a fair comparison, we always test on a set of images balanced between both genders. The benchmark test sets are of the following sizes – **CelebSET**: 480 and **CFD**: 128. The test sets for our datasets are of sizes ($|\mathbf{S} - \mathbf{Img}_{test}| = |\mathbf{Img}_{test}^{1-to-1}|$): $S - \mathbf{PARI}_{test}$: 124 images and $S - \mathbf{CheSS}_{test}$: 50 images.

6.2.1 Performance improvement

From Tables 3 and 4, we can see that the zero-shot accuracies on all datasets are below 85%, but interestingly, the accuracies for the synthetic images $\mathbf{S} - \mathbf{Img}_{test}$ are comparable, and in some cases "higher", to their real counterparts. This reinforces our claims in Section 6.1 that the generated images are realistic and diverse, and can serve as good auditing benchmarks for FRSs.

Fine-tuning with $S - \mathbf{CheSS}_{train}$ (Table 3) improves accuracy across all test sets under both generation methods, with peak gains of +19.89% (**B**) and +18.28% (SemCLoRA) on $S - \mathbf{PARI}_{test}$, accompanied by disparity reductions of 46.23% and 43.01% respectively. Here, Standard LoRA and SemCLoRA perform

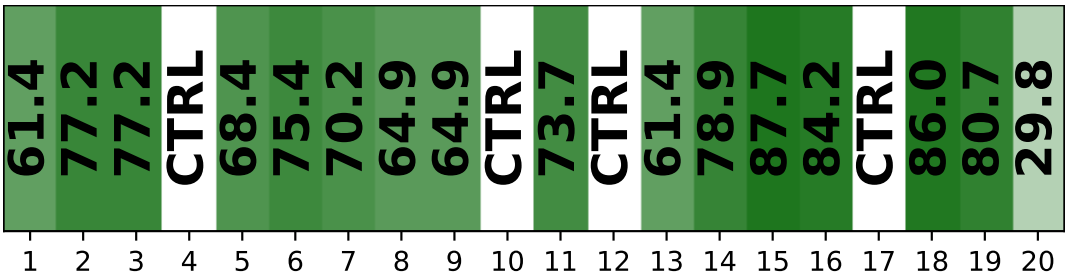

Figure 4: Heatmap with the percentage of respondents who correctly mark the non-control questions in the user survey.

comparably, with Standard LoRA holding a marginal edge on disparity reduction for several real and OOD test sets.

Fine-tuning with $S - \mathbf{PARI}_{train}$ (Table 4), however, reveals a consistent and meaningful advantage for SemCLoRA. Across every test set, SemCLoRA outperforms Standard LoRA (**B**) on both accuracy and disparity: peak accuracy gains reach +21.51% (SemCLoRA) vs. +15.06% (**B**) on $S - \mathbf{PARI}_{test}$, while disparity is reduced by as much as 50.53% (SemCLoRA) vs. 33.33% (**B**) on the same set. This pattern holds without exception across all eight test sets, suggesting that SemCLoRA generates synthetic training data with richer and more balanced demographic coverage than Standard LoRA, particularly when the source dataset itself ($S - \mathbf{PARI}_{train}$) is demographically diverse.

6.2.2 Generalizability on other datasets

While the improvements on $\mathbf{S} - \mathbf{Img}_{test}$ and $\mathbf{Img}_{test}^{\text{1-to-1}}$ are understandable and expected [3] when training on $\mathbf{S} - \mathbf{Img}_{train}$, it is interesting to note the consistent improvements for other synthetic and real datasets, all of which are out-of-distribution (OOD) inputs.

When fine-tuning on $S - \mathbf{CheSS}_{train}$, OOD generalization is robust under both methods: **CelebSET** sees accuracy gains of +7.98% (Standard LoRA) and +7.01% (SemCLoRA) with disparity reductions of 17.36% and 15.14%, while **CFD** improves by +6.25% and +4.17% in accuracy and 12.50% and 8.33% in disparity respectively. When fine-tuning on $S - \mathbf{PARI}_{train}$, the SemCLoRA advantage that is evident in-domain carries through to OOD sets as well: on **CelebSET**, SemCLoRA achieves +5.97% accuracy and −13.34% disparity compared to +3.40% and −7.92% for Standard LoRA; on **CFD**, SemCLoRA yields an accuracy improvement of +5.21% and disparity reduction of 10.41% against +2.86% and −5.73%, respectively for Standard LoRA. This is especially notable given that **CFD** and **CelebSET** contain face images of individuals from a completely different geography—USA—and a different racial distribution (White, Black, Asian, and Latinx). Hence, not only do our models improve performance on face images from the same distribution (different sociocultural and geographic backgrounds) through improved accuracy and reduced disparity, but also perform similarly well for face images from other parts of the world—with SemCLoRA proving the stronger generation method when the source data is sufficiently diverse—thereby authoritatively validating the effectiveness of our framework and answering RQ2.

6.3 Manual evaluation

For the qualitative analysis, we conduct a manual evaluation (as described in Section 5.2.3) through a user survey. In Figure 4, we plot a heatmap of the participant responses for the 20 pairs of images. Cells 4, 10, 12 and 17 had control questions with either both synthetic or both real images. All other cells indicate the % of participants who correctly identified the synthetic image. We observe that the human success rate is 71.38% (better than random chance), which means that even though SemCLoRA does not generate perfectly indistinguishable synthetic images, humans are fooled very often. Infact, the per-question values range between 30-88%, thus proving that there is no synthetic image that *all* participants are able to correctly

[3] See Discussion on FID scores in the previous section

identify. For the control questions, the expected response was *"I cannot determine which one is synthetic."* but the majority of participants chose either the left or the right image (with an accuracy of 26.32%), showing their uncertainty in identifying the synthetic images and effectively proving that our generated images are realistic and indistinguishable.

A one-sample t-test on participant responses yielded statistically significant results ($p < 0.001$). Participants reported uncertainty about whether an image was synthetic in approximately 15% of cases, suggesting that the visual artefacts were sufficiently subtle to prevent reliable discrimination in roughly one out of seven instances. To assess inter-rater reliability, we computed Fleiss' κ and Krippendorff's α, both of which were $\simeq 0.305$, indicating fair agreement among participants. This supports our claim that the generated images effectively confound human observers.

To examine demographic effects, we conducted Kruskal–Wallis and ANOVA tests across age groups. The absence of significant differences suggests that the observed effects generalize across ages, with older participants proving no less susceptible to confusion than younger, more genAI-familiar cohorts (Draxler et al., 2023). Additionally, a per-question χ^2 goodness-of-fit analysis revealed highly non-uniform response distributions $p < 0.001$ for 15 of the 16 active questions, indicating that responses were driven by image content rather than random guessing. Collectively, these findings demonstrate that the generated images exhibit a high degree of realism from both computational and human-evaluation perspectives. Participants were also asked to briefly justify their decisions; a summary of these responses and the associated word cloud are provided in the Appendix.

In the next section, we discuss the implications of the results presented in this section.

7 Discussion

In this study, our primary aim is to develop a framework for generating realistic and diverse synthetic face images (see Figure 3). To do this, we first leverage vision-language models (VLMs) for their multimodal capabilities and align them to generate descriptive, objective textual prompts at scale. These prompts are then used as inputs for diffusion models. We develop a novel contrastive loss that performs intermediate semantic supervision during the denoising step, guiding the model to generate images that are more aligned with the input prompt. Such synthetic face images not only address concerns surrounding toxicity (Birhane et al., 2023) and representation bias (Jaiswal et al., 2024) in existing benchmark datasets but also prove capable of enhancing the performance of downstream applications, such as FRSs, as we demonstrate in our experiments.

Image quality. From our experiments in Sec. 6.1, it is clear that while extremely powerful, both off-the-shelf VLMs and DMs are still incapable of generating realistic outputs simply through human-supplied prompts, especially when the prompts are sufficiently detailed. Thus, we prepare a framework that leverages the power of both these models, coupled with resource-efficient fine-tuning strategies viz. DPO & LoRA to achieve a better result. Our SemCLoRA framework results in the lowest FID score, which is a standard metric that indicates the closeness of the generated set of images to a real world distribution. It is especially important to note that SemCLoRA beats the baseline IDiffFace by as much as $\simeq 52$ points (on the **CheSS** dataset), when both of them are trained on the same dataset. This represents a $\simeq 47\%$ decrease in distributional divergence, effectively bringing the FID down to just over half the baseline value. Similarly, the TIFA (Hu et al., 2023) metric shows the benefit of using DPO aligned prompts over subjective or rudimentary prompts (see SDFD (Stypulkowski et al., 2024)). Fine-tuning the DM with such prompts aligns the image better with the text, especially for the SemCLoRA objective. Moreover, SemCLoRA also provides the best tradeoff between realism and possibility of data leakage from source to generated outputs. Even though IDiffFace-generated images are not realistic, they learn many biometric markers that enable high face match accuracy. Our fine-tuning regime, which attempts to make the output images more faithful to *only* the prompts, retains enough realism, while ensuring that a majority of the synthetic faces cannot be mapped back to their *exact* real counterparts. Thus, our framework can be used to scalably generate synthetic face images that can be reliably, albeit responsibly, used for audits and fine-tuning web-based FRSs.

Downstream application. Finally, we display the effectiveness of our strategy by using the generated images to perform SFT on a deep CNN-based FRS for the task of binary gender classification. The models fine-tuned on $\mathbf{S-Img}_{train}$ result in improved accuracy on respective $\mathbf{S-Img}_{test}$ and $\mathbf{Img}_{test}^{1-to-1}$ datasets as well as when testing for transfer learning (for example, training on $\mathbf{S-PARI}_{train}$ and testing on $\mathbf{S-CheSS}_{test}$. We also evaluate the performance of our fine-tuned models on three popular benchmark datasets composed of faces from different ethnicities and geographies – **CelebSET** (White & Black Hollywood celebrities) and **CFD** (White, Black, Asian and Latinx citizens of USA). There is a marked improvement in accuracy and reduction in disparity in all cases, compared to a zero-shot baseline. Thus, training FRSs on such synthetic and realistic images not only improves model accuracy but also reduces bias, thereby, addressing one of the most pressing social concerns of AI. Existing research has shown that models collapse if recursively trained on synthetic data (Shumailov et al., 2024), which is why we restrict our fine-tuning to exactly one iteration. An in-depth analysis of the potential for model collapse in this scenario is deferred to future work. Finally, we note that since our model of synthetic data generation is based entirely on open-source, small-sized models, it can be deployed on-premises in accordance with local governance and legal regulations, without having to use third-party black-box services.

Manual evaluation. Our user study with 57 participants supplements our quantitative evaluation and shows that real-world users are also unable to confidently disambiguate synthetic images from real ones. Interestingly, there is no synthetic image that *all* participants identify correctly. We do note that the images are not "completely" devoid of visual artefacts, with participants noting "skin", "texture" and "smoothness" as some distinguishing parameters. We attribute this to the significantly larger training set (Rombach et al., 2022), which is known to be biased (Luccioni et al., 2023; Ghosh et al., 2024a). For example, generated faces have the following "quirks" which act as a giveaway to human observers– nose pin on both nostrils and multiple "bindis" on the forehead. All images have a smooth skin texture, irrespective of gender.

Guidelines for responsible use of the synthetic dataset. We note some important guidelines that should be followed for the responsible use of our and other synthetic face image datasets– (1) *Clarity*: The users should know when and why they need the synthetic face data (e.g., auditing, data augmentation, privacy protection, balancing representation). One needs to ensure synthetic data is necessary, and not a shortcut for responsible real-data collection (Whitney & Norman, 2024). We maintain that synthetic data will always be derivative and inherit all ethical, social and demographic constraints of its seed datasets. Thus, participatory frameworks and governance mechanisms need to be established to dismantle existing power asymmetries, especially in data curation and model design. We unequivocally state that our framework should not be viewed as a replacement for consent-based data curation and governance, but as a technical intervention that requires ethical and accountable institutional support. (2) *Temporal audits*: There should be an oversight and governance mechanism in place that ensures regular monitoring for the performance and societal impact of models trained with such synthetic data; periodic audits to monitor emergent bias or misuse over time should be in place. (3) *Sandboxing*: For novel applications, especially where outcomes are not normative/well established, e.g., visual search on e-commerce (Jaiswal & Mukherjee, 2022), synthetic data should be used in sandboxed mode for detailed evaluation to ensure no performance degradation or unfair treatment takes place. We commit to releasing the dataset for research purposes only after proper vetting of the requests.

8 Conclusion

In this work, we proposed a new framework that aligns a VLM on a curated face dataset using DPO fine-tuning to obtain descriptions that can be used as prompts to generate realistic face images. A deep analysis of image realism, prompt alignment and face matching reveals that our synthetically generated images are highly realistic, faithful to the prompt and are not easily mapped back to the source images. We also conducted a user survey with 57 participants to disambiguate between real and synthetic images and observed that humans are confused about the true synthetic image 15% of the time on average. The survey also reveals which features compel the participants to choose an image as synthetic. We plan to address these in future. Finally, to prove the effectiveness of our approach, we fine-tune an FRS for the task of gender classification using the synthetic images and observe not only an improvement of accuracy but also a reduction in disparity, across multiple benchmark datasets. As part of future work, we plan to scale up our

dataset and expand to other FRS tasks. We also plan to study the potential of model collapse and generative biases when using synthetic images.

Broader Impact Statement

In this work, we propose SemCLoRA, a framework for generating a synthetic face dataset to audit and mitigate biases in Face Recognition Systems. Our framework carries a risk of unintended misuse, and we aim to mitigate it through proper vetting of requests for access to the prompts and datasets, and through the transparent release of model weights and a model card. We expect users to adhere to the responsible use guidelines stated above.

References

Stability AI. Stable diffusion. `https://stability.ai/`, 2025. Accessed: 2025-04-04.

Georgia Baltsou, Ioannis Sarridis, Christos Koutlis, and Symeon Papadopoulos. Sdfd: Building a versatile synthetic face image dataset with diverse attributes. In *IEEE FG*, 2024.

Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. *NeurIPS*, 2023.

Vishnu Naresh Boddeti. Secure face matching using fully homomorphic encryption. *IEEE BTAS*, 2018.

Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *IEEE/CVF ICCV*, 2023.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT**, 2018.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023.

Hong-You Chen and Chi-Jen Lu. Nested variance estimating vae/gan for face generation. In *IJCNN*, 2019.

Nour Dados and Raewyn Connell. The global south. *Contexts*, 2012.

Alakh Desai and Nuno Vasconcelos. Improving image synthesis with diffusion-negative sampling. In *Springer ECCV*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.

Fiona Draxler, Daniel Buschek, Mikke Tavast, Perttu Hämäläinen, Albrecht Schmidt, Juhi Kulshrestha, and Robin Welsch. Gender, age, and technology education influence the adoption and appropriation of llms. *arXiv preprint arXiv:2310.06556*, 2023.

Facebook. Testing new ways to combat scams and help restore access to compromised accounts. `https://tinyurl.com/34s47vbj`, 2024. Accessed: 2024-12-12.

Sourojit Ghosh. Interpretations, representations, and stereotypes of caste within text-to-image generators. In *AAAI/ACM AIES*, 2024.

Sourojit Ghosh, Nina Lutz, and Aylin Caliskan. "i don't see myself represented here at all": User experiences of stable diffusion outputs containing representational harms across gender identities and nationalities. In *AAAI/ACM AIES*, 2024a.

Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach. In *AAAI/ACM AIES*, 2024b.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

Google. Nano banana. `https://gemini.google/in/overview/image-generation/`, 2025. Accessed: 2025-11-11.

Alexandros Graikos, Srikar Yellapragada, and Dimitris Samaras. Conditional generation from unconditional diffusion models using denoiser representations. In *BMVC*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *IEEE/CVF ICCV*, 2023.

Ahsan Imran, Riad Ahmed, Md Mehedi Hasan, M Helal Uddin Ahmed, AKM Azad, and Salem A Alyami. Faceengine: A tracking-based framework for real-time face recognition in video surveillance system. *SN Computer Science*, 2024.

Siddharth Jaiswal and Animesh Mukherjee. Marching with the pink parade: Evaluating visual search recommendations for non-binary clothing items. In *ACM CHI Extended Abstracts*, 2022.

Siddharth Jaiswal, Karthikeya Duggirala, Abhisek Dash, and Animesh Mukherjee. Two-face: Adversarial audit of commercial face recognition systems. In *AAAI ICWSM*, 2022.

Siddharth Jaiswal, Animesh Ganai, Abhisek Dash, Saptarshi Ghosh, and Animesh Mukherjee. Breaking the global north stereotype: A global south-centric benchmark dataset for auditing and mitigating biases in facial recognition systems. In *AAAI/ACM AIES*, 2024.

Amina Kammoun, Rim Slama, Hedi Tabia, Tarek Ouni, and Mohmed Abid. Generative adversarial networks for face generation: A survey. *ACM Computing Surveys*, 2022.

Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan. Examining the expanding role of synthetic data throughout the ai development pipeline. In *ACM FAccT*, pp. 45–60, 2025.

Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE CVPR*, 2016.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF CVPR*, 2022.

Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *IEEE/CVF CVPR*, 2023.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Sahil Kuchlous, Marvin Li, and Jeffrey G Wang. Bias begets bias: The impact of biased embeddings on diffusion models. *arXiv preprint arXiv:2409.09569*, 2024.

Anjana Lakshmi, Bernd Wittenbrink, Joshua Correll, and Debbie S Ma. The india face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in Psychology*, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *ACM CCS*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF CVPR*, 2024.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *NeurIPS*, 2023.

Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 2015.

Wafa Mellouk and Wahida Handouzi. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, 2020.

Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations. In *IEEE/CVF ICCV*, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Dall-e 3. `https://openai.com/index/dall-e-3/`, 2023. Accessed: 2025-04-04.

Lorenzo Papa, Lorenzo Faiella, Luca Corvitto, Luca Maiano, and Irene Amerini. On the use of stable diffusion for creating realistic faces: from generation to detection. In *IEEE IWBF*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Abidur Rahman, Faiyaz Al-Mamoon, Mohammad Nazmus Saquib, Mukarrama Tun Fabia, Farhan Bin Bahar, and Md Khalilur Rhaman. Implementation of diffusion model in realistic face generation. In *ICIVC*, 2024.

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *AAAI/ACM AIES*, 2020.

Vera Lúcia Raposo. Facial recognition ai technology in healthcare and the law. In *Research Handbook on Health, AI and the Law*. 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF CVPR*, 2022.

P. Sainath. Pari (people's archive of rural india). `https://ruralindiaonline.org/en/categories/faces/`, 2020. Accessed: 2024-09-09.

Morgan Klaus Scheuerman, Katy Weathington, Tarun Mugunthan, Remi Denton, and Casey Fiesler. From human to data to dataset: Mapping the traceability of human subjects in computer vision datasets. *ACM CSCW*, pp. 1–33, 2023.

Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *IEEE ICEET*, 2021.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 2024.

Enis Simsar, Thomas Hofmann, Federico Tombari, and Pinar Yanardag. Loraclr: Contrastive adaptation for customization of diffusion models. In *IEEE/CVF CVPR*, 2025.

Rachel Skilton and Alison Cardinal. Inclusive prompt engineering: A methodology for hacking biased ai image generation. In *ACM SIGDOC*, 2024.

Maximilian Spitzley and Jenny Gesle. Netherlands: Face-Recognition company Clearview AI fined for violating EU's general data protection regulation. `https://www.loc.gov/item/global-legal-monitor/2024-10-16/netherlands-face-recognition-company-clearview-ai-fined-for-violating-eus-general-data-protection-regulation/`, 2024. Accessed: 2025-04-04.

Michal Stypulkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *IEEE/CVF WACV*, 2024.

Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun. Multi-caption text-to-face synthesis: Dataset and algorithm. In *ACM MM*, 2021.

Rajesh Kumar Verma, Praveen Singh, Chhabi Rani Panigrahi, and Bibudhendu Pati. Iss: intelligent security system using facial recognition. In *ICACIE*. 2020.

Hengfei Wang, Zhongqun Zhang, Yihua Cheng, and Hyung Jin Chang. Textgaze: gaze-controllable face generation with natural language. In *ACM MM*, 2024.

Cedric Deslandes Whitney and Justin Norman. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *ACM FAccT*, pp. 1733–1744, 2024.

Wenlong Xiang, Shuzhen Xu, Cuicui Lv, and Shuo Wang. A customizable face generation method based on stable diffusion model. *IEEE Access*, 2024.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *ECCV*, 2024.

Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM CSUR*, 2003.

Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li, and Ran He. A survey of deep facial attribute analysis. *IJCV*, 2020.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.

A Appendix

A.1 Experimental Details

The Diffusion Model is fine-tuned with SemCLoRA for 200 epochs on a server with 1 H100 80GB GPU. All other fine-tuning and inference are performed on a server with 2 L40 48GB GPUs.

A.2 User Study

Figure 5 presents an example question from the Google Form that is shared with the participants of the User Study. The participants are provided with proper instructions, and consent is obtained before the study commences. We do not collect any PII details and only present aggregate results here.

We ask participants in the user study to briefly explain the reasons behind their survey choices. We plot a word cloud of their responses in Figure 6. Here, we note that the majority of participants consider the

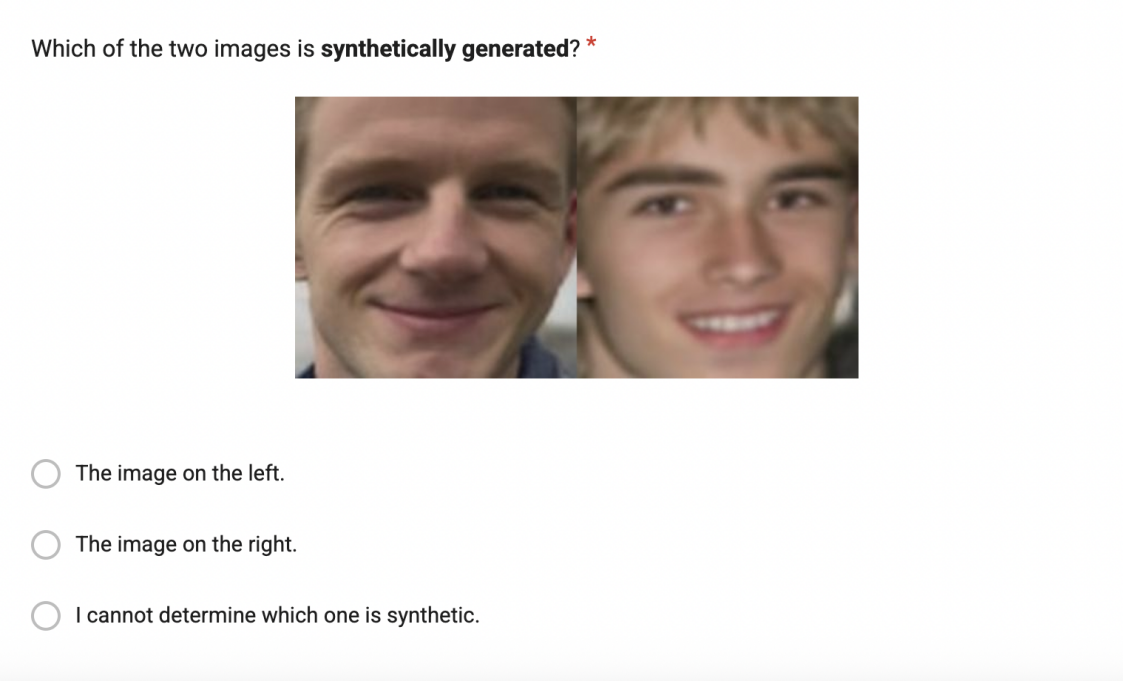

Figure 5: Example question from the Google form used in the user study.

Figure 6: Wordcloud of participant explanations for their choices in the user study.

"texture", *"smooth/-ness"*-of the face, *"lighting"*, and *"background"* as reasons for their choices. These parameters help participants disambiguate between real and synthetic images; however, if the parameters are not strongly visible in the images, they are unable to identify the synthetic image with certainty.