Limitations of Face Image Generation

Harrison Rosenberg*, Shimaa Ahmed*, Guruprasad Ramesh*,
Kassem Fawaz, Ramya Korlakai Vinayak

Electrical and Computer Engineering Department
University of Wisconsin–Madison
hrosenberg@ece.wisc.edu, {ahmed27,viswanathanr,kfawaz}@wisc.edu, ramya@ece.wisc.edu

Abstract

Text-to-image diffusion models have achieved widespread popularity due to their unprecedented image generation capability. In particular, their ability to synthesize and modify human faces has spurred research into using generated face images in both training data augmentation and model performance assessments. In this paper, we study the efficacy and shortcomings of generative models in the context of face generation. Utilizing a combination of qualitative and quantitative measures, including embedding-based metrics and user studies, we present a framework to audit the characteristics of generated faces conditioned on a set of social attributes. We applied our framework on faces generated through state-of-the-art text-to-image diffusion models. We identify several limitations of face image generation that include faithfulness to the text prompt, demographic disparities, and distributional shifts. Furthermore, we present an analytical model that provides insights into how training data selection contributes to the performance of generative models. Our survey data and analytics code can be found online at https://github.com/wi-pi/Limitations_of_Face_Generation

Introduction

Text-to-image (TTI) diffusion models have become popular due to their unprecedented image-generation capability. Taking a textual prompt as input, these models generate realistic images which align with user intentions. Their ability to synthesize and modify human faces has spurred research into using generated face images in training data augmentation and model performance assessments (Dixit et al. 2017; Trabucco et al. 2023). For example, face recognition systems can benefit from synthetic datasets that exhibit more demographic diversity than existing natural datasets (Smith et al. 2023; Friedrich et al. 2023).

This work analyzes the quality of synthetic datasets for facial recognition applications and whether they exhibit demographic disparities. Achieving this objective requires generating a large set of identities belonging to diverse demographic groups and generating multiple (different) faces for each identity. Existing diffusion models are incapable of meeting this objective for two reasons. First, aligning the generated faces with the provided prompt is challenging (Tsaban and Passos 2023; Brack et al. 2023). Second, generating multiple faces with the same identity in a one-shot fashion is typically infeasible (Tsaban and Passos 2023; Brack et al. 2023). Limited research exists in this space. Previous works either optimize the diffusion model to a particular demographic group, generate faces without a notion of identity, or limit their objectives to frequency analysis of the demographics of the generated images (Perera and Patel 2023).

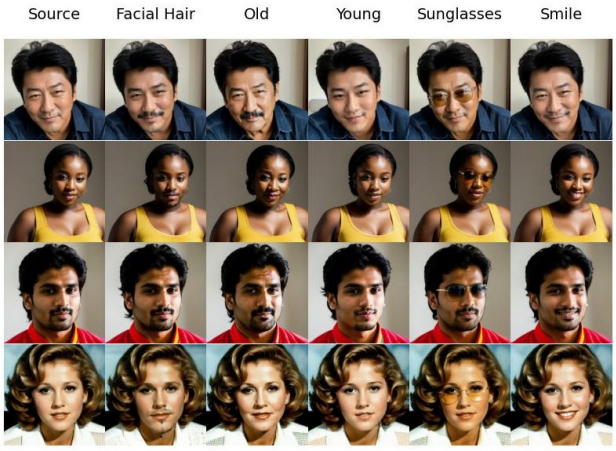


Figure 1: Samples of the non-celebrities dataset using Realism model for four demographic groups: 'East Asian Male', 'Black Female', 'Indian Male', and 'White Female'. 'Source' refers to images generated from Realism using the prompt template. The second to sixth columns show transformed images when an attribute is applied to 'Source' using SEGA.

In this work, we propose a new framework to generate synthetic face images, as shown in fig. 1. Our face-generation pipeline takes as input demographic attributes, applies custom prompts to generate identities for each demographic attribute, and utilizes image editing models (Brack et al. 2023) to generate diversified faces for each identity. The resulting dataset, which we manually verify, resembles a natural face image dataset, albeit demographically balanced by design.

We then apply a three-pronged approach to assess the synthetic face image dataset's quality through face verification (Schroff, Kalenichenko, and Philbin 2015), quantitative

*These authors contributed equally.

quality metrics (Ruiz et al. 2023), and a user study. Our evaluation shows that generated images exhibit demographic disparities in the eyes of face recognition systems. Results from our user studies show disparity in quality of the generated faces for different demographics, with images belonging to majority demographics rated as higher quality. We also study the efficacy of edit correctness metrics built on CLIP and DINO (Brooks, Holynski, and Efros 2023; Zhang et al. 2023). We find these metrics do not correlate with human preferences in facial semantic changes. Research is needed to develop perceptually aligned metrics.

Finally, our findings suggest that methods intended to mitigate bias exhibit demographic disparities in the quality of generated images. Through an analytical model, we show that generative models mimic the demographic disparities in the existing data, typically sampled from the Internet. We also develop sample complexity and data sampling conditions to overcome this inherent bias.

To the best of our knowledge, this paper is the first work that: (1) *provides an end-to-end pipeline*, utilizing a TTI diffusion model, to generate batches of synthetic faces annotated with fine-grained attributes; (2) *evaluates the quality of large-scale batch-generated faces* using a user study; and (3) *assesses the fidelity* of recently proposed TTI quality metrics on face images. Our findings are further discussed in our long-form report[1].

Related Work

In the following, we describe recent works in the context of synthetic face image generation and associated biases.

Synthetic Face Image Generation

TTI diffusion models, such as DALL-E (Ramesh et al. 2021) and Stable Diffusion (Rombach et al. 2021), rely on internal randomness to generate high-quality examples through denoising steps. They employ CLIP (Radford et al. 2021) or its variants as text encoders. Thus, a text prompt is sufficient to control the output of a TTI diffusion model.

Two challenges arise in prompt-based face generation. The first is aligning the generated faces with the provided prompt (Tsaban and Passos 2023; Brack et al. 2023). The second is generating multiple faces belonging to the same identity in a one-shot fashion (Tsaban and Passos 2023; Brack et al. 2023). There exist methods to better control image generation. These methods include segmentation masks and inpainting (Zhang et al. 2023); text-inversion, which learns a text token that corresponds to certain image concept (Gal et al. 2022a); model fine-tuning and embedding optimization (Kawar et al. 2023). While these techniques are generally effective, they are unsuitable for large-scale generation of diverse faces. They often disrupt the fast and natural interface that differentiates TTI diffusion models.

In this work, we aim to analyze the synthetic faces generated by TTI diffusion models. This objective requires generating a large set of identities belonging to diverse demographic groups and generating multiple (different) faces for each identity. We devise a novel pipeline that employs semantic guided attention (SEGA) (Brack et al. 2023), fixed seeds, and specialized prompts. The pipeline, described within the Framework section, depends on neither inversion nor fine-tuning.

Bias in Face Image Generation

Recent works (Friedrich et al. 2023; Seshadri, Singh, and Elazar 2023; Smith et al. 2023) have studied the bias of TTI face generation by analyzing the proportions of demographics in generated images. Seshadri et al. found that generative models amplify the discrepancies in training data (Seshadri, Singh, and Elazar 2023). One example is gender-occupation bias, where Stable Diffusion can generate highly biased face distributions from a gender-neutral prompt about occupations. Friedrich et al. mitigated these biases with a post-processing technique called Fair Diffusion (Friedrich et al. 2023). When the user inputs their prompt, a model detects the potential bias in the prompt and steers the output to a fairer region, leveraging a lookup table of instructions and the semantic image-editing technique SEGA (Brack et al. 2023). Similarly, Smith et al. utilized InstructPix2Pix (Brooks, Holynski, and Efros 2023), an instruction-based image editing model, to edit existing images to be demographically balanced. While this dataset debiasing technique results in finer-grained control over demographic attributes, it introduces a distribution shift between natural and synthetic images. It also stacks the biases of different models (Smith et al. 2023).

Luccioni et al. performed a different bias characterization that relies on correlating model outputs in the embedding space with social attributes (Luccioni et al. 2023). The authors found three popular TTI models are biased toward masculine and white concepts. Struppek et al. studied another source of bias resulting from non-Latin scripts (Struppek, Hintersdorf, and Kersting 2022). They found that using special non-Latin characters better exposes the internal biases of models and proposed using homoglyphs to mitigate this bias. Muñoz et al. analyzed the bias in relatively older face generation models trained on the CelebA and FFHQ datasets (Muñoz et al. 2023). Using quantitative metrics, including demographic frequencies, face recognition verification, and Fréchet inception distance, they found that the generative models are biased.

These conclusions are consistent with earlier GAN literature, where Maluleke et al. found them to generate racially biased distributions of faces (Maluleke et al. 2022). Maluleke et al. went one step further by analyzing the quality of generated faces through a user study, where generated faces from minority groups (e.g., Black) exhibited lower quality.

In summary, existing works focus primarily on frequency analysis to characterize bias in TTI models, propose embedding-based metrics to evaluate the quality of generated images, and utilize synthetic data to mitigate bias. In our work, we characterize the synthetic datasets, showing that methods intended to mitigate bias exhibit demographic disparities in the quality of generated images. We go beyond frequency analysis by rating image quality in a user study. We also utilize the user study results to assess embedding-based metrics in characterizing the quality of the generated images.

[1] https://arxiv.org/abs/2309.07277

Framework

We develop a framework, as depicted in fig. 2, to audit the characteristics of generated face images. This framework consists of choosing the demographic conditions, prompting diffusion models to generate identities according to these conditions, followed by evaluating the generated images both quantitatively and qualitatively.

Notation

We first prescribe the notation used within this paper. There exists a sample space $\mathcal{X} \subseteq \mathbb{R}^d$ and label set $\mathcal{Y}$. A sample $\boldsymbol{x} \in \mathcal{X}$ is a d-dimensional vector. If $\boldsymbol{x}$ is an RGB image, then d equals $3 \times h \times w$, which corresponds to the number of channels multiplied by the number of pixels in the image. In the context of face recognition, we assume each face image $\boldsymbol{x}$ depicts an identity $y \in \mathcal{Y}$. A face recognition model $\mathrm{f} : \mathcal{X} \to \mathcal{Y}$ is trained on a finite dataset $S \in \mathcal{X} \times \mathcal{Y}$. S is drawn i.i.d. from distribution $\mathcal{D}$. Sometimes, when clear from context, S refers to an unlabeled dataset. A metric embedding network $\mathrm{f}_k : \mathcal{X} \to \mathbb{R}^k$ is often internal to deep-network based classifiers. Metric embedding network f_k maps inputs to a k-dimensional embedding space. If two samples have low pairwise distance, they are assumed to be more similar in the associated label space.

We analyze disparities in generative models across social attributes. To analyze these disparities, we examine synthetic face image quality and the performance of generated images in face recognition tasks. A common class of social attributes is demographics. With respect to demographics, we use terminology consistent with Buolamwini and Gebru (Buolamwini and Gebru 2018), a work among the most cited in the space of face recognition fairness. Face images are annotated by sex and ethnicity. Sex annotations are "Male" and "Female." Ethnicity annotations are "White," "Black," "East Asian," and "Indian." The set of demographic groups is denoted as G, where g is a placeholder for a demographic group in G. In this paper, we study eight demographic groups, corresponding to sex-ethnicity combinations.

Given a text prompt p from the space of prompts $\mathcal{P}$, a text-to-image model $\mathrm{h}_q : \mathcal{P} \to \mathcal{X}$ returns the image prescribed by its textual prompt p, where the random seed q is a real number. Because diffusion models have internal randomness, each q generates a different realization of the same prompt p. In our framework, we encode the identity y and its demographic group g in the textual prompt p, and we vary the seed q to generate multiple images of the same identity. We use a fixed set of seeds to ensure the reproducibility of generated images.

Generative Models

We generate synthetic faces using two TTI Diffusion models: the open-source Stable Diffusion v2.1 model by Stability AI and the finetuned Realistic Vision Model[2], hereafter referred to as *SDv2.1* and *Realism*, respectively. We analyze the images generated by these models individually to assess their efficacy in face-generation pipelines.

[2]https://huggingface.co/SG161222/Realistic_Vision_V4.0_noVAE

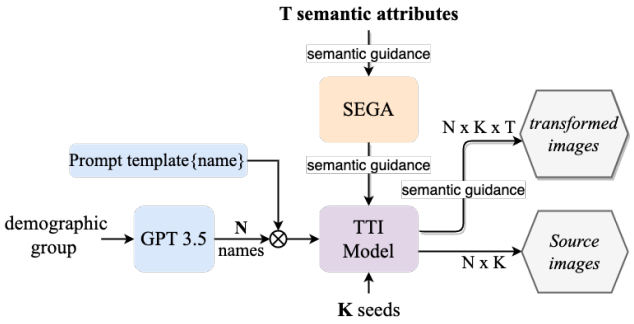


Figure 2: Our data generation pipeline: (1) generate N names (identities) belonging to each demographic group $\mathsf{g} \in \mathsf{G}$ and insert them into the prompt template p, (2) TTI generates K images per identity, using K seeds, (3) SEGA steers the TTI generation to incorporate each of the T semantic attributes.

SDv2.1 is finetuned from the Stable Diffusion v2 (SDv2) checkpoint, which was trained from scratch on a subset of the LAION-5B dataset. SDv2.1's training dataset contains more faces than that of SDv2[3]. Hence, SDv2.1 performs better in generating faces than SDv2. Realism is among the many openly available fine-tuned models from the checkpoints of Stable Diffusion. However, its exact implementation details are not known. We treat both SDv2.1 and Realism as grey-box models. Both models are capable image generators with differing performance characteristics, and our framework is agnostic to their implementation details. The design of the system shown in fig. 2 can be used with any relevant text-to-image generative model to synthesize scalable batches of facial data useful for training data augmentation or as tailored test sets for face recognition applications.

To diversify generated faces, we employ the semantic-guidance image generation technique SEGA (Brack et al. 2023). SEGA steers the TTI model towards generating images that incorporate semantic concepts based on user-provided textual edits while keeping the rest of the image semantics intact, all without the need for fine-tuning the TTI models. This technique proves valuable in creating faces with diverse attributes, such as incorporating sunglasses. Moreover, recent works (Friedrich et al. 2023; Smith et al. 2023) leverage SEGA and similar methods for fair face image generation by introducing demographics as semantic concepts during image generation. Thus, we study the efficacy of incorporating SEGA in the image generation pipeline.

Data Generation Pipeline

To generate our facial datasets, we design a prompt that specifies a demographic group and an identity associated with that group. The prompt guides the model to generate a set of diverse face images for each of these identities.

Identity. We found that when we explicitly mention the demographic group in the prompt, like *an Indian man*, the generated images exhibit limited diversity; i.e. identities look

[3]https://stability.ai/blog/stablediffusion2-1-release7-dec-2022

quite similar. To encourage the generation of more varied identities, we employed *names* as indicators of different identities within demographic groups. We found that TTI models interpret names as proxies for ethnicity and sex, and each name carries a unique identity despite the randomness of the generation process.

For each of the eight demographic groups we study in this paper, we instructed GPT-3.5 to create two separate lists of names—one comprising 'celebrity' names and the other 'non-celebrity' names. For the non-celebrity (celebrity) collection, we generated 20 (30) names per demographic group. The two lists reflect different levels of knowledge within the TTI model: celebrity images are more likely to exist in the training data of TTI, while non-celebrities are more likely to be indirectly learned by the model.

Prompt. We desire prompts that guide the model to generate multiple and diverse face images with user-specified semantics. Including a name within a prompt encodes both identity and demographic information. Trial and error, combined with our user study, led us to the below approach.

For Realism, we experimented with a set of prompts, and we found this template to generate face images of high quality: "A photo of the face of $\{identity\}$." We vary the TTI generator seed to generate multiple images per identity and prompt. We also add a set of *negative* prompts that steer the model away from unrealistic, cartoon, or low-quality image generation. These negative prompts are frequently used in face image generation. For a fair comparison, we use this prompt template along with a set of pre-selected five seeds to generate images of all identities and demographic groups.

For SDv2.1, we observed that the prior template generates images of poor quality on both celebrity and non-celebrity identities. Thus, we expanded the prompt template as follows: "A photo of the face of ($\{identity\}$:2.0). (realistic:2.0). (Face shot only:2.0)." This revised prompt improved the generated image quality of celebrity identities. However, it did not have the same effect on non-celebrity images. Thus, we decided to evaluate only the celebrity identities for the SDv2.1 model.

We manually validated that the generated images from both models contain a face image, different seeds generate diverse images of the same identity, and that identities are distinct and belong to the intended demographic group. We provide further details on the challenges of high-quality face image generation in the long-form report.

Attributes. Using SEGA with Realism and SDv2.1, we induce five attributes to the generated data: 'young', 'old', 'facial hair', 'sunglasses', and 'smile.' The details of SEGA's hyperparameters are in the long-form report. We refer to the images obtained without SEGA as *source images* and with SEGA as *transformed images*. All the synthesized images are of 512×512 resolution. For SDv2.1, to ensure better quality, we generate the images at 768×768 and then downsample them to 512×512. Figure 1 shows a sample of the non-celebrity images synthesized using Realism and SEGA.

In total, we generate 800 source and 4000 transformed non-celebrity images, and we generate 1200 source and 6000 transformed celebrity images per model.

Evaluation Methods

We use three independent evaluation methods to assess the quality of the generated datasets: quantitative metrics, face verification, and user study.

Quantitative Metrics The metrics below are used to evaluate overall quality of the source and the transformed images.

- **Image-Image Metrics:** These are mainly used to verify identity retention under SEGA transformation. CLIP-I and DINO-I measure the cosine similarity between the source and transformed images' CLIP (Radford et al. 2021) and DINO-v2 (Oquab et al. 2023) embeddings, respectively. Higher similarity implies that the identity is preserved.
- **CLIP-directional**: CLIP-directional (Gal et al. 2022b) intends to identify the correctness of the semantic change in the transformed image. It measures the similarity of the change between the embeddings of the source and transformed images and the change between their captions.

Face Verification Face verification accuracy utilizes pairwise face comparisons to measure embedding space quality. The embeddings of two faces depicting the same identity are expected to be close to each other. We analyze face verification performance on Facenet (Schroff, Kalenichenko, and Philbin 2015), a well-studied face recognition network.

Our analysis of face recognition models focuses on verification accuracy. Given two face images $\boldsymbol{x}, \boldsymbol{x}'$, verification accuracy VER is computed as:

$$\begin{aligned}\mathrm{VER}(\boldsymbol{x}, y, \boldsymbol{x}', y') \triangleq\ & \mathbb{1}[y = y'] \cdot \mathbb{1}\left[\rho(\mathrm{f}_k(\boldsymbol{x}), \mathrm{f}_k(\boldsymbol{x}')) < t\right] \\ & + \mathbb{1}[y \neq y'] \cdot \mathbb{1}\left[\rho(\mathrm{f}_k(\boldsymbol{x}), \mathrm{f}_k(\boldsymbol{x}')) \geq t\right]\end{aligned} \tag{1}$$

where $\mathbb{1}$ denotes the indicator function and threshold t is chosen heuristically to minimize false verification rate. Further, y and y' are identities associated with $\boldsymbol{x}$ and $\boldsymbol{x}'$, respectively. We report the average verification accuracy as computed across sets of pairs. Within our evalutation, sets of pairs are constructed so that half of the pairs correspond to the same identity. When analyzing verification accuracy for the user study, we implicitly assume that humans can perfectly distinguish identities of generated faces.

To study the effect of demographics on verification, we report two notions of verification accuracy: same group and any group. For a specified group g, same group verification accuracy refers to the evaluation of VER on lists of pairs in which both images $\boldsymbol{x}, \boldsymbol{x}'$ belong to group g. Any group verification accuracy refers to the evaluation of VER where only each pair's first image $\boldsymbol{x}$ must be in group g.

We utilize the Labeled Faces in the Wild (LFW) dataset as a baseline for natural faces verification. LFW is a canonical dataset for face recognition tasks. The LFW dataset contains 13233 images and a total of 5749 unique identities. Demographic annotations for images in LFW were obtained from the system introduced by Kumar et al. (Kumar et al. 2009).

User Study We conducted a human evaluation of the generated images from both models combined with SEGA. Toward that end, we designed an online Qualtrics survey for each model-identity collection pair, resulting in three surveys:

(SDv2.1 Celebrities, Realism Celebrities, and Realism Non-Celebrities). The surveys are approved by our IRB and are conducted on the Prolific platform.

For each survey, we randomly sampled 15 identities per demographic group, one image per identity; 120 images in total. We paired each source image with its 5 transformed images corresponding to the 5 semantic attributes. This results in 600 source-transformed image pairs per survey. We presented each participant with a set of 21 blocks. Each block shows two images: one source image (without SEGA), and one transformed image (using SEGA), along with the transform instruction used by SEGA. For each block, the participant answers three questions: (1) whether the two images depict the same person, (2) the consistency of the transformed image with the edit instruction on a 5-point scale, and (3) how they rate the quality of the two images on a 5-point Likert scale.

For each study, we recruited 85 participants, and each image pair received three ratings on average. Each participant was compensated $3.5 for their effort, with an average completion time of 15 minutes. The study was distributed evenly to male and female participants. The participants' demographic distribution is discussed in our long-form report.

Evaluation

After generating the datasets, we apply the evaluation methods to analyze the associated demographic discrepancies. Three questions guide this evaluation:

1. *How does face verification on synthetic data compare to natural data and does it exhibit demographic disparities?*
2. *Does the quality of synthetic face images depend on the demographic group?*
3. *Can quantitative metrics replace expensive user studies to assess the quality of synthetic face images?*

Face Verification

Face verification performance is depicted in fig. 3. The figure shows the verification accuracy measured on LFW and synthetic datasets. Across all demographics and datasets, with one exception, we observe that generated faces perform worse than natural faces (LFW). Only in the White demographic does a synthetic dataset, Realism Celebrities, have better face verification performance than natural data. We also observe that for each demographic and dataset pair, same-demographic verification accuracy is often notably less than its any-demographic counterpart. Hence, we conclude that face recognition systems are demographically aware on generated faces in a manner similar to natural faces.

Synthetic Face Image Quality

Table 1 presents the average survey scores in terms of image quality and transformation correctness across all demographics and datasets. The scores suggest that image quality depends on the identity's demographic group. Moreover, SEGA transformations drop the quality of all images, and the drop is also demographic-dependent. We use one-way ANOVA in an attempt to reject null hypotheses of forms:

	Dataset	Demographic group					
		E Asian	Black	Indian	White	Female	Male
M1	D1	4.463	4.401	**4.394**	**4.515**	4.428	4.459
	D2	**4.253**	4.240	**4.045**	4.097	4.166	4.140
	D3	4.121	**4.149**	4.112	**4.039**	4.112	4.099
M2	D1	0.195	0.122	0.187	0.184	**0.244**	**0.098**
	D2	0.190	**0.085**	0.114	**0.592**	0.364	0.140
	D3	0.100	**0.156**	0.123	0.123	0.154	**0.098**
M3	D1	4.020	3.972	**4.144**	**3.945**	3.954	4.092
	D2	3.624	3.367	**3.717**	**2.636**	3.203	3.455
	D3	**3.747**	**3.197**	3.516	3.325	3.492	3.412
M4	D1	87.5	84.6	**90.3**	**83.8**	85.5	87.7
	D2	78.4	69.5	**80.7**	**51.2**	66.0	73.6
	D3	**79.9**	**65.2**	72.3	69.9	72.2	71.8

Table 1: User survey average answers to the following measures: M1: source image quality on a 5-point scale, M2: drop in image quality after SEGA transformation, M3: SEGA transformation correctness on a 5-point scale, and M4: percentage (%) of correct transformation (transformation correctness score $\geq$ 3 out of 5). D1: Realism Non-Celebrities, D2: Realism Celebrities, D3: SDv2.1 Celebrities. Highest and lowest scores are highlighted in bold. E Asian denotes the East Asian demographic group.

Null Hypothesis 1. The per-demographic distributions of source image quality in ⟨Dataset⟩ are identical.

Null Hypothesis 2. The per-demographic distributions of transformed image quality in ⟨Dataset⟩ are identical.

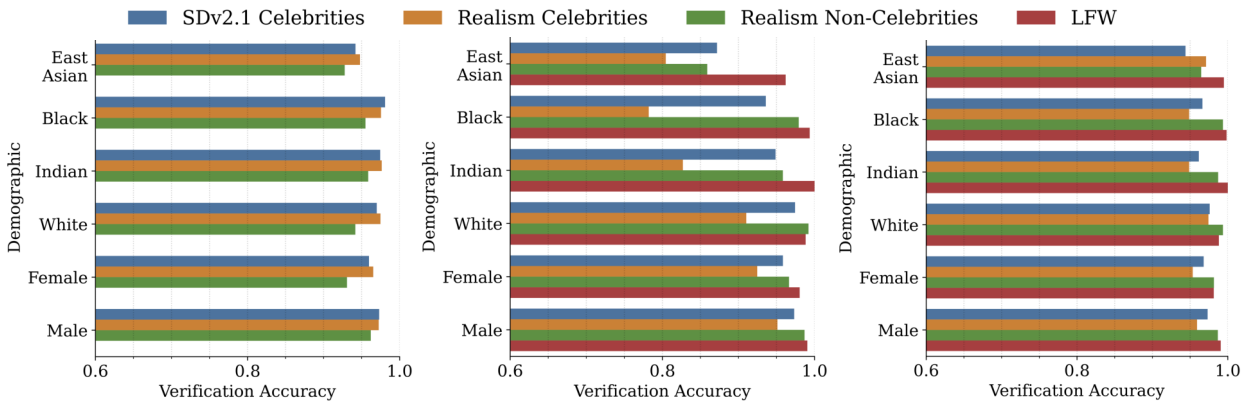
Null Hypothesis 3. The per-demographic distributions of quality difference between source and transformed images in ⟨Dataset⟩ are identical.

On the Realism Non-Celebrities and Realism Celebrities datasets, one-way ANOVA rejects null hypotheses 1 to 3 with p-values less than 0.05; corresponding p-values appear in the long-form report. This test tells us that for these two datasets, source image quality, transformed image quality, and the difference between source and transformed image quality have a dependence on demographics. The only dataset for which image quality does not conclusively depend on demographics is SDv2.1 Celebrities.

The same observation of demographic dependence applies to the transformation correctness measures (M3, M4). It is interesting to note that demographic groups that have higher source image quality are not consistent with groups of higher transformation correctness. This suggests that SEGA introduces its own biases in the generative pipeline.

Quantitative Metrics vs. User Study

User studies are the most direct way to measure human perception of generated faces. Unfortunately, they are prohibitively expensive when implemented at scale. If we have a metric serving as a proxy for human sentiment toward generated face quality, costs associated with generating realistic face data could be drastically reduced. We analyze the correlation between the different metrics and the questions posed



(a) Survey Responses (b) Same Demographic Pairs (c) Any Demographic Pairs

Figure 3: Verification Accuracy is plotted across four datasets. Each row is a demographic, and each dataset is depicted with a different hue. Note that each plot is x-axis limited between 0.6 and 1.

Dataset	Null hypothesis 4		Null hypothesis 5
	CLIP-I	DINO-I	CLIP Directional
SDv2.1 Celebrities	0.147	0.107	0.128
Realism Celebrities	0.197	0.122	0.348
Realism Non-Celebrities	0.245	0.142	0.0908

Table 2: Spearman correlation coefficients for null hypotheses 4 and 5. Each correlation coefficient is statistically significant. Corresponding p-values appear in the long-form report.

in the user study regarding the quality of the source and transformed images, the presence of semantic change, and identity retention after applying the semantic change. We calculate the Spearman correlation coefficients between the metrics and the scores to the user-study questions and once again make use of one-way ANOVA tests to reject null hypotheses:

Null Hypothesis 4. On ⟨Dataset⟩, there is no monotonic relationship between image-image ⟨similarity metric⟩ and maintenance of identity post application of semantic change.

Null Hypothesis 5. On ⟨Dataset⟩, there is no monotonic relationship between CLIP-directional and appearance of the semantic change.

On all three datasets, one-way ANOVA tests enable us to reject null hypothesis 4 on image similarity metrics CLIP-I and DINO-I. We also similarly reject null hypothesis 5 on the CLIP Directional metric. Despite rejecting null hypotheses, each Spearman coefficient is low, as evident from table 2. Hence, in the context of face recognition, image quality metrics are not a suitable proxy for humans in performing both identity verification and transformation verification tasks. This result is partially surprising: the DINO-I metric is designed to recognize differences between images of similar description. However, CLIP-I metric exhibits difficulty in distinguishing images with similar text descriptions (Ruiz et al. 2023). Our findings also indicate a low correlation between this metric and human assessment.

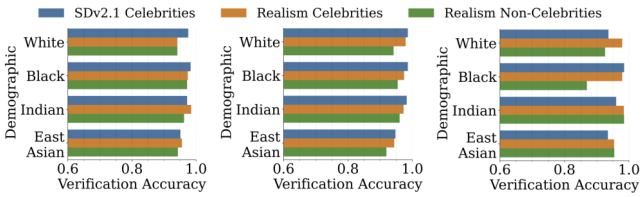
Analytical Model

We observed that verification accuracy degrades on synthetic images. We attribute this to machine learning models being trained on finite data. Typically these datasets are drawn from the internet. Generative models, such as diffusion models, thusly learn to generate images patterned on their internet-based dataset. Because the internet is well-understood to be a biased sample of the universe, a diffusion network trained on an internet-sourced dataset is itself a biased sample generator. To understand how a biased, finite training set can yield biased sample generation, we utilize a Gaussian Mixture Model (GMM). A GMM is theoretically tractable proxy through which we gain intuition about generative models.

In that model, an image in demographic group g is drawn from $\mathcal{N}(\boldsymbol{\mu}_{\mathrm{g}}, \boldsymbol{\Sigma}_{\mathrm{g}})$. For brevity, we denote the distribution of examples in group g by $\mathcal{D}_{\mathrm{g}}$. Without loss of generality, our analysis considers two groups: a and b. Group a occurs with probability α where $\alpha \in (0, 1)$. Thus, group b occurs with probability $1-\alpha$. The universe's distribution can be written as $\mathcal{D} = \alpha\mathcal{D}_a + (1-\alpha)\mathcal{D}_b$. As previously identified, generative models are typically trained on a biased dataset. To model this bias, we assume training dataset S is drawn i.i.d. from biased data distribution $\mathcal{D}_S$. Samples in S are assumed to be d-dimensional. The biased data distribution $\mathcal{D}_S$ is a possibly re-weighted mixture of Gaussians $\mathcal{D}_a$ and $\mathcal{D}_b$. That is to say, $\mathcal{D}_S = \beta\mathcal{D}_a + (1-\beta)\mathcal{D}_b$ where $\beta \in (0, 1)$. Distributions $\mathcal{D}_S$ and $\mathcal{D}$ are only equivalent if $\alpha = \beta$. For notational brevity, S_{g} denotes examples in S drawn from $\mathcal{D}_{\mathrm{g}}$.

The estimator of $\mathcal{D}$ learned from S is denoted $\hat{\mathcal{D}}_S$. The quality of estimator $\hat{\mathcal{D}}_S$ is measured with total variation distance, a notion of distributional discrepancy. Our use of total variation distance as a notion of distribution estimator quality is motivated by its use in generative model literature (Lin et al. 2018; Sajjadi et al. 2018).

Definition 1 (Discrepancy)**.** Consider a measure space $(\Omega, \mathcal{F})$. If ν_1 and ν_2 are continuous probability distributions,



(a) Black (b) White (c) Hispanic

Figure 4: User Verification Accuracy. The y-axis captures queried image demographics. Each subfigure depicts a respondent demographic. Note that each plot is x-axis limited between 0.6 and 1.

then the discrepancy is computed as

$$\rho_{\mathrm{TV}}(\nu_1, \nu_2) = \sup_{z \in \mathcal{F}} |\nu_1(z) - \nu_2(z)| \tag{2}$$

Utilizing definition 1, we can show that two distinct factors contribute to discrepancy. The first factor is the finite size of S. The second factor is S being a non-representative sample of $\mathcal{D}$. Both factors are formalized in proposition 1, its proof is in our long-form report. We assume the process yielding $\hat{\mathcal{D}}_S$ is an empirical Bayes estimator, such as the Expectation-Maximization algorithm. This process learns five parameters about the distribution: the group proportion β, the means and covariances of groups a and b: $\boldsymbol{\mu}_{\mathsf{a}}$, $\boldsymbol{\Sigma}_{\mathsf{a}}$, $\boldsymbol{\mu}_{\mathsf{b}}$, and $\boldsymbol{\Sigma}_{\mathsf{b}}$.

Proposition 1 (Proposition)**.** *Let $\delta \in (0, 1)$. If*

$$|S| = O\left(d^2 \log(2) \log(2/\delta) \left[H^2(\mathcal{D}_a, \mathcal{D}_b)\right]^{-4}\right) \tag{3}$$

then we have

$$\mathbb{P}\left[\rho_{\mathrm{TV}}(\hat{\mathcal{D}}_S, \mathcal{D}) > \frac{|\alpha - \beta|}{2} H^2(\mathcal{D}_a, \mathcal{D}_b)\right] \geq 1 - \delta \tag{4}$$

where H^2 is the squared Hellinger distance, and

$$\begin{aligned} H^2(\mathcal{D}_a, \mathcal{D}_b) &= \left(1 - \frac{|\boldsymbol{\Sigma}_a|^{1/4} |\boldsymbol{\Sigma}_b|^{1/4}}{\left|\frac{\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b}{2}\right|^{1/2}}\right) \\ &\times \exp\left\{-\frac{1}{8}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^{\mathsf{T}} \left(\frac{\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b}{2}\right)^{-1} (\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)\right\} \end{aligned} \tag{5}$$

if $\mathcal{D}_a, \mathcal{D}_b$ are each multivariate Gaussians.

This proposition suggests that even with a large number of samples in S, it can be impossible to learn $\mathcal{D}$ exactly. This is due to non-representative proportions being drawn from each demographic group, i.e. when $|\alpha - \beta| > 0$. On the other hand, when $\alpha = \beta$, and the number of samples in S is infinite, $\hat{\mathcal{D}}_S$ and $\mathcal{D}$ are equal, so $\rho_{\mathrm{TV}}(\hat{\mathcal{D}}_S, \mathcal{D})$ tends to 0. The bound also has a dependence on dimension squared: large dimension inputs require many more samples to train high-fidelity generative models. Thus, we conjecture faces generated by diffusion models trained upon larger datasets should close the observed gap in verification accuracy between LFW and datasets synthesized in this paper.

Discussion

Our user study provides direction for follow-on research relating to the Own Race Effect (ORE). ORE refers to the documented tendency of individuals to better recognize faces from within their racial group (Tanaka, Kiefer, and Bukach 2004; Meissner and Brigham 2001). We observed that our user survey seems to disagree with ORE as shown in fig. 4. Hence, a rigorous study of perceived identity of images under semantic transformations would be of research value.

As evidenced by the user study, mechanisms of human face perception present unique challenges to the application of generative models in face recognition. Moreover, automated prompt design strategies require access to a metric quantifying the quality of generated images. This does not detract from techniques evaluating generative model performance, rather, it opens a new research avenue: tuning face quality metrics to better align with human preferences.

Though our analysis techniques generalize to other domains, they assume CLIP functions as intended. Unfortunately, CLIP and similar semantic-visual embedding models are trained on internet data. Hence, their embedding space contains biases. Further, CLIP is known to have trouble constructing embeddings for uncommon or otherwise niche words and phrases. Niche words and phrases, such as "inter-eye distance" and "eyebrow slant", which can affect human perceived face identity (Tsao and Livingstone 2008), are problematic for CLIP. Further analysis of semantic-visual embeddings is necessary to gain a full picture of text-to-image generative models. Additionally, CLIP's understanding of cultural constructs is not entirely understood. For example, it is unclear what an "intelligent face" or "beautiful face" means to CLIP. Thus, the semantic transformations we study are explicit face attributes.

Finally, our study does not negate the value of generative approaches in model analysis. Generative examples can serve as a targeted curated dataset. Tailored generation has the potential to mitigate inherent biases found in existing datasets; however, the effectiveness of this approach is closely tied to the data used to train the generative model. It's important to remember that the generated examples are not i.i.d. samples from the natural distribution. Instead, they represent i.i.d. samples from a possibly skewed estimate derived from a finite pool of realized examples within the training set.

Conclusion

Generative models have been the subject of much recent societal interest. Synthesized examples achieve near-realistic quality. Though recent advances have increased the expressive power of generative models, their performance characteristics remain opaque, especially for face image generation. We put forth a new framework to synthesize diverse face images and evaluate them from multiple perspectives. Our findings are boosted with intuition from an analytical model. Our work demonstrates the need for further research into properties of semantic-visual embeddings and human perception mechanisms upon generated faces.

Acknowledgments

This work is partially supported by the DARPA GARD program under agreement number 885000, the NSF through award CNS-1942014, and the Wisconsin Alumni Research Foundation. The authors would also like to thank Yue Gao[4] for his thoughtful insights that led to the research in this work.

References

Brack, M.; Friedrich, F.; Hintersdorf, D.; Struppek, L.; Schramowski, P.; and Kersting, K. 2023. Sega: Instructing diffusion using semantic dimensions. *arXiv preprint arXiv:2301.12247*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Dixit, M.; Kwitt, R.; Niethammer, M.; and Vasconcelos, N. 2017. Aga: Attribute-guided augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7455–7463.

Friedrich, F.; Schramowski, P.; Brack, M.; Struppek, L.; Hintersdorf, D.; Luccioni, S.; and Kersting, K. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022a. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022b. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*.

Lin, Z.; Khetan, A.; Fanti, G.; and Oh, S. 2018. Pacgan: The power of two samples in generative adversarial networks. *Advances in neural information processing systems*, 31.

Luccioni, A. S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*.

Maluleke, V. H.; Thakkar, N.; Brooks, T.; Weber, E.; Darrell, T.; Efros, A. A.; Kanazawa, A.; and Guillory, D. 2022. Studying bias in gans through the lens of race. In *European Conference on Computer Vision*, 344–360. Springer.

[4]gy@cs.wisc.edu

Meissner, C. A.; and Brigham, J. C. 2001. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1): 3.

Muñoz, C.; Zannone, S.; Mohammed, U.; and Koshiyama, A. 2023. Uncovering Bias in Face Generation Models. *arXiv preprint arXiv:2302.11562*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.

Perera, M. V.; and Patel, V. M. 2023. Analyzing bias in diffusion-based face generation models. *arXiv preprint arXiv:2305.06402*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Sajjadi, M. S.; Bachem, O.; Lucic, M.; Bousquet, O.; and Gelly, S. 2018. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Seshadri, P.; Singh, S.; and Elazar, Y. 2023. The Bias Amplification Paradox in Text-to-Image Generation. *arXiv preprint arXiv:2308.00755*.

Smith, B.; Farinha, M.; Hall, S. M.; Kirk, H. R.; Shtedritski, A.; and Bain, M. 2023. Balancing the Picture: Debiasing Vision-Language Datasets with Synthetic Contrast Sets. *arXiv preprint arXiv:2305.15407*.

Struppek, L.; Hintersdorf, D.; and Kersting, K. 2022. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*.

Tanaka, J. W.; Kiefer, M.; and Bukach, C. M. 2004. A holistic account of the own-race effect in face recognition: evidence from a cross-cultural study. *Cognition*, 93(1): B1–B9.

Trabucco, B.; Doherty, K.; Gurinas, M.; and Salakhutdinov, R. 2023. Effective Data Augmentation With Diffusion Models. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Tsaban, L.; and Passos, A. 2023. LEDITS: Real Image Editing with DDPM Inversion and Semantic Guidance. arXiv:2307.00522.

Tsao, D. Y.; and Livingstone, M. S. 2008. Mechanisms of face perception. *Annu. Rev. Neurosci.*, 31: 411–437.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. arXiv:2306.10012.