Exploring Social Bias in Downstream Applications of Text-to-Image Foundation Models

Adhithya Saravanan[1,2], **Rafal Kocielnik**[2], **Roy Jiang**[2], **Pengrui Han**[3], **Anima Anandkumar**[2,4]
[1]University of Cambridge, [2]California Institute of Technology, [3]Carleton College, [4]Nvidia
{aps85@cam.ac.uk, rafalko@caltech.edu}

Abstract

Text-to-image diffusion models have been adopted into key commercial workflows, such as art generation and image editing. Characterising the implicit social biases they exhibit, such as gender and racial stereotypes, is a necessary first step in avoiding discriminatory outcomes. While existing studies on social bias focus on image generation, the biases exhibited in alternate applications of diffusion-based foundation models remain under-explored. We propose methods that use synthetic images to probe two applications of diffusion models, image editing and classification, for social bias. Using our methodology, we uncover meaningful and significant inter-sectional social biases in *Stable Diffusion*, a state-of-the-art open-source text-to-image model. Our findings caution against the uninformed adoption of text-to-image foundation models for downstream tasks and services.

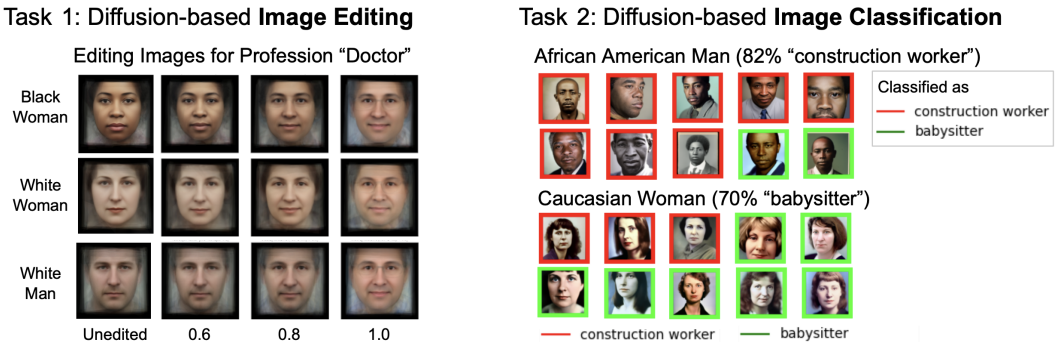

Figure 1: Impact of social bias in diffusion-based foundation models on downstream tasks uncovered using synthetic test images. Task 1: Diffusion-based editing of images for different intersectional groups results in stereotyped gender flips and skin tone changes (we depict average faces using Facer [1], example individual images can be found in §7.1.2). Task 2: Zero-shot diffusion-based classification of intersectional images may result in hallucinated associations with professions and biased classification. Here we depict a small representative subset in this classification task, the full set of images can be found in §7.2.4. Aggregate results are shown in Figure 3 and details in Table 1 and 5.

1 Introduction

Recent advances in generative text-to-image models have been fueled by the application of denoising diffusion probabilistic models [2]. Notably, DALL-E [3, 4], Imagen [5], and Stable Diffusion [6] have emerged as prominent examples, showcasing their strong visio-linguistic understanding through the production of high-resolution images across diverse contexts.

Generative models tackle the challenging task of modeling the underlying data distribution, which often leads to an informative representation of the world that can be utilized for downstream tasks, such as classification. In natural language processing, many successful pre-trained models are generative (i.e., language models). Generative pre-training is also being increasingly adopted for downstream vision tasks [7, 8], with recent works achieving competitive results against CLIP on

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

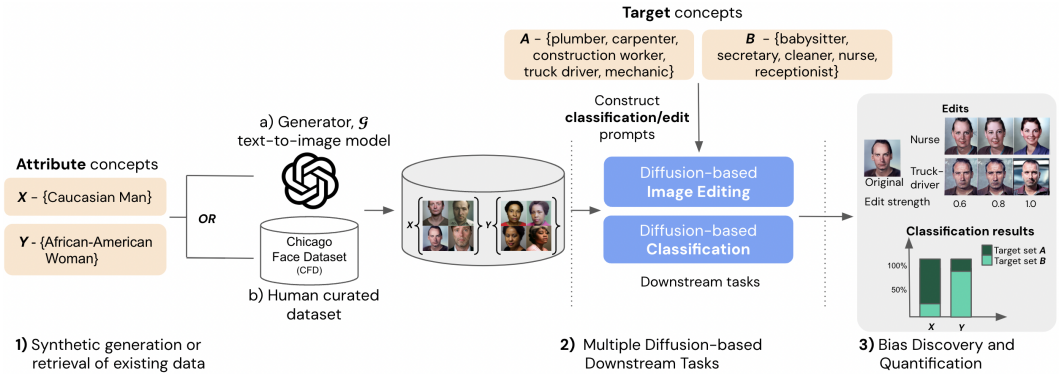

Figure 2: Overview of our approach: Our method involves defining two sets of attribute concepts, X and Y, and using either a) synthetically generated images or b) images from curated datasets to represent these concepts. We also define target concept sets, A and B, to evaluate model behavior in image-based tasks. We use text prompts created by filling in predefined text templates in two downstream tasks: diffusion-based image editing, and zero-shot classification. Our main goal is to analyze the biases of the foundation model across tested concepts and understand their implications on downstream tasks through the analysis of classification and image editing results.

zero-shot image classification, using text-to-image foundation models with no additional training. Other downstream tasks include segmentation [9], dense correspondence [10], image retrieval [11], as well as generative tasks, such as text-guided image editing [12, 13] and in-painting.

Simultaneously, a growing concern has been raised by works such as [14] and [15], which underscore the presence of various social biases—ranging from social and religious to sexual orientation—embedded within these models. These biases can be attributed to the contrastive pre-training of CLIP (encoders of most text-to-image models) and generative training of the text-to-image models. This is as the internet-scale datasets used in both these stages reflect and compound the biases in society [16], though the tendency of models to amplify imbalances in training data has also been audited [17]. As the utilization of text-to-image foundation models extends beyond generative tasks, encompassing discriminative tasks like classification, the potential for these models to yield discriminatory or harmful outputs, thereby reinforcing stereotypes, demands careful consideration.

Our Approach: In this work, we probe social bias in two applications of text-to-image foundation models, image editing [18, 19] and zero-shot classification [7, 8], using bias testing methods designed to resemble downstream workflows. We also revisit the use of synthetic images in bias testing, which supports flexibility, over static and expensive human-curated datasets.

Prior work: Recent works predominantly assess bias in text-to-image models using two methods: 1) comparisons in CLIP embedding space [20, 15], and 2) attribute (e.g. race, gender) classifiers [14] . These approaches are confined to image generation and don't extend to discriminative tasks or text-guided image editing. Krojer et al. [11] present biases in image retrieval but rely on human-curated datasets. Perera et al. [21] investigate the impact of training data on social bias in diffusion-based face generation models. The utilization of synthetic image data as supplementary training data to address fairness discrepancies across social groups in recognition tasks has been explored in previous studies [22, 23, 24]. There has also been efforts to benchmark recognition models using synthetic data by perturbing attributes, using GANs, to assess accuracy [25]. Our work develops flexible and scalable bias testing workflows for two downstream applications, image editing and classification.

Findings: In our experiments, we use a neutral[1] photo representing a social identity and prompt the model to edit it into a specific profession, mimicking real-world applications such as professional head-shot generation [26]. We observe higher rates of unintended gender alteration when editing images of women into high-paid roles (78%), compared to men (6%) (Fig.3-Left). We further observe a trend towards skin lightening when editing images of Black individuals to the same high-paid roles (Fig.3-Middle), and to a lesser extent when editing to low-paid roles.

We also analyzed the use of *Stable Diffusion* as a classifier, following [8]. Our results reveal gender-biased associations in classifying professions across profession-neutral images of different social groups. For instance, in binary classification, between a male- and female-dominated profession, the male-dominated profession was selected for synthetic images of Males 64% of the time compared to 28% for images of Females (Fig. 3-Right). This indicates a strong learned relationship between

[1]A photo without any aspects revealing the tested attributes (e.g., clothing indicative of a particular profession)

Figure 3: Left: Percentage of flips in gender (CLIP) from editing Male and Female images to high-paid roles in diffusion-based image editing. Middle: Skin Color Changes (↑ change towards lighter skin color using an established methodology described in §3) from editing images of White and Black individuals using high-paid prompts in diffusion-based image editing. Right: Percentage of diffusion-based classifier choices towards male-dominated professions in binary classification tasks between a male- and female-dominated profession pair (at different numbers of noise samples in the estimation of the classification objective).

visual cues concerning attributes, such as gender, and target concepts, such as professions. The bias towards stereotyped professions also amplifies when the number of noise samples used to calculate the classification objective is increased — a hyper-parameter linked to higher classification accuracy [8, 7] (Fig. 3-Right). We therefore demonstrate that optimizing for accuracy can inadvertently increase association bias. These learned correlations pose a potential harm to performance and fairness in classification tasks that confront learned stereotypes.

Contributions: In this work we offer the following contributions:

- To our best knowledge, we are the first to define bias testing methods for two downstream applications of text-to-image foundation models: image-editing and zero-shot classification. We leverage synthetic images to support flexibility and scalability.
- We run experiments on *Stable Diffusion* with these downstream tasks and show the presence of severe social biases across professions for various intersectional groups.
- We show that increasing hyper-parameters that improve performance in downstream tasks, including the number of noise samples (classification), also inadvertently amplifies social bias.

2 Preliminaries

Social Bias in ML: Intersectional social bias refers to the overlapping and inter-dependent forms of discrimination that individuals face due to any combination of their race, gender, class, sexuality or any other identity factors. Several works have studied how intersectionality affects the manifestation of bias in ML, including in word embeddings ([27, 28], language ([29, 30, 31]) and image-generation ([15, 14] models. Another consideration is the distinction between extrinsic and intrinsic bias, described in [32] as the biases that originate from pre-training and fine-tuning, respectively. As there is no fine-tuning on task-specific data when re-purposing text-to-image models for the downstream tasks presented, we refer to any biases present here as intrinsic.

Diffusion models: Details regarding Denoising Diffusion Models are found in [33, 2, 34, 35].

Diffusion-based Image Editing: In CLIP latent space models (e.g., [4, 6]), image generation initializes diffusion from a random latent vector, whereas image editing initializes from an embedding of the image to be edited [12, 13]. Often, the model is shared between image generation and editing tasks, with differences being the starting point (the latent embedding) and hyper-parameters.

A crucial hyper-parameter, "strength", defaulting to 0.8 (max: 1.0), controls noise addition to the reference image. Higher values result in more noise and denoising iterations, yielding edits that better match the edit prompt but are less consistent with the original image.

Diffusion-based Classification: The key idea in diffusion-based classification is that a diffusion model's ability to denoise a noised image, given a text description of a label, is a proxy for that label's likelihood [7, 8]. The classification objective, the evidence lower bound (ELBO), is defined as a Monte-Carlo estimate of the expected noise reconstruction losses (ϵ-prediction loss), obtained by repeatedly adding Gaussian random-noise, ϵ, to the image.

Among various hyper-parameters in the design of the classifier, the most critical one is the number of noise samples employed to compute the classification objective. For an input image $\boldsymbol{x}$ and class set $\boldsymbol{C}$, the text prompt $\boldsymbol{c_i}$ that minimizes noise reconstruction loss is chosen:

$$\arg\min_{c_i \in C} E_{t,\epsilon} \left\| \epsilon - \epsilon_\theta(x_t, c_i) \right\|_2^2 \tag{1}$$

3 Methodology

Synthetic data generation: We use *stable-diffusion-2-1* (SD v2.1) for synthetic dataset creation. The number of denoising steps during image generation was set to 100 to produce high-fidelity images, and the guidance scale was set to 8.5. We fixed the dataset generation prompt template as: *"A color photograph of a _____, headshot, high-quality."* based on [36].

Downstream tasks setup: We demonstrate our bias testing methods on downstream applications of Stable Diffusion model *stable-diffusion-2-1*. For classification, we use the default set-up [8]. We vary the number of noise samples in the calculation of the classification objective, ELBO (1, 10 and 100 samples). Zero-shot classification prompts follow the template: *"A portrait of a _____."*

We use *StableDiffusionImg2ImgPipeline* from Hugging Face, which uses the diffusion-denoising mechanism proposed in [37], for text-guided image editing. We vary edit strength and report results for 0.6, 0.8 (default) and 1. We use default values for number of inference steps (50) and guidance scale (7.5). We construct edit prompts using the template: *"A color photograph of a _____, headshot, high-quality."*, in line with synthetic dataset generation prompts.

Bias in Diffusion-based Image Editing: Bias testing in image generation focuses on determining attributes of the images generated for a target concept prompt, while bias testing in editing must examine changes in pre-existing visual attributes. We quantify changes during editing through zero-shot gender classification using CLIP, between *'man'* and *'woman'*, as in [38]. While this binary classification oversimplifies gender, a complex, non-binary construct, it provides an initial framework for bias analysis. We employ Facer [1], an open-source Python package, to compute the average face of sets of original and edited images. Predicting race based on visual cues is challenging, especially through CLIP [39]. Instead, we focus on skin color as a quantifiable metric, employing the Individual Typology Angle (ITA) [40] as a proxy. We use the YCbCr algorithm [41] to determine skin pixels from the average faces, and calculate the ITA, a statistical dermatology value, from their RGB values, through the implementation used in [42, 43]. ITA is versatile as it is also commonly mapped to discrete skin-tone classes, such as the Fitzpatrick Scale [44].

Bias in Diffusion-based Classification: We introduce attribute sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ (e.g., terms for male and female) and target sets $\boldsymbol{A}$ and $\boldsymbol{B}$ (e.g., professions dominated by each gender). We consider image datasets $\boldsymbol{\mathcal{D}_X}$ and $\boldsymbol{\mathcal{D}_Y}$, which can be synthetic, generated by a generator G, or human-curated, and assume neutrality concerning the concepts in $\boldsymbol{A}$ and $\boldsymbol{B}$. By classifying images into profession pairs from $\boldsymbol{A}$ and $\boldsymbol{B}$ and averaging the results, we gauge the attribute-to-target concept association. We introduce an association measure, and a differential variant, to quantify the differences in the associations of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Note that c is the decision of the classifier.

$$S(\boldsymbol{\mathcal{D}}, \boldsymbol{A}, \boldsymbol{B}) = \underset{x \in \boldsymbol{\mathcal{D}}}{avg} \underset{(a,b) \in \boldsymbol{A} \times \boldsymbol{B}}{avg} p(c = a | \{a, b\}, x) \tag{2}$$

$$S(\boldsymbol{\mathcal{D}_X}, \boldsymbol{\mathcal{D}_Y}, \boldsymbol{A}, \boldsymbol{B}) = S(\boldsymbol{\mathcal{D}_X}, \boldsymbol{A}, \boldsymbol{B}) - S(\boldsymbol{\mathcal{D}_Y}, \boldsymbol{A}, \boldsymbol{B}) \in \ [-1, 1] \tag{3}$$

4 Datasets and Results

4.1 Datasets

Human-Curated Dataset: We run our analyses on the human-curated Chicago Face Dataset (CFD) [45]. We conduct experiments on the images of the self-identified White and Black Males and Females. We use the whole dataset for classification, and randomly sampled 25 neutral facial-expression images, for each social group, for image-editing.

Synthetic Data: We also generate synthetic datasets containing 256 images for a range of intersectional social identities (Caucasian and African-American men and women). We use the whole dataset for classification, and randomly sampled 25 images, for each social group, for image-editing.

Dataset	Social Identity (X)	Edit concepts	Δ Gender (CLIP)			Δ Skin-Color (ITA)		
Edit strength			0.6	0.8	1.0	0.6	0.8	1.0
CFD	White Female	High-paid professions	0.18	0.48	**0.76**	↑0.32	↓4.25	↓1.65
	White Male		0.20	0.20	0.08	↓1.16	↓6.04	↓1.40
	Black Female		0.20	0.42	**0.72**	↑1.71	↑8.60	↑**37.39**
	Black Male		0.10	0.04	0.08	↑0.56	↑6.08	↑**35.56**
SD v2.1	Caucasian-Woman		0.04	0.24	**0.84**	↑3.99	↓4.30	↓4.33
	Caucasian-Man		0	0.02	0.06	↑3.68	↑1.51	↑**22.52**
	African-Amer. Woman		0.02	0.36	**0.78**	↑6.62	↓0.19	↑**19.18**
	African-Amer. Man		0	0	0.02	↑12.31	↓1.61	↑**20.22**
CFD	White Female	Low-paid professions	0.02	0.08	0.30	↓0.42	↓4.47	↓8.24
	White Male		0.38	**0.62**	**0.56**	↑1.32	↓1.08	↓9.04
	Black Female		0.02	0.16	0.28	↑1.71	↑4.62	↑**24.06**
	Black Male		0.22	0.42	**0.58**	↑0.54	↑3.05	↑**20.36**
SD v2.1	Caucasian Woman		0.06	0.20	0.48	↑3.29	↓3.86	↓13.03
	Caucasian Man		0.02	0.20	0.36	↑8.55	↑7.33	↑**17.05**
	African-Amer. Woman		0.06	0.38	**0.50**	↑10.11	↑3.69	↑14.05
	African-Amer. Man		0.22	0.32	0.44	↑12.26	↑8.97	↑**16.35**

Table 1: For each row, we edit 25 original images into two professions. The high-paid professions are doctor and CEO. The low-paid professions are *'dishwasher'* and *'fastfood-worker'*. This results in 50 edited images, per edit strength, in each row. 'Change in gender (CLIP)' column: Percentage of edited images that are different in gender from original image. We embolden results where more than half the edits alter the gender. 'Change in skin-color (ITA)' column: Change in ITA between the average face of the edited set and the original set of images (↓in ITA - skin becomes darker, ↑in ITA - skin becomes lighter). We embolden changes over ± 15 points. Absolute ITA values are found in Appendix 7.1.4.

Biases: We focus on professions, common for testing social biases in generative models [46]. For image editing, we focus on the two highest paid professions: *'doctors'* and *'CEOs'*, and the two lowest paid professions, *'dishwashers'* (*'dishwasher-worker'* used to avoid generations of the appliance) and *'fast-food workers'*, as per US Labour Statistics [47]. For classification, we pick the five top male and female-dominated professions, according to US Labor Statistics [46]. Male-dominated roles include *'carpenters'*, *'plumbers'*, *'truck drivers'*, *'mechanics'*, and *'construction workers'* and female-dominated include *'babysitters'*, *'secretaries'*, *'housekeepers'*, *'nurses'*, and *'receptionists'*.

4.2 Results

4.2.1 Social Bias in Text-guided Image Editing

We varied social groups, target concepts, and edit strengths (refer to Table 1).

Change in Gender: Editing images of women towards high-paying careers results in a higher rate of gender alteration than in images of men, at all strengths and in both datasets. This is prominent at the maximum edit strength, 1.0 (81% vs. 4% for synthetic and 74% vs. 8% for CFD). Editing towards low-paying careers induces a lower rate of gender alteration in images of women and a higher rate in images of men, compared to their respective rates for high-paying careers, at all strengths and in both datasets. This difference is also most prominent at the maximum edit strength (49% vs. 40% for synthetic and 57% vs 29% for CFD).

Change in Skin-Color: Positive changes in ITA value between the average face of the original and edited set of images indicate a shift towards lighter tones, while negative changes the opposite. We found an average trend towards skin lightening (M=6.85 for synthetic and M=4.51 for CFD), which is particularly prominent for non-white individuals (M=10.16 for synthetic and M=12.02 for CFD). The shift for non-white individuals is more pronounced at higher edit strengths (0.6: M=10.33, 1.0: M=17.45 for synthetic and 0.6: M=1.13, 1.0: M=29.34 for CFD). High-paid edits have a notably greater increase than low-paid edits at the max edit strength in both datasets (high-paid: M=14.40 , low-paid: M=8.61 for synthetic and high-paid: M=17.48, low-paid: M=6.04 for CFD). These trends are qualitatively validated in the average faces in Appendix 7.1.3.

4.2.2 Social Bias in Diffusion-based Classification

In Table 5, we analyze gender bias in a *stable-diffusion-2-1*-based classifier, by measuring association of profession-neutral intersectional female and male image datasets, towards male- and female-dominated profession sets. Association values of 0.0, 0.50, and 1.0 indicate female-only, unbiased,

and male-only classifications, respectively. The classifier shows lower-than-neutral association (average across ELBO steps) towards male professions for images of women (0.36 for synthetic and 0.30 for CFD) and higher for images of men (0.65 for synthetic and 0.56 for CFD). The differential association of the male and female datasets intensifies with increased ELBO samples. In the CFD dataset, it changes from 0.16 to 0.36, and in the synthetic dataset, from 0.22 to 0.37, as the ELBO samples is increased from 1 to 100. CFD images of Black females exhibit a less pronounced bias (0.27) at 100 ELBO samples compared to images of White females (0.10), but this interestingly coincides with a much lower gender identification accuracy, 52% vs. 96% (see Appendix 7.2.1).

5 Discussion

Social Bias in Diffusion-based Image Editing: Frequent unintended alterations to gender and skin color highlights the strong associations between social identities and professions, as reported by prior works in the image generation context [15, 14]. When the editing prompt challenges prevailing stereotypes associated with the image's identity, "protected attributes"—characteristics like gender or ethnicity legally safeguarded from discrimination—are frequently modified. The prevalent trend towards lighter skin tones, pronounced in non-white individuals and when editing to high-paying professions, align with prior works that suggest a *"white default"* in image generation models [48]. The presented biases are acute, as even with visual guidance on protected attributes in the embedding of the original image, edits produce biased and stereotypical results concerning the target prompt.

Social Bias in Diffusion-based Classification: We observe strong differences in the association of neutral images of different genders with particular professions. Increasing the number of ELBO samples improves classification accuracy in [8, 7] and 7.2.1, but also escalates social bias. Enhanced proficiency in recognizing protected attributes like gender (7.2.1) inadvertently intensifies biased associations. This consistent but misleading correlation, in the absence of concrete profession identifiers, raises questions regarding the robustness of the *stable-diffusion-2-1*-based classifiers.

Broader Impact and Deployment: The framework for evaluating the social impact of generative AI systems presented in [49] suggests two modes of evaluation— 1) evaluating the technical 'base' system and 2) the impact of context-specific deployment on people and society. Our work tackles the former, albeit it through hypothetical in-context applications, and raises concerns across several criteria used to evaluate the technical system including 'Bias, Stereotypes, and Representational Harms' and 'Disparate Performance'. Further, the differential ease of applying certain edits to different groups, whilst preserving identity (gender, race but also facial features, for example) makes these systems susceptible to misuse, including for the perpetuation of negative stereotypes. We defer to [49] for further discussion on evaluation areas for potential impacts on people and society.

Limitations: One limitation is the potential for the generator text-to-image model to inject it's own social biases into the test images. We assume generated images are profession neutral and diverse. For attribute concepts where this is not the case, prompt engineering should be explored [50]. We inherit CLIP's fairness and accuracy limitations, by using it for gender classification. However, CLIP achieved a 100% accuracy on the unedited images, and similar performance is expected for the edited images. Further, image based skin-tone calculation is susceptible to artefacts and low dynamic ranges (note the low initial ITA value for synthetic Caucasian-man images from the occasionally murky generated images). Future work could verify the data quality, by exerting tight control over synthetic image generation (e.g. make attribute level changes to human-curated images) and post-generation normalization (e.g. filter for well-lit images, center faces and normalize viewpoints).

6 Conclusion

Methods derived from Stable Diffusion showcased pronounced intersectional biases across gender and skin-color indicating a pressing need for bias testing methods that are aligned with downstream tasks, in order to facilitate ethical deployment. Our work serves as an initial foray into methodology that supports flexible bias testing at scale in two such downstream tasks. In future work, the effectiveness of refined editing or classification prompts [50] and the fairness of diffusion-based classifiers, especially concerning a larger range of intersectional social groups, should be explored to further understand and improve robustness and reliability.

References

[1] johnwmillr/facer: Simple face averaging in python. `https://github.com/johnwmillr/Facer`. (Accessed on 09/27/2023).

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[5] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[7] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers, 2023.

[8] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier, 2023.

[9] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation, 2023.

[10] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence, 2023.

[11] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners?, 2023.

[12] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023.

[13] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions, 2023.

[14] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens, 2023.

[15] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models, 2023.

[16] Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps, 2023.

[17] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation, 2023.

[18] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *CoRR*, abs/2111.05826, 2021.

[19] Robert Wolfe and Aylin Caliskan. American == white in multimodal language-and-image ai, 2022.

[20] Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation, 2023.

[21] Malsha V Perera and Vishal M Patel. Analyzing bias in diffusion-based face generation models. *arXiv preprint arXiv:2305.06402*, 2023.

[22] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification, 2023.

[23] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition?, 2023.

[24] Seyma Yucer, Samet Akçay, Noura Al Moubayed, and Toby P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. *CoRR*, abs/2004.08945, 2020.

[25] Hao Liang, Pietro Perona, and Guha Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation, 2023.

[26] AppleNews. Editing images for resume / linkedin bias. `https://apple.news/ASm78Oj8oR1GWvQWBbwL6vQ`. (Accessed on 09/27/2023).

[27] Bhavya Ghai, Md Naimul Hoque, and Klaus Mueller. Wordbias: An interactive visual tool for discovering intersectional biases encoded in word embeddings, 2021.

[28] Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21. ACM, July 2021.

[29] Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23. ACM, August 2023.

[30] Rafal Kocielnik, Shrimai Prabhumoye, Vivian Zhang, Roy Jiang, R. Michael Alvarez, and Anima Anandkumar. Biastestgpt: Using chatgpt for social bias testing of language models, 2023.

[31] Liam Magee, Lida Ghahremanlou, Karen Soldatic, and Shanthi Robertson. Intersectional bias in causal language models, 2021.

[32] Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *CoRR*, abs/2112.07447, 2021.

[33] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[34] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *CoRR*, abs/2102.09672, 2021.

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[36] BloombergTechnology+Equality. Humans are biased. generative ai is even worse. `https://www.bloomberg.com/graphics/2023-generative-ai-bias/`. (Accessed on 09/27/2023).

[37] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *CoRR*, abs/2108.01073, 2021.

[38] Robert Wolfe and Aylin Caliskan. Markedness in visual semantic ai, 2022.

[39] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan. Evidence for hypodescent in visual semantic ai, 2022.

[40] Del Bino, S Sok, J Bessac, and E Bernerd. Relationship between skin response to ultraviolet exposure and skin color type. *Pigment Cell Res*, 19(6):606–614, 2006.

[41] Seema Kolkur, Dhananjay R. Kalbande, P. Shimpi, C. Bapat, and Janvi Jatakia. Human skin detection using rgb, HSV and ycbcr color models. *CoRR*, abs/1708.02694, 2017.

[42] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021.

[43] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *arXiv preprint arXiv:2207.02942*, 2022.

[44] Newton M. Kinyanjui, Timothy Odonga, Celia Cintas, Noel C. F. Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R. Varshney. Estimating skin tone and effects on classification performance in dermatology datasets. *CoRR*, abs/1910.13268, 2019.

[45] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4):1122–1135, December 2015.

[46] Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frédéric A. Dreyer, Aleksandar Shtedritski, and Yuki Markus Asano. How true is gpt-2? an empirical analysis of intersectional occupational biases. *CoRR*, abs/2102.04130, 2021.

[47] Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity : U.s. bureau of labor statistics. `https://www.bls.gov/cps/cpsaat11.htm`. (Accessed on 09/27/2023).

[48] Robert Wolfe and Aylin Caliskan. American == white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 800–812, New York, NY, USA, 2022. Association for Computing Machinery.

[49] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III au2, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. Evaluating the social impact of generative ai systems in systems and society, 2023.

[50] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions?, 2022.

Acknowledgments and Disclosure of Funding

We would like to thank the Caltech SURF program for contributing to the funding of this project. Anima Anandkumar is Bren Professor at Caltech and Senior Director of AI Research at NVIDIA. This material is based upon work supported by the National Science Foundation under Grant # 2030859 to the Computing Research Association for the CIFellows Project.

7 Appendix

7.1 Image Editing

We present the edits for a subset of the edit prompts tested, namely 'doctor' (high-paid) and 'fast-food worker' (low-paid), across both the human-curated (CFD) and synthetic (SD v2.1) sets, for White/Caucasian Men and Black/African-American Women. This provides a qualitative sense of the differential shift of protected attributes, including gender and skin color, when the edit prompt concerns a stereotype or an anti-stereotype concept.

7.1.1 Human-curated (CFD) Image Edits

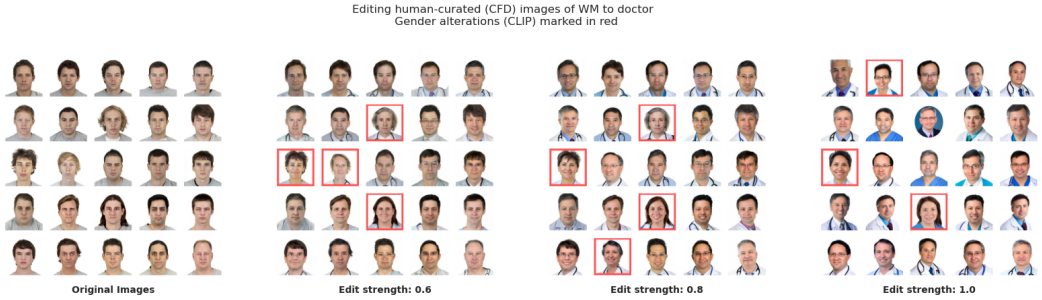

Figure 4: Edits of human-curated (CFD) images of 'White Male' to doctor

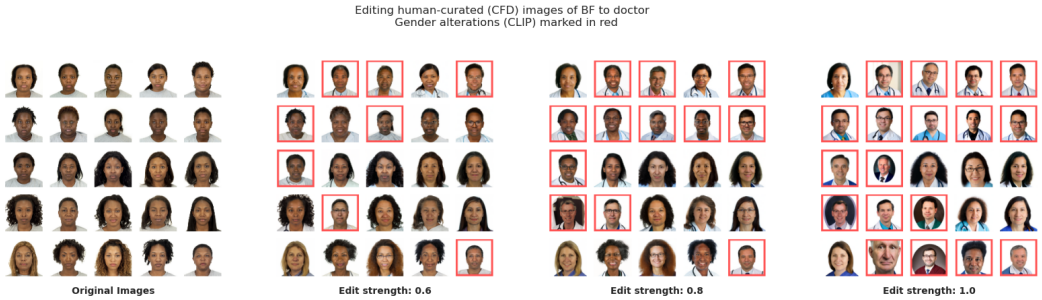

Figure 5: Edits of human-curated (CFD) images of 'Black Female' to doctor

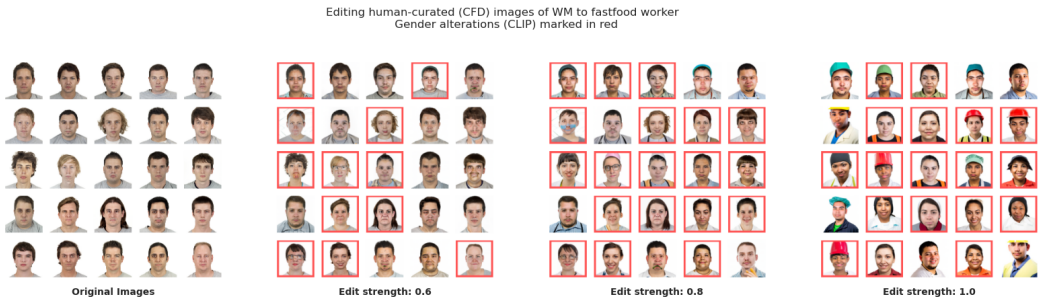

Figure 6: Edits of human-curated (CFD) images of 'White Male' to fastfood worker

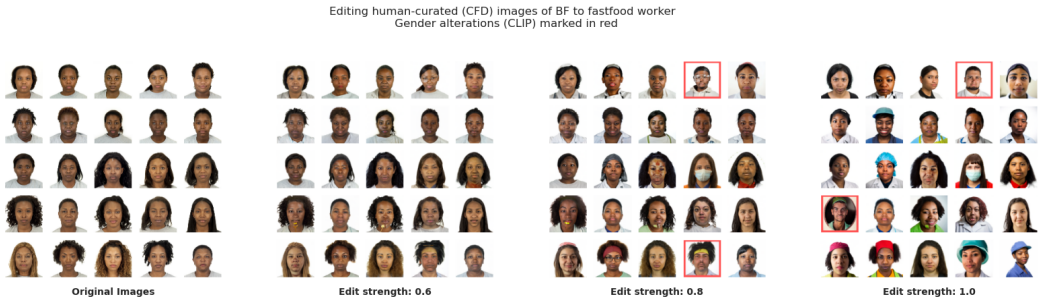

Figure 7: Edits of human-curated (CFD) images of 'Black Female' to fastfood worker

7.1.2 Synthetic (SD v2.1) Image Edits

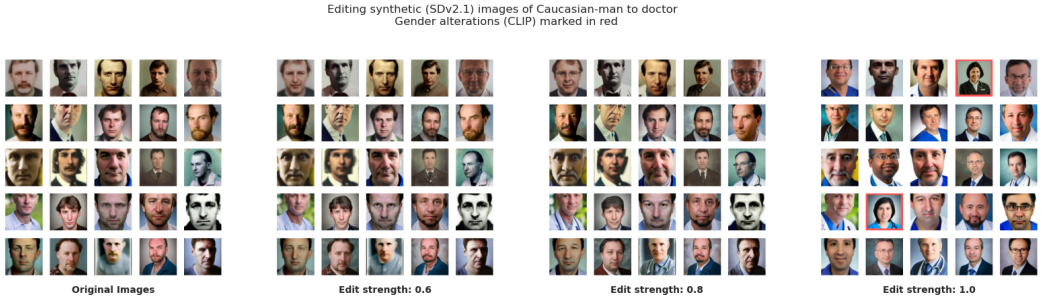

Figure 8: Edits of synthetic (SD v2.1) images of 'Caucasian Man' to doctor

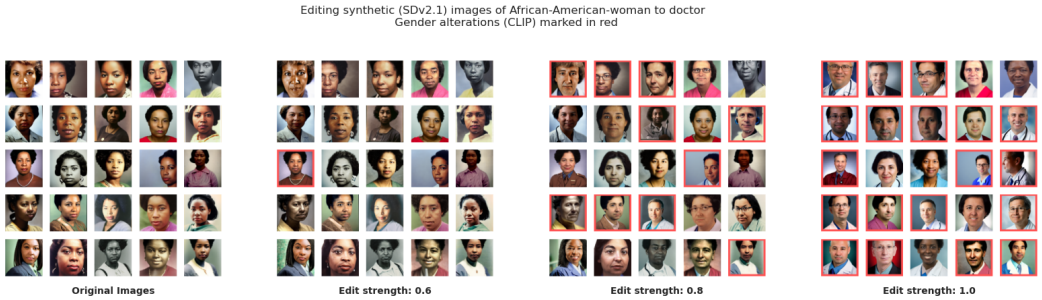

Figure 9: Edits of synthetic (SD v2.1) images of 'African-American Woman' to doctor

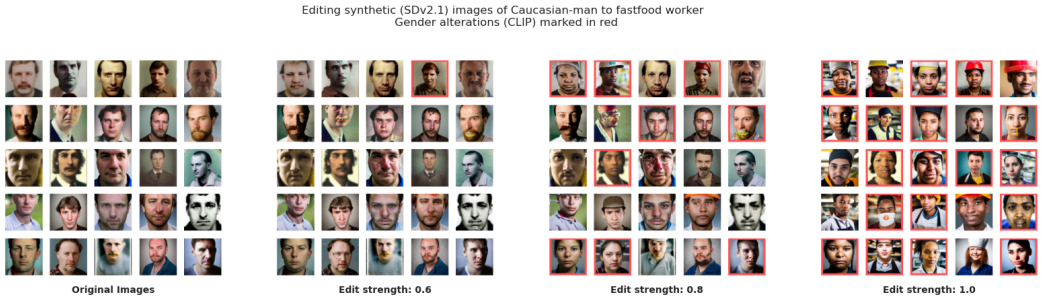

Figure 10: Edits of synthetic (SD v2.1) images of 'Caucasian Man' to fastfood worker

Figure 11: Edits of synthetic (SD v2.1) images of 'African-American Woman' to fastfood worker

7.1.3 Average Faces

We present the average faces of the original images, of those edited, and the edits, of different edit strengths and concepts, for all social groups, for both human-curated (CFD) and synthetic data.

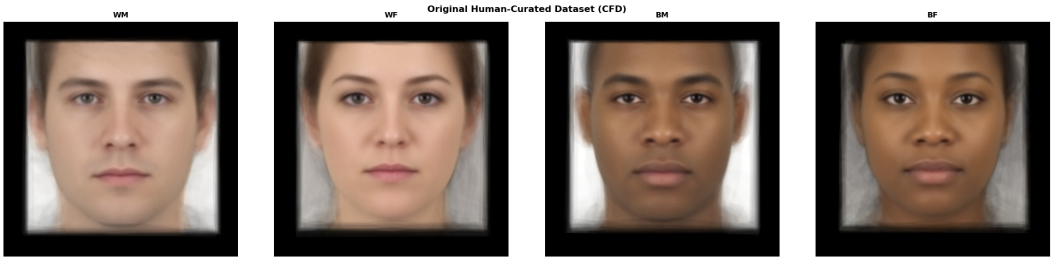

Figure 12: Average faces of the images that were edited (Original) - Human-curated (CFD)

Figure 13: Average faces of the images that were edited (Original) - Synthetic (SD v2.1)

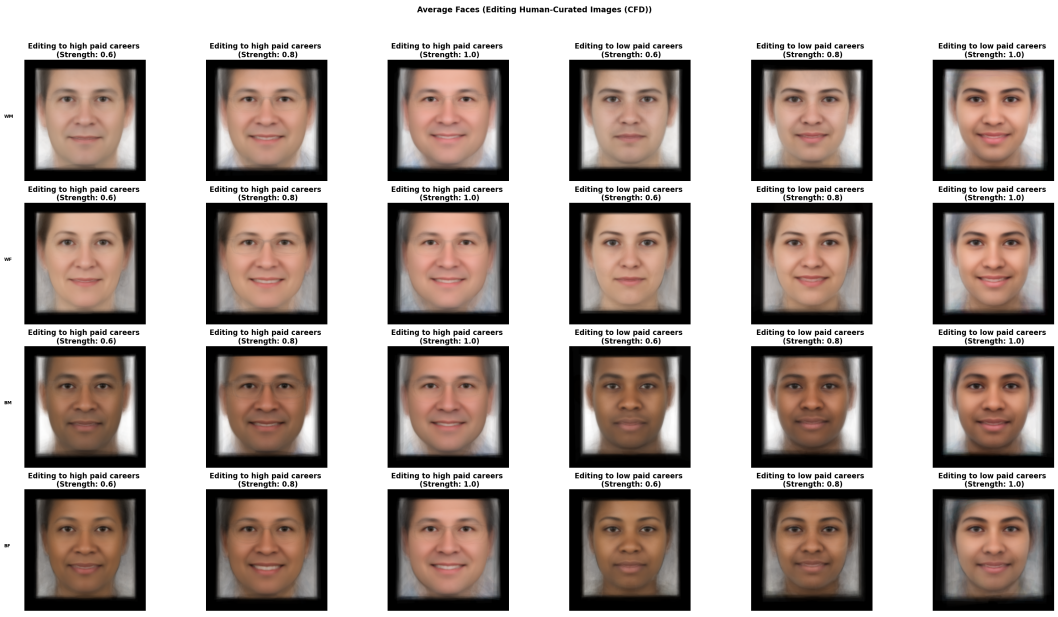

Figure 14: Average faces of edited sets of images, for human-curated (CFD) images, when edited towards high- and low-paying careers. Note that each average face is comprised of 50 images, 25 edits towards each profession (low-paying: dishwasher-worker and fastfood-worker, high-paying: doctor and CEO).

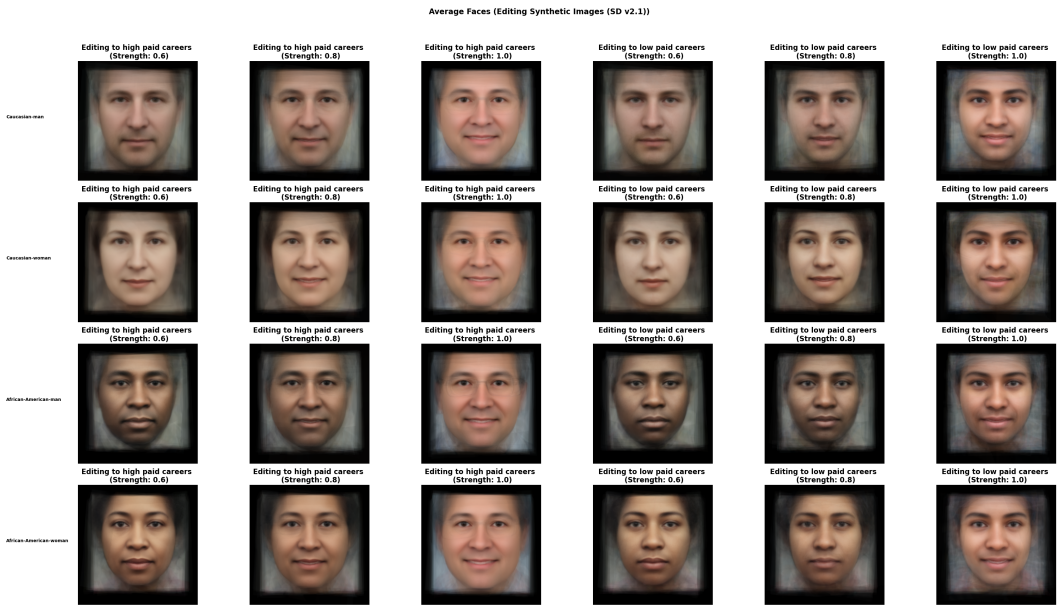

Figure 15: Average faces of edited sets of images, for synthetic (SD v2.1) images, when edited towards high- and low-paying careers. Note that each average face is comprised of 50 images, 25 edits towards each profession (low-paying: dishwasher-worker and fastfood-worker, high-paying: doctor and CEO).

7.1.4 Average Faces - Skin-color (ITA) Values

Dataset	Social group	ITA
CFD	Black Male	-17.73
	White Male	41.91
	Black Female	-15.66
	White Female	39.39"
SD v2.1	Caucasian-woman	26.63
	Caucasian-man	2.52
	African-American-man	-10.29
	African-American-woman	-3.81

Table 2: We report the absolute ITA values for the average face of the original images that were edited, for each social group, for both human-curated (CFD) and synthetic (*SD v2.1*) datasets (averages faces shown in 12 and 13).

Dataset	Social group	Edit Concept	ITA (0.6)	ITA (0.8)	ITA (1.0)
CFD	BM	low-paid	-17.18	-14.68	2.63
		high-paid	-17.16	-11.64	17.84
	BF	low-paid	-13.94	-11.03	8.4
		high-paid	-13.94	-7.06	21.74
	WM	low-paid	43.23	40.83	32.87
		high-paid	40.76	35.87	40.51
	WF	low-paid	38.97	34.91	31.15
		high-paid	39.71	35.14	37.74
SD v2.1	African-American-man	low-paid	1.97	-1.32	6.06
		high-paid	2.02	-11.9	9.93
	African-American-woman	low-paid	6.29	-0.12	10.24
		high-paid	2.81	-4.01	15.36
	Caucasian-man	low-paid	11.07	9.85	19.57
		high-paid	6.2	4.03	25.04
	Caucasian-woman	low-paid	29.93	22.78	13.61
		high-paid	30.62	22.33	22.31

Table 3: We report the absolute ITA values for the average face of the edited sets of images, for both human-curated (CFD) and synthetic (*SD v2.1*) datasets, at all edit strengths, for low and high-paid professions (average faces shown in 14 and 15).

7.2 Classification

7.2.1 Gender Classification

Dataset	**Social group**	**Accuracy**		
Number of noise samples in ELBO estimation		1	10	100
CFD	White Female	0.54	0.65	0.96
	White Male	0.75	0.93	0.95
	Black Female	0.35	0.35	0.52
	Black Male	0.93	0.98	1.00
SD v2.1	Caucasian-Woman	0.68	0.87	0.96
	Caucasian-Man	0.83	0.86	0.97
	African-American-Woman	0.71	0.87	0.97
	African-American-Man	0.69	0.80	0.97

Table 4: We report accuracy in gender classification, into 'A portrait of a man.' and 'A portrait of a woman.', as a reference of the classification fidelity of the SD v2.1-based classifier. We report results for 1, 10 and 100 noise samples in the estimation of the classification objective (ELBO). Accuracy increases across the board as the number of the noise samples is increased.

7.2.2 Profession Associations with Intersectional Social Identities

Dataset	**Social Identity** (X)	Target Set 1 (A)	Target Set 2 (B)	$S(\boldsymbol{\mathcal{D}_X}, \boldsymbol{A}, \boldsymbol{B})$		
Number of noise samples in ELBO estimation				1	10	100
CFD	White Female	Male-dominated professions	Female-dominated professions	0.42	0.18	0.10
	White Male			**0.54**	0.47	**0.51**
	Black Female			0.47	0.39	0.27
	Black Male			**0.67**	**0.62**	**0.58**
SD v2.1	Caucasian Woman			0.39	0.30	0.35
	Caucasian Man			**0.67**	**0.66**	**0.83**
	African-Amer. Woman			0.43	0.33	0.39
	African-Amer. Man			**0.59**	**0.51**	**0.65**
Mean across races	**Female**			0.43	0.30	0.28
	Male			**0.62**	**0.57**	**0.64**

Table 5: Association measure towards male- and female-dominated profession sets, for human-curated (CFD) and synthetic (SD v2.1) datasets across inter-sectional social identities. We embolden values greater than 0.50, which suggests a biased association towards the male-dominated professions set, A.

7.2.3 Profession Associations - Individual Comparisons

We present the inter target-set comparisons from the CFD and synthetic images' association tests in which one of the two profession classes is picked > 75% of the time. We report the results for 100 ELBO samples, the set-up that corresponds to the greatest classifier fidelity of those tested (1, 10, 100).

X	a	b	%a
BF	mechanic	babysitter	0.0
BF	mechanic	receptionist	0.0
BF	mechanic	housekeeper	0.09
BF	mechanic	nurse	0.16
BF	plumber	babysitter	0.0
BF	plumber	receptionist	0.01
BF	plumber	housekeeper	0.2
BF	carpenter	babysitter	0.0
BF	carpenter	secretary	0.77
BF	carpenter	receptionist	0.01
BF	carpenter	housekeeper	0.24
BF	construction worker	babysitter	0.0
BF	construction worker	secretary	0.83
BF	construction worker	receptionist	0.04
BF	truck driver	babysitter	0.0
BF	truck driver	secretary	0.77
BF	truck driver	receptionist	0.04
BM	mechanic	babysitter	0.06
BM	mechanic	secretary	1.0
BM	mechanic	receptionist	0.12
BM	mechanic	housekeeper	0.77
BM	mechanic	nurse	0.79
BM	plumber	babysitter	0.1
BM	plumber	secretary	1.0
BM	plumber	receptionist	0.16
BM	plumber	housekeeper	0.87
BM	plumber	nurse	0.81
BM	carpenter	babysitter	0.05
BM	carpenter	secretary	1.0
BM	carpenter	receptionist	0.13
BM	carpenter	housekeeper	0.83
BM	carpenter	nurse	0.75
BM	construction worker	babysitter	0.14
BM	construction worker	secretary	0.99
BM	construction worker	receptionist	0.21
BM	construction worker	housekeeper	0.85
BM	construction worker	nurse	0.78
BM	truck driver	babysitter	0.16
BM	truck driver	secretary	1.0
BM	truck driver	receptionist	0.24
BM	truck driver	housekeeper	0.87
BM	truck driver	nurse	0.84
WF	mechanic	babysitter	0.0
WF	mechanic	secretary	0.21
WF	mechanic	receptionist	0.0
WF	mechanic	housekeeper	0.03
WF	mechanic	nurse	0.0
WF	plumber	babysitter	0.0
WF	plumber	receptionist	0.0
WF	plumber	housekeeper	0.06
WF	plumber	nurse	0.21
WF	carpenter	babysitter	0.0
WF	carpenter	receptionist	0.0
WF	carpenter	housekeeper	0.01
WF	carpenter	nurse	0.03
WF	construction worker	babysitter	0.0
WF	construction worker	receptionist	0.01
WF	construction worker	housekeeper	0.04
WF	construction worker	nurse	0.14
WF	truck driver	babysitter	0.0
WF	truck driver	receptionist	0.01
WF	truck driver	housekeeper	0.05
WF	truck driver	nurse	0.08
WM	mechanic	babysitter	0.0
WM	mechanic	secretary	0.93
WM	mechanic	receptionist	0.13
WM	plumber	babysitter	0.17
WM	plumber	secretary	0.96
WM	plumber	housekeeper	0.78
WM	plumber	nurse	0.83
WM	carpenter	babysitter	0.02
WM	carpenter	secretary	0.96
WM	carpenter	receptionist	0.18
WM	construction worker	babysitter	0.04
WM	construction worker	secretary	0.97
WM	construction worker	receptionist	0.21
WM	truck driver	babysitter	0.08
WM	truck driver	secretary	0.97
WM	truck driver	nurse	0.81

Table 6: Inter target-set comparisons with >75% decisions towards one profession for human-curated (CFD) images (100 ELBO samples)

X	a	b	%a
African-American-man	mechanic	babysitter	0.87
African-American-man	mechanic	secretary	0.85
African-American-man	mechanic	housekeeper	0.8
African-American-man	mechanic	nurse	0.24
African-American-man	plumber	receptionist	0.15
African-American-man	plumber	nurse	0.12
African-American-man	carpenter	babysitter	0.94
African-American-man	carpenter	secretary	0.93
African-American-man	carpenter	housekeeper	0.85
African-American-man	construction worker	babysitter	0.82
African-American-man	construction worker	secretary	0.83
African-American-man	construction worker	housekeeper	0.82
African-American-man	truck driver	babysitter	0.94
African-American-man	truck driver	secretary	0.88
African-American-man	truck driver	receptionist	0.82
African-American-man	truck driver	housekeeper	0.88
African-American-woman	mechanic	receptionist	0.24
African-American-woman	mechanic	nurse	0.01
African-American-woman	plumber	babysitter	0.02
African-American-woman	plumber	secretary	0.15
African-American-woman	plumber	receptionist	0.01
African-American-woman	plumber	nurse	0.0
African-American-woman	carpenter	secretary	0.81
African-American-woman	carpenter	housekeeper	0.77
African-American-woman	carpenter	nurse	0.05
African-American-woman	construction worker	nurse	0.05
African-American-woman	truck driver	nurse	0.12
Caucasian-man	mechanic	babysitter	0.94
Caucasian-man	mechanic	secretary	0.95
Caucasian-man	mechanic	receptionist	0.96
Caucasian-man	mechanic	housekeeper	0.86
Caucasian-man	plumber	babysitter	0.75
Caucasian-man	plumber	secretary	0.86
Caucasian-man	plumber	receptionist	0.78
Caucasian-man	plumber	housekeeper	0.79
Caucasian-man	carpenter	babysitter	0.95
Caucasian-man	carpenter	secretary	0.99
Caucasian-man	carpenter	receptionist	0.98
Caucasian-man	carpenter	housekeeper	0.91
Caucasian-man	construction worker	babysitter	0.86
Caucasian-man	construction worker	secretary	0.92
Caucasian-man	construction worker	receptionist	0.93
Caucasian-man	construction worker	housekeeper	0.84
Caucasian-man	truck driver	babysitter	0.96
Caucasian-man	truck driver	secretary	0.96
Caucasian-man	truck driver	receptionist	0.99
Caucasian-man	truck driver	housekeeper	0.9
Caucasian-woman	mechanic	nurse	0.01
Caucasian-woman	plumber	babysitter	0.03
Caucasian-woman	plumber	secretary	0.12
Caucasian-woman	plumber	receptionist	0.06
Caucasian-woman	plumber	nurse	0.0
Caucasian-woman	carpenter	nurse	0.03
Caucasian-woman	construction worker	nurse	0.02
Caucasian-woman	truck driver	nurse	0.05

Table 7: Inter target-set comparisons with >75% decisions towards one profession for synthetic (SD v2.1) images (100 ELBO samples)

7.2.4 Visualisation

A significant inter target-set comparison/classification task, between classes a: construction-worker and b: babysitter, from the synthetic images' (SD v2.1) association test is visualised below. This provides a qualitative sense of the spread of professions assigned to images of different social identities. The figure is generated from classification results in which 100 samples were used in ELBO estimation.

Figure 16: Synthetic (SD v2.1) 'Caucasian Man' images classified into classes 'babysitter' and 'construction worker'

Figure 17: Synthetic (SD v2.1) 'African-American man' images classified into classes 'babysitter' and 'construction worker'

Figure 18: Synthetic (SD v2.1) 'Caucasian Woman' images classified into classes 'babysitter' and 'construction worker'

Figure 19: Synthetic (SD v2.1) 'African-American woman' images classified into classes 'babysitter' and 'construction worker'

7.3 Aggregated Bias Results - Human-curated (CFD) images

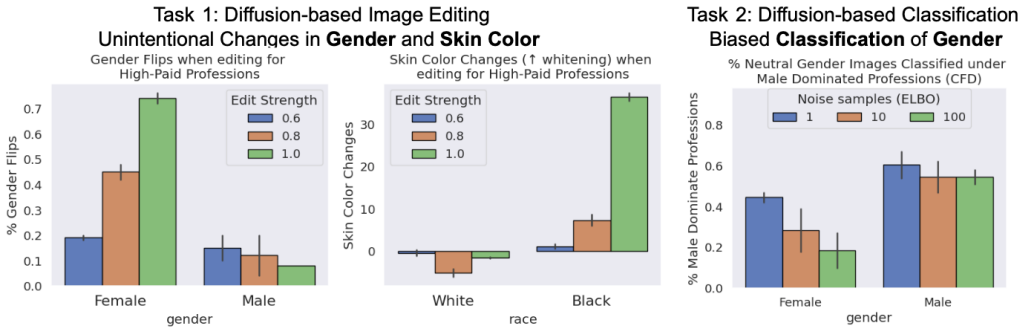

Figure 20: Left: Percentage of flips in gender (CLIP) from editing human-curated (CFD) Male and Female images using high-paid prompts in diffusion-based image editing. Middle: Skin Color Changes (↑ change towards lighter skin color using an established methodology described in §3) from editing human-curated (CFD) images of White and Black individuals using high-paid prompts in diffusion-based image editing. Right: Percentage of diffusion-based classifier choices towards male-dominated professions in binary classification tasks between a male- and female-dominated profession pair (at different numbers of noise samples in the estimation of the classification objective) for human-curated (CFD) images.

7.4 Aggregated Bias Results - Synthetic (SD) images

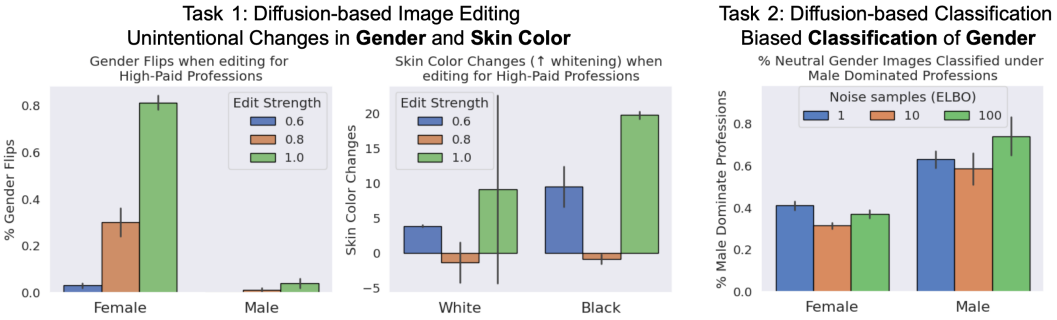

Figure 21: Left: Percentage of flips in gender (CLIP) from editing synthetic (SD) Male and Female images using high-paid prompts in diffusion-based image editing (across different levels of edit strength). Middle: Skin Color Changes (↑ change towards lighter skin color using an established methodology described in §3) from editing synthetic (SD) images of White and Black individuals using high-paid prompts in diffusion-based image editing (across different levels of edit strength). Right: Percentage of diffusion-based classifier choices towards male-dominated professions in binary classification tasks between a male- and female-dominated profession pair (at different numbers of noise samples in the estimation of the classification objective) for synthetic (SD) images.