Evaluating the Fairness of Diffusion-based Face Generation from Conversational Text - A Pilot Study

VISHNU SASHANK DORBALA and ANGELOS MAVROGIANNIS

1 INTRODUCTION

Generative AI is becoming an integral part of our daily lives, showing up in various forms such as chatbots powered by Large Language Models (LLMs) [Brown et al. 2020], or realistic high-quality images synthetically generated using diffusion models [Rombach et al. 2022]. As data-driven models that were pre-trained on large-scale internet corpora, these models are biased and often amplify dangerous stereotypes [Bianchi et al. 2023; Nadeem et al. 2020]. Besides text-to-text and text-to-image generation, generative models serve as the foundation for downstream tasks with additional data modalities, giving rise to the emerging generalized term of pre-trained Foundation Models [Zhou et al. 2023].

While massive deployment of these models has been positively contributing towards making these powerful technologies available to the people, emerging biases lurking beneath them can lead to a mass dissemination of stereotypes. An interesting multi-modal task where these biases might arise is face generation from speech in the context of voice profiling. Voice profiling aims to identify various parameters of a human (e.g. age, gender etc.) from their voice and has been explored in multiple scientific fields [Singh 2019]. Studies have shown that voice profiling is often used maliciously to promote racial discrimination [Baugh 2005]. While there have been some approaches on face generation from speech [Oh et al. 2019; Wen et al. 2019], a sample in the input data of these approaches consists of an audio recording of a single person speaking which is then embedded to a vector. Given that an increasing amount of companies have been integrating AI-powered chatbots into their services, replacing verbal communication with human agents, data in the form of audio recordings might become scarce, which raises the need for face generation methods that are based on conversational data in textual form.

In a more practical scenario, as embodied agents in the form of cognitive and personal robots become increasingly popular in household environments [Bylieva et al. 2020], we could expect stricter constraints on privacy based on user opinions [Ray et al. 2008; Sung et al. 2008]. As such, in the absence of a camera due to privacy-related issues, having an estimate on the appearance of a person could potentially help with better understanding the choices and personality of the user. Other potential use cases of a text-to-face model include identifying a person of interest in a criminal investigation based on textual evidence, generating synthetic datasets for face detection and age estimation, or generating a realistic user avatar in an online game. In the case of a criminal investigation, the model could act as an AI-based sketch artist, utilizing witness statements or transcribed conversations including potential suspects to generate an estimate of their face. Synthetic face generation could augment existing datasets by mitigating racial biases and minority underrepresentation.

To this end, inspired by work exposing emerging biases of text-to-image models [Friedrich et al. 2023], we look at measuring the fairness of diffusion-based face generation models powered by conversational input data in textual form. We design a pilot study where participants answer LLM-generated non-intrusive questions about themselves with the goal of implicitly extracting facial features to build an informed prompt for face generation.

2 BACKGROUND AND RELATED WORK

Wen et al. [2019] introduced the task of generating faces from voice in the context of voice profiling using Generative Adversarial Networks (GANs) [Goodfellow et al. 2014]. In this work, a voice embedding network generates an embedding vector from an audio recording, which is then passed to a GAN that is trained to generate a face image that matches the identity of the corresponding speaker in the training set. Speech2Face [Oh et al. 2019] is another approach generating faces from voices. The authors trained a model on short audio-visual recordings in a self-supervised manner, learning voice-face correlations towards producing images that capture various physical attributes of the speakers (e.g. age, gender, ethnicity). The model consists of two separately trained networks: a voice encoder that receives a spectrogram as input and a 4096-dimensional face descriptor from the penultimate layer of a VGG-Face network [Parkhi et al. 2015], and a face decoder [Cole et al. 2017] that receives this descriptor and constructs an image of a face in a canonical form (front-facing, neutral expression).

Similarly, early work on text-to-face models [Ayanthi and Munasinghe 2022; Khan et al. 2020; Nasir et al. 2019] is based on an encoder-decoder architecture, projecting textual description of faces to a latent space, and then using GANs to generate images of faces conditioned on the encoded text. Recent works [Borji 2022] are leveraging diffusion-based models such as OpenAI's DALL-E [Ramesh et al. 2022] and Stability AI's Stable Diffusion [Rombach et al. 2022] for face generation from text, or even diffusion models with multimodal input signals, as proposed by Huang et al. [2023]. The textual descriptions come from datasets like CelebFaces Attributes (CelebA) [Liu et al. 2015] or Labeled Faces in the Wild (LFW) [Learned-Miller 2014]. A common set of metrics used to evaluate these models for realism are is FID and LPIPS scores, which compare the actual face of the person with the generated samples.

When it comes to uncovering biases in text-to-image generative models, Friedrich et al. [2023] define fairness as a joint probability of an image along with its label given a protected attribute of interest. They look at the probabilities $P(x, y = 1|a = 1)$ and $P(x, y = 1|a = 0)$, where x is a given image, y is the corresponding label, for example "firefighter", and a is the protected attribute, in this example "gender". On the other hand, [Bianchi et al. 2023] evaluate fairness qualitatively by designing carefully hand-engineered prompts and visually evaluating the generated images resulting from a stable diffusion model [Rombach et al. 2022]. Studying emerging biases of diffusion-based models is an active area of research. Schramowski et al. [2023] focus on inappropriate generated images and propose a strategy to mitigate them. Their study is focused on the stable diffusion model, while Luccioni et al. [2023] also study DALL-E and explore social biases focusing on gender and ethnicity. The most relevant work to ours is by Perera and Patel [2023], who utilize the FairFace dataset [Karkkainen and Joo 2021] to train diffusion models and GANs and provide quantitative results focusing on gender-, racial-, and age-related biases. However, these methods depend on datasets including captions or sets of attributes that describe the images succinctly and accurately. On the other hand, we orchestrate a pilot study and capture conversational data through a Q&A of an LLM with a group of human participants to implicitly mine their facial features.

3 METHOD

Figure 1 presents an overview of our proposed pipeline. We first use an LLM (GPT-4) to both generate queries and summarize a conversation with a person. The summary is then used as a prompt to face generation models. We finally measure the fairness of our approach by comparing traits from the generated face with the actual face of the person. Notice that our scheme is driven by the Large Language Model, as the prompt given to the face generator is the LLM's summary of a conversation that the LLM itself wrote dialogues for.

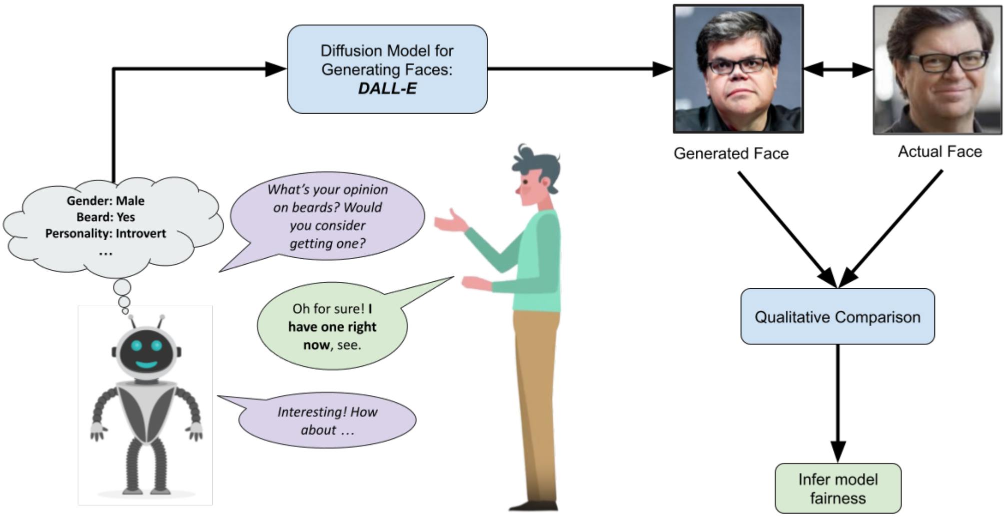

Diffusion Model for Generating Faces: DALL-E
Generated Face
Actual Face
Gender: Male
Beard: Yes
Personality: Introvert
...
What's your opinion on beards? Would you consider getting one?
Oh for sure! I have one right now, see.
Interesting! How about ...
Qualitative Comparison
Infer model fairness

Fig. 1. We investigate the fairness of Face Generation models. We employ GPT-4 to summarize a conversation, providing it context of what our task is, and asking it to generate a *personality prompt* for a face generation model. We then extract features and use it to compare the generated and actual faces, helping us infer the fairness of the model. The prompt for Yann LeCun's generated face here was 'Nerdy guy with glasses'.

Our experimental setup for evaluating fairness has 3 components :-

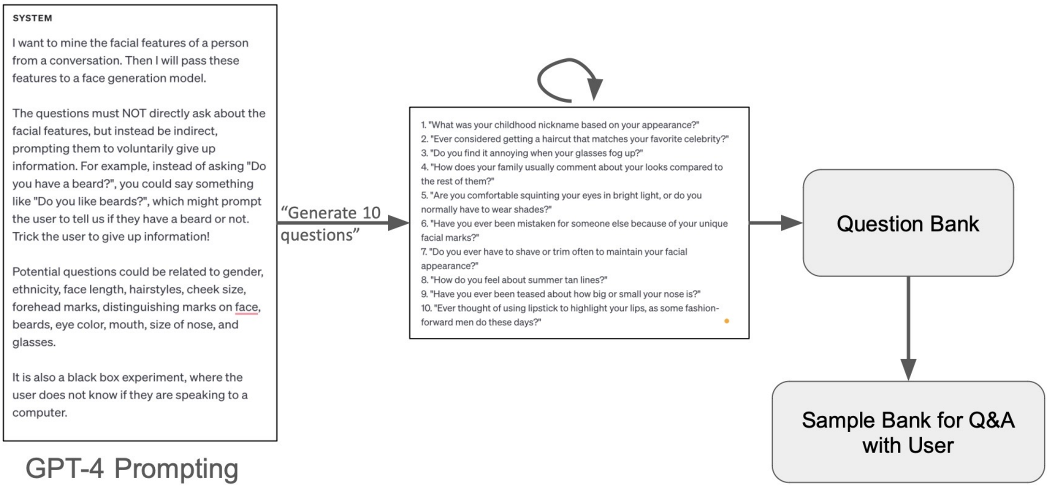
3.1 Question Generation

The objective here is to mine physical characteristics and personality traits from a one-to-one personal conversation with a human that could be used to accurately reconstruct their face. A key objective is also to mine information in a way that is not too direct or intrusive. For instance, in Figure 1, notice that the agent phrases a query asking the person for their opinion on beards, rather than directly asking them if they have one. In response, the person voluntarily gives up information about their appearance. We wish to maintain this objective, as an important goal in human-robot interaction is for agents to converse with humans in a manner that seems natural, allowing a seamless integration into human environments [André and Pelachaud 2010; Bonarini 2020; Skantze 2021].

To create a diverse set of non-invasive questions, we create a system prompt providing context to GPT-4 and prompt it to generate 10 questions with the goal of implicitly extracting a user's facial features. We repeat this procedure until we acquire a list of 50 questions from which we can then sample during the conversation of the LLM with a human. Figure 2 illustrates the question generation pipeline.

3.2 Pilot Study

Our problem setting requires a user to have a conversation with an embodied agent utilizing an LLM to interpret dialogue. To simplify experimentation, we design a pilot study resembling a Turing Test [Turing 2009] style setup, where the human test subject is speaking to a computer-generated voice translating the LLM's queries via *text-to-speech*. The study has the form of a Wizard of Oz



SYSTEM

I want to mine the facial features of a person from a conversation. Then I will pass these features to a face generation model.

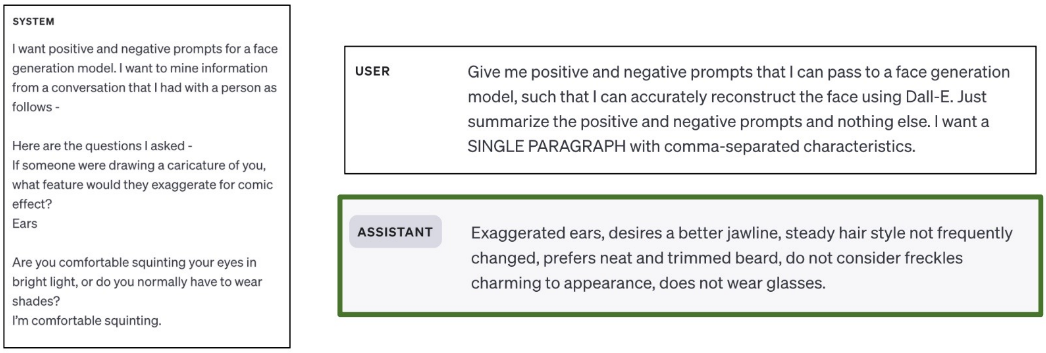
The questions must NOT directly ask about the facial features, but instead be indirect, prompting them to voluntarily give up information. For example, instead of asking "Do you have a beard?", you could say something like "Do you like beards?", which might prompt the user to tell us if they have a beard or not. Trick the user to give up information!

Potential questions could be related to gender, ethnicity, face length, hairstyles, cheek size, forehead marks, distinguishing marks on face, beards, eye color, mouth, size of nose, and glasses.

It is also a black box experiment, where the user does not know if they are speaking to a computer.

1. "What was your childhood nickname based on your appearance?"
2. "Ever considered getting a haircut that matches your favorite celebrity?"
3. "Do you find it annoying when your glasses fog up?"
4. "How does your family usually comment about your looks compared to the rest of them?"
5. "Are you comfortable squinting your eyes in bright light, or do you normally have to wear shades?"
6. "Have you ever been mistaken for someone else because of your unique facial marks?"
7. "Do you ever have to shave or trim often to maintain your facial appearance?"
8. "How do you feel about summer tan lines?"
9. "Have you ever been teased about how big or small your nose is?"
10. "Ever thought of using lipstick to highlight your lips, as some fashion-forward men do these days?"

Fig. 2. We build a question bank by prompting GPT-4 to generate questions with the goal of implicitly mining a user's facial features. We sample from this question bank during the live conversation at the pilot study.



SYSTEM

I want positive and negative prompts for a face generation model. I want to mine information from a conversation that I had with a person as follows -

Here are the questions I asked -
If someone were drawing a caricature of you, what feature would they exaggerate for comic effect?
Ears

Are you comfortable squinting your eyes in bright light, or do you normally have to wear shades?
I'm comfortable squinting.

USER Give me positive and negative prompts that I can pass to a face generation model, such that I can accurately reconstruct the face using Dall-E. Just summarize the positive and negative prompts and nothing else. I want a SINGLE PARAGRAPH with comma-separated characteristics.

ASSISTANT Exaggerated ears, desires a better jawline, steady hair style not frequently changed, prefers neat and trimmed beard, do not consider freckles charming to appearance, does not wear glasses.

Fig. 3. We pass the transcript of a conversation to GPT-4 and instruct it to generate positive and negative prompts for a diffusion-based face generation model.

experiment [Riek 2012] as the participants assume they are conversing with an autonomous agent, but in reality the questions are pre-generated via GPT-4, and a moderator has to manually advance to the next question or skip a question if it's deemed potentially uncomfortable for the participant. To make participants feel more comfortable speaking about themselves, but also to maintain the the Wizard of Oz setting, we add a separator between the participant and the moderators. In addition to that, we add a whiteboard with instructions for the participant, as well as a standing mirror, since some of the generated questions require a bit of pondering about the physical characteristics of a participant, or a potential change in their appearance (e.g. a change of hair style). A bluetooth speaker is used for asking the questions via *text-to-speech*. We record the audio of the answers given via a microphone and convert it to text via *speech-to-text*. The context of our experimental setup has been defined in the system prompt(See Figure 2). At the same time, we also grab an image of the face of the subject for evaluating our face-generation model later on. The setting of our pilot study can be seen in Figure **??**.

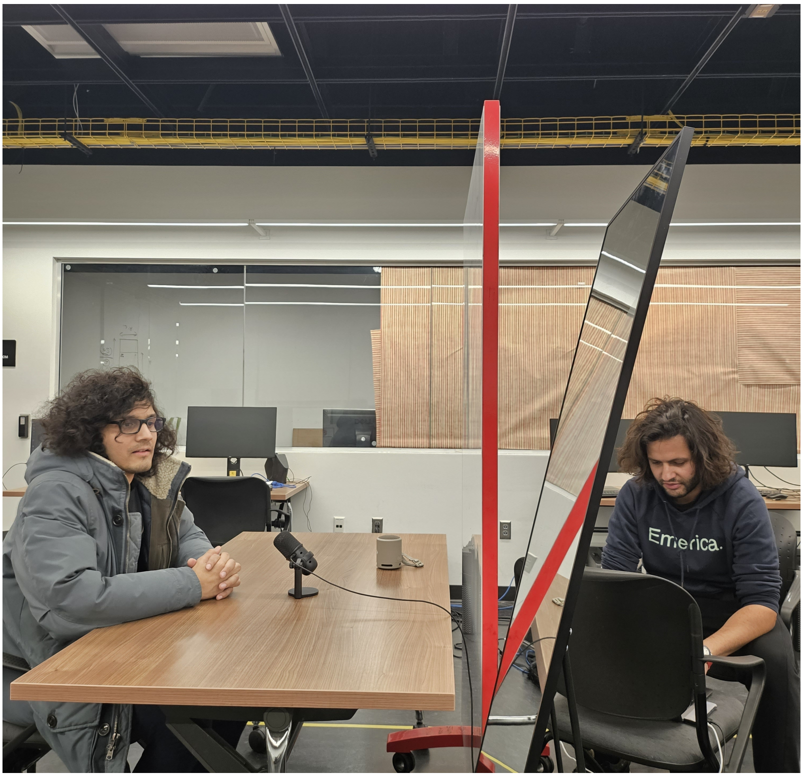


Fig. 4. The setting of our pilot study. The participant is sitting on the left, along with a speaker and microphone for posing the questions through *text-to-speech* and capturing the answers with *speech-to-text*. The moderator is sitting on the right side. In between, there is a separator with a whiteboard including instructions for the experiment, and a mirror for the participant.

3.3 Face Generation

After acquiring a transcript of the conversation between the LLM and the participant, we listen to the recording to correct any incorrectly captured parts of the transcript from the *speech-to-text*. We then pass it to GPT-4 and instruct it to generate positive and negative prompts for a diffusion model that generates images of faces based on a textual conversational input. The prompt for this task, along with an example answer from GPT-4 is shown in Figure 3.

The objective of our Face Generation Model is to produce diverse face samples based on the summary of the physical characteristics and personality that we obtain during conversation. We use DALL-E 3 [Betker et al. 2023; Ramesh et al. 2022] to generate faces. This model takes text as input to generate various types of images following a latent diffusion-based approach. we use the following prompt template: *A hyper realistic waist-up portrait of a <GENDER> face with <Comma Separated Features>*. Overall, our method can be summarized in the following steps:

(1) Generate 50 non-intrusive questions to implicitly extract user facial features
(2) Randomly sample 10 questions per user and ask with text-to-speech
(3) Record answers with a microphone and transcribe using speech-to-text
(4) Post-process transcript and summarize with GPT-4
(5) Prompt DALL-E with the summary for face generation

(6) Qualitatively evaluate the generated faces

4 RESULTS & DISCUSSION

We run the pilot study with 7 participants (2 females, and 5 males). We ask 10 randomly picked queries from the Question Bank generated (shown in figure 2), and transcribe each of the recordings manually. We then pass these transcriptions to GPT's system prompt and ask for a summary containing positive and negative prompts that we can pass to our face generation model Dall-E 3 [Betker et al. 2023]. We make the following inferences:

4.1 Strange Annotations

We notice that in some of the generated images, there are annotations pointing either to specific text, or zoomed in parts of the face. This is illustrated in figure 5. We notice this is prominent mainly with Asian faces (Chinese, Korean and Indian), where data might have been mostly scraped from publicly available medical or legal documents due to country-specific internet restrictions.

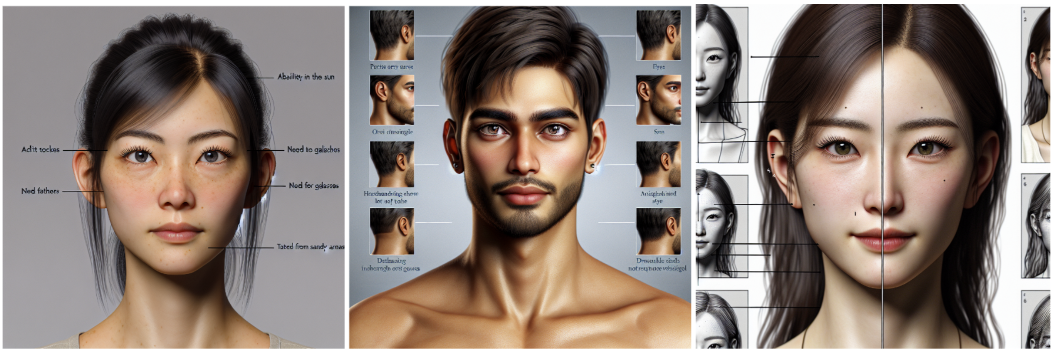


Fig. 5. **Strange Annotations**: We notice that generated images sometimes have strange annotations associated with them. Moreover, we notice this to be prominent with people of Asian ethnicities. This indicates a bias in the training data, where access to Asian-related content might have been limited, forcing data to be scraped from publicly available medical or legal documents where such annotations are common.

4.2 *"Beard-Bias"*

On specifying the geographic origins of our test candidates, we notice a "*Beard-Bias*" on male participants. Despite explicitly mentioning clean-shaven appearance in prompts, DALL-E tends to not pay attention to this and generates beards for people from a region where beards a popular. This is illustrated in figure 6. The prompt for this generation explicitly states *"Male face that DOES NOT HAVE ANY FACIAL HAIR"*, while also mentioning the origin as *"from Visakhapatnam, India"*, a region where beards are very common. This shows bias, in that the model overlooks the fact that the individual in question does not have facial hair, and generates one anyway due to his place of origin. This seems to be a common occurrence and has been cited in some prior studies as well [1][2].

4.3 Feature Exaggeration

In some cases, DALL-E tends to place too much emphasis on marks or unique identification features that a participant describes their face. For instance, if the participant has mentioned having some

[1] https://www.96layers.ai/p/ai-loves-beards

[2] https://edwardv.com/cs229_project.pdf

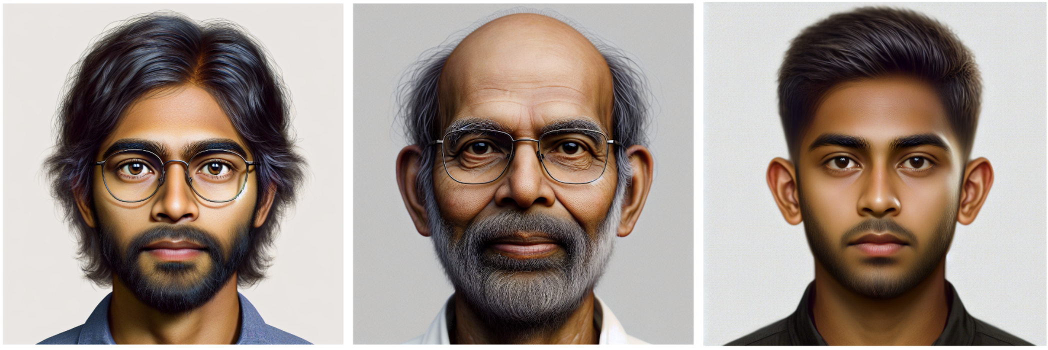


Fig. 6. **Beard Bias**: All these images were created from similarly worded prompts, modifying the explicit specification of the lack of facial hair in different ways. Notice that the model for the most part tends to overlook this, and instead produces stereotypical images of a male face from the native region of the participant (Visakhapatnam, India in this case).

moles on their face, the model tends to create a face with too many moles. Examples of this are illustrated in figure 7. We notice that even explicitly specifying the type of mark (number or size) is a hit and a miss.

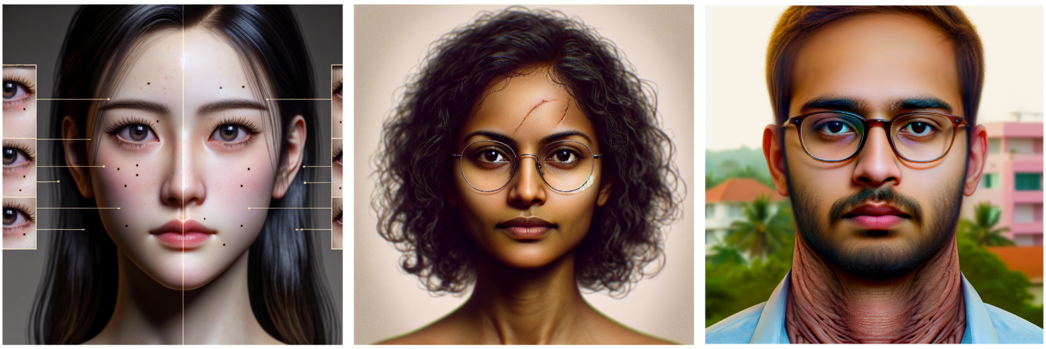


Fig. 7. **Feature Exaggeration**: We notice that DALL-E tends to exaggerate *marks* or *identifying facial features* that the participant describes. Notice the excessive number of moles in the left image, the longer scar marks in the center image, and the horrific looking *"circular scar on neck"* on the rightmost image. We can infer that the model lacks an accurate sense of what a real-world face might look like, instead over-hallucinating on features that were given to it, leading to unnatural images.

4.4 Positive Outcomes

We sought to uncover if providing the name of the participant would bring about bias and influence the facial generation in any way. DALL-E however ensures that the names of the people in the prompt are anonymized in its reformulation, and warns us about sharing intimate details. We believe this to be a positive inference, in OpenAI taking some responsibility to mitigate bias in their models.

Moreover, not all the generated images were bad, with some of them even resembling the participant's faces to good extent. Some examples of this are shown in figure 8.

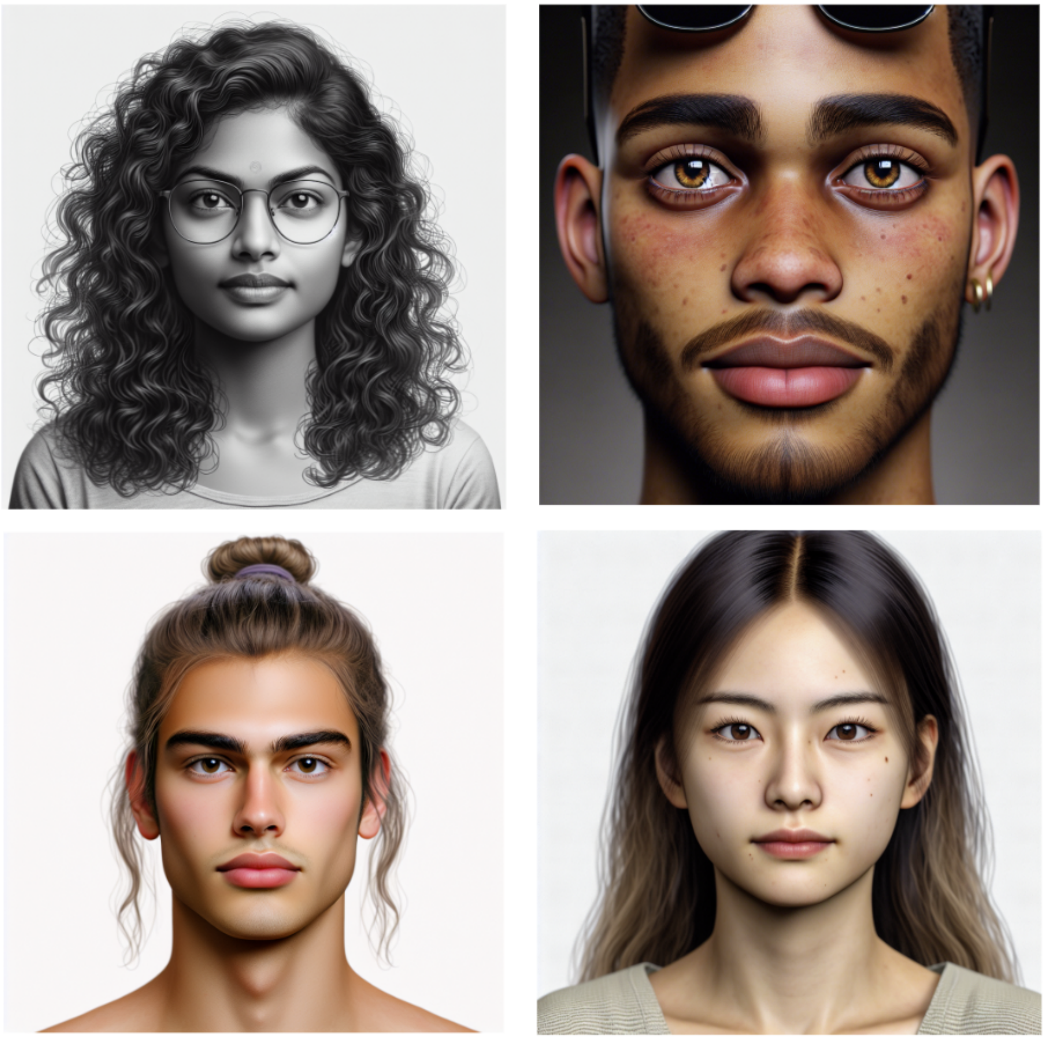


Fig. 8. **Positive Inferences**: Not all the images generated were bad. We did receive some positive feedback from participants about how much the generated face correlates with their true appearance. For instance, the participant on the top right captured that the model got the beard correctly, but was not accurate with the eyes — in part because they mentioned an anecdote about them having swollen eyes at one point. The participant on the top left was also surprised with how accurately it captured the hairstyle that they spoke about. While biases may always exist, we believe prompting for more descriptive answers would improve performance.

5 CONCLUSION

Our experimental pilot study provided some preliminary insights on the fairness of diffusion-based models on face generation. It also highlighted that there is room for improvement on several parts of our pipeline. Our first step would entail restructuring the question generation strategy. Because the LLM was instructed to generate non-intrusive questions, some of them ended up leading to user answers that did not contribute any significant information towards reconstructing their facial features. Furthermore, randomly sampling from a set of questions might neglect salient question categories that are essential for obtaining an accurate face generation. We observe that the generated questions tend to fall into 5 categories - *Ethical, Expressions, Fitness, Grooming, and Look-alikes.* Therefore, we could categorize the generated questions into these 5 types and ensure that there is a representative number of questions for each of these categories per user.

We also discover multiple biases in the face generation model that correlate both directly and indirectly with the prompts given to it. Exploring other open-source face generation models and in-context learning for prompts is part of future study.

Evaluation is currently based on visual observation and qualitative comparison of the generated image with the original for a small number of participants (7). Drawing statistically significant conclusions would require a larger number of participants and metrics quantifying the bias reflected by the generated images.

Finally, it is hard to implicitly capture adequate information for all the salient facial features required for accurate face generation in such a limited amount of time. In a household environment, a personal robot will interact with a human user for a significantly larger amount of time, so we expect that more conversational data gathered will lead to more accurate face generation.

REFERENCES

Elisabeth André and Catherine Pelachaud. 2010. Interacting with embodied conversational agents. *Speech technology: Theory and applications* (2010), 123–149.

DMA Ayanthi and Sarasi Munasinghe. 2022. Text-to-face generation with stylegan2. *arXiv preprint arXiv:2205.12512* (2022).

John Baugh. 2005. Linguistic profiling. In *Black linguistics*. Routledge, 167–180.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* (2023).

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1493–1504.

Andrea Bonarini. 2020. Communication in human-robot interaction. *Current Robotics Reports* 1 (2020), 279–285.

Ali Borji. 2022. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586* (2022).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

Daria Bylieva, Zafer Bekirogullari, Victoria Lobatyuk, and Natalia Anosova. 2020. Home assistant of the future: What is it like?. In *Proceedings of the International Scientific Conference-Digital Transformation on Manufacturing, Infrastructure and Service*. 1–8.

Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. 2017. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3703–3712.

Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893* (2023).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. 2023. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6080–6090.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1548–1558.

Muhammad Zeeshan Khan, Saira Jabeen, Muhammad Usman Ghani Khan, Tanzila Saba, Asim Rehmat, Amjad Rehman, and Usman Tariq. 2020. A realistic image generation of face from text description using the fully trained generative adversarial networks. *IEEE Access* 9 (2020), 1250–1260.

Gary B. Huang Erik Learned-Miller. 2014. *Labeled Faces in the Wild: Updates and New Reporting Procedures.* Technical Report UM-CS-2014-003. University of Massachusetts, Amherst.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV).*

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Evaluating Societal Representations in Diffusion Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

Osaid Rehman Nasir, Shailesh Kumar Jha, Manraj Singh Grover, Yi Yu, Ajit Kumar, and Rajiv Ratn Shah. 2019. Text2facegan: Face generation from fine grained textual descriptions. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM).* IEEE, 58–67.

Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. 2019. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 7539–7548.

Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015.* British Machine Vision Association.

Malsha V Perera and Vishal M Patel. 2023. Analyzing bias in diffusion-based face generation models. *arXiv preprint arXiv:2305.06402* (2023).

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

Céline Ray, Francesco Mondada, and Roland Siegwart. 2008. What do people expect from robots?. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems.* IEEE, 3816–3821.

Laurel D Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 10684–10695.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 22522–22531.

Rita Singh. 2019. *Profiling humans from their voice.* Vol. 41. Springer.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.

Ja-Young Sung, Rebecca E Grinter, Henrik I Christensen, and Lan Guo. 2008. Housewives or technophiles? Understanding domestic robot owners. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction.* 129–136.

Alan M Turing. 2009. *Computing machinery and intelligence.* Springer.

Yandong Wen, Bhiksha Raj, and Rita Singh. 2019. Face reconstruction from voice using generative adversarial networks. *Advances in neural information processing systems* 32 (2019).

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419* (2023).