Fairness in Diffusion Models

David Yang

Project in collaboration with Amy Feng, Alexander Goslin, and Selena She @ REU-CAAR (University of Maryland, College Park).

Abstract

- Generative Artificial Intelligence models have quickly grown in popularity, and their use has become very widespread. One such model, a diffusion model, generates an image for a given text prompt. Since these models are trained on large-scale datasets, they inherit many biases implicit in this data, thereby enforcing societal stereotypes.
- In this project, we focus on images of people generated by Stable Diffusion. We aim to address gender and race stereotypes by modifying the image output of these diffusion models to yield a more fair and general representation of people of all genders and races.

Motivation

- Generative AI such as Diffusion Models suffer from implicit biases (gender, race, age stereotypes).
- Previous work such as FairDiffusion focuses on individual fairness, as the most commonly used large face dataset, FairFace, contains just single-person images.

Figure: A previous approach, FairDiffusion [2], handles individual images well, using Sega [1].

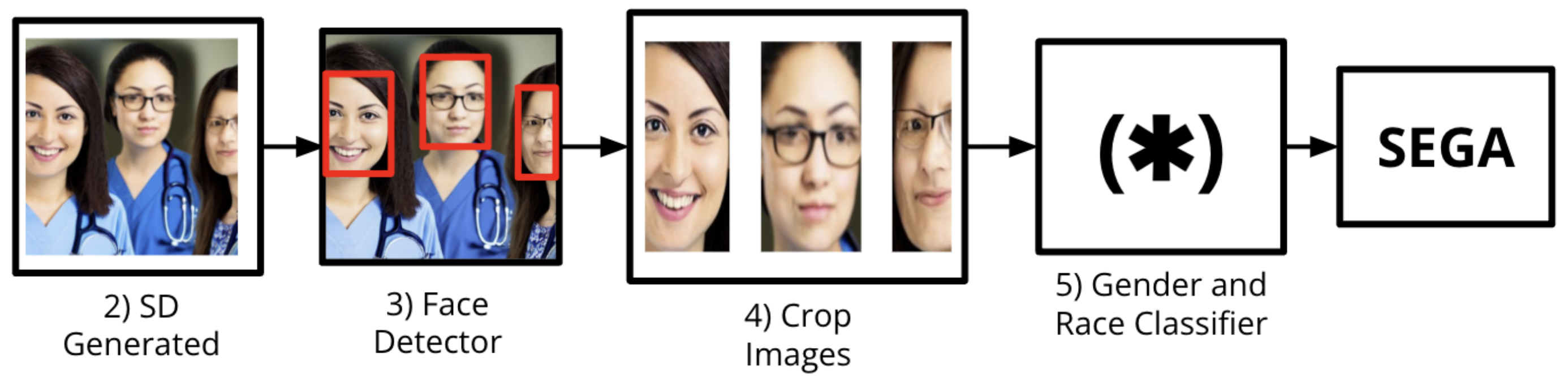
- We are interested in improving fairness in diffusion models and extending previous work to solve fairness in group images.

Acknowledgments

Thank you to my mentor Furong Huang and the REU-CAAR Program for making this research possible.

General Pipeline

1) prompt: *"a photo of three female doctors"*



(✱) Average Confidences

Man	Woman
29.5	70.5

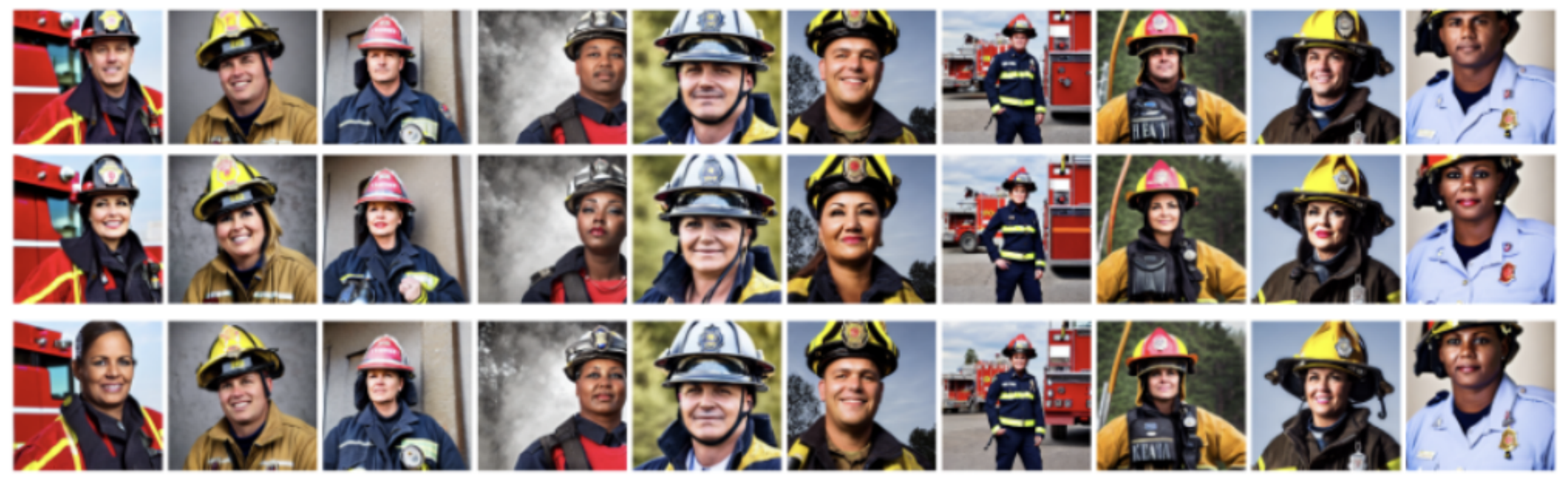
Gender

White	Latino Hispanic	Black	Asian	Middle Eastern	Indian
37.2	28.2	13.6	0.8	16.3	3.9

Race

Results (Individual Images)

prompt: *"a portrait photo of a firefighter"*



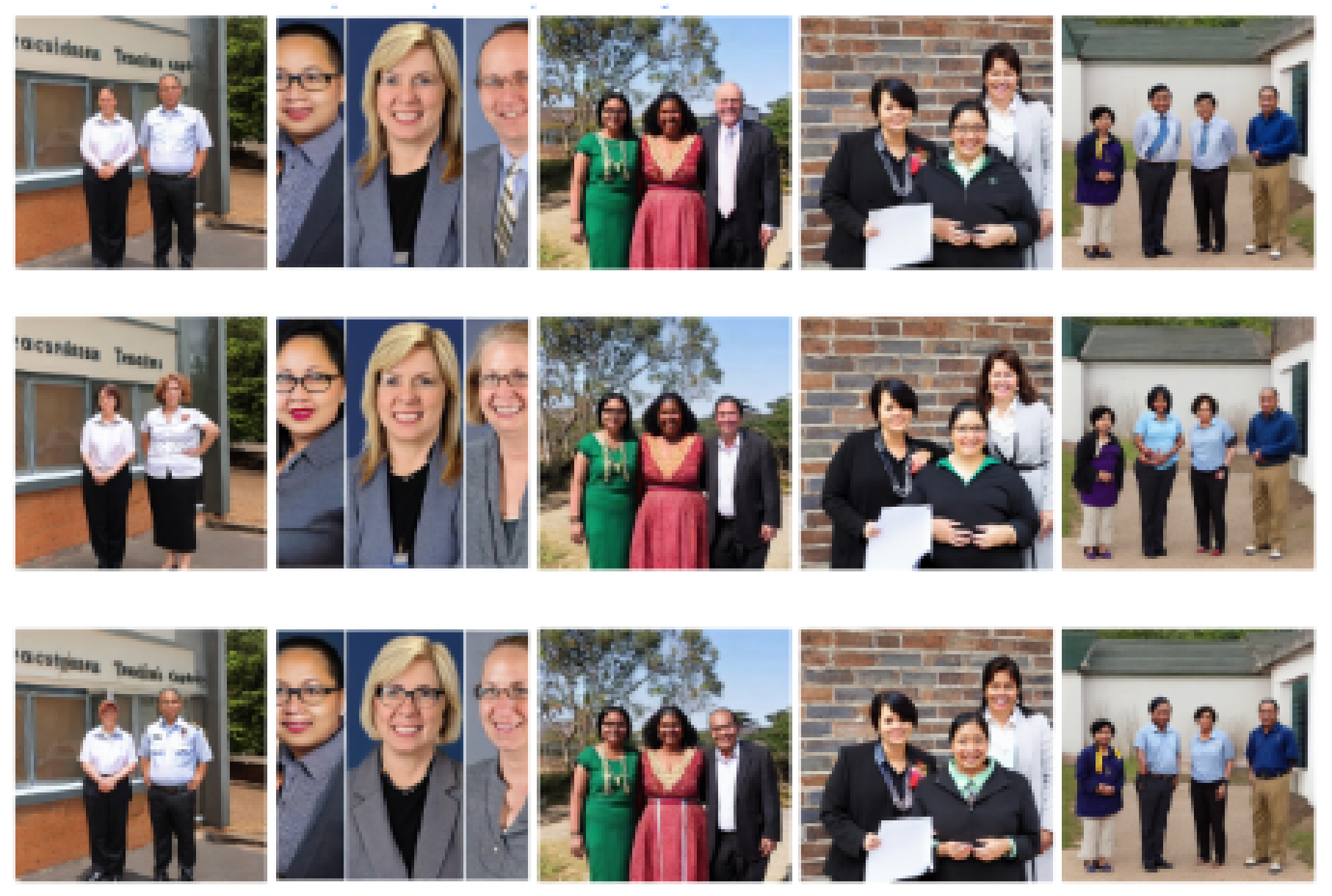
Row 1: SD Generated | *Row 2*: "Fair Diffusion" Generated | *Row 3*: Our approach.

+ female + male
edit weights: [1, -1]

+ female + male
edit weights: [0.98, 0.02]

Results (Group Images)

prompt: *"a photo of three teachers"*



Methodology and Approach

- The main tool we use is known as Semantic Guidance (Sega) [1]. This approach uses text descriptions to "guide" concepts in the text embedding space. The changes are isolated to the guided concepts and do not affect the remainder of the image.

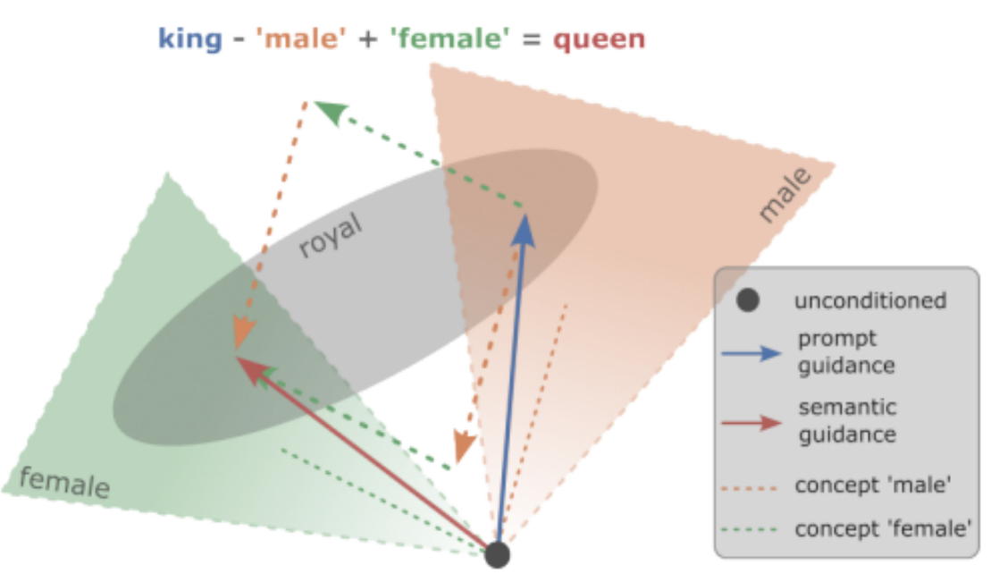


Figure: Sega's effect on the embedding space

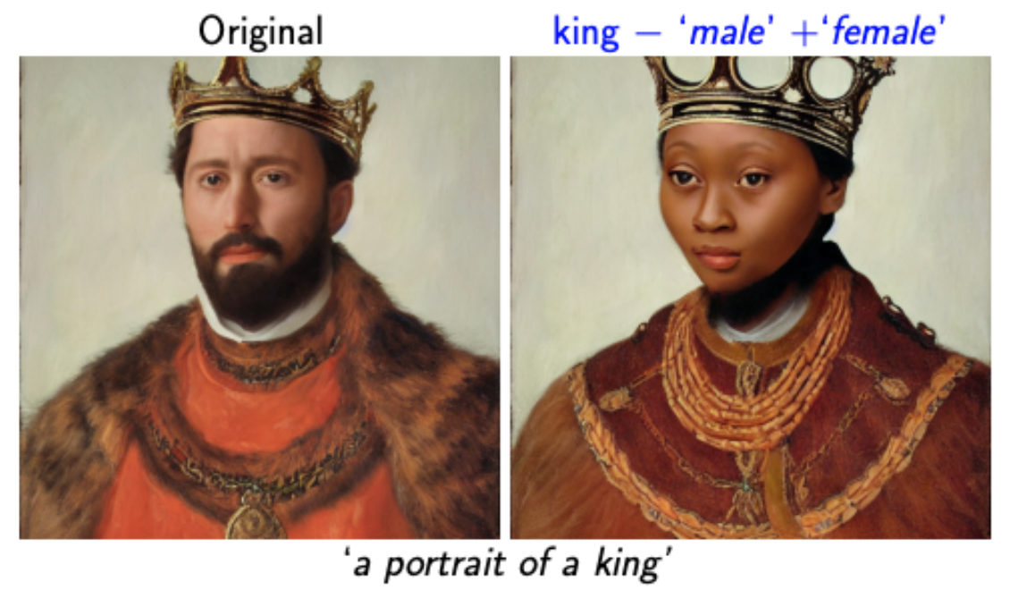


Figure: Example Output of Sega

Limitations

In the generative process, we find

- Deformed/Blurry Faces in Group Images
- Trying to improve fairness seems to adversely affect image quality
- Sometimes cannot properly determine race/gender to evaluate approaches

In our general pipeline, we use DeepFace as a gender and race classifier; this is itself biased and faulty with respect to classification.

Conclusion & Future Work

- Our current approach is slow and only works as post-processing, with varying degrees of success.
- Ultimately, we want an approach that makes the model itself more fair; this could be done by fine-tuning the model's internal parameters.

References

[1] Manuel Brack et al. *SEGA: Instructing Text-to-Image Models using Semantic Guidance*. 2023. arXiv: 2301.12247 [cs.CV].

[2] Felix Friedrich et al. *Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness*. 2023. arXiv: 2302.10893 [cs.LG].