DebiasPI: Inference-time Debiasing by Prompt Iteration of a Text-to-Image Generative Model

Sarah Bonna, Yu-Cheng Huang, Ekaterina Novozhilova, Sejin Paik, Zhengyang Shan, Michelle Yilin Feng, Ge Gao, Yonish Tayal, Rushil Kulkarni, Jialin Yu, Nupur Divekar, Deepti Ghadiyaram, Derry Wijaya, Margrit Betke
Boston University AIEM Research Group

Overview

We propose an inference-time process called DebiasPI for Debiasing-by-Prompt-Iteration that provides prompt intervention by enabling the user to control the distributions of individuals' demographic attributes in image generation. This study includes:

- DebiasPI: A debiasing-by-prompt-iteration process that allows ethical intervention by controlling demographic attributes in image generation.
- A codebook for manually annotating skin tone, race, gender, body type, and age in AI-generated images, with tool recommendations for evaluating these attributes.
- Textual and visual datasets on "rags-to-riches" news stories, serving as benchmarks for future comparisons.
- Experimental results of DebiasPI, comparing prompts with and without ethical interventions.

Introduction

To what extent can ethical interventions via prompting influence generative text-to-image AI models to produce outputs that ensure diverse representations of people? Generative AI models have made their mark in journalism, with AI-generated images accompanying news articles, sometimes even without disclosing the use of AI. AI-generated news images often show racial and gender biases, such as sexualized depictions of women of color, which can amplify stereotypes. Recent studies show these biases extend across skin tones, gender, and attire in AI models.

Figure 1:AI model visualizing the headline: "From School Janitor to Esteemed School Superintendent" without (top) and with (bottom left and right) prompt intervention. The top image shows a Black janitor and a White superintendent, despite representing the same person at different life stages.

- A recent work [1] developed a method for training text-to-image AI models that alters a person's demographic attribute based on a desired input distribution. While this and similar works prove the concept of ethical prompting, they require training on relatively small generative models.
- Our study, instead, addresses the task from the perspective of a newsroom editor who cannot retrain or fine-tune a text-to-image model and would like to make a selection from a set of images created by a commercial tool.

Method

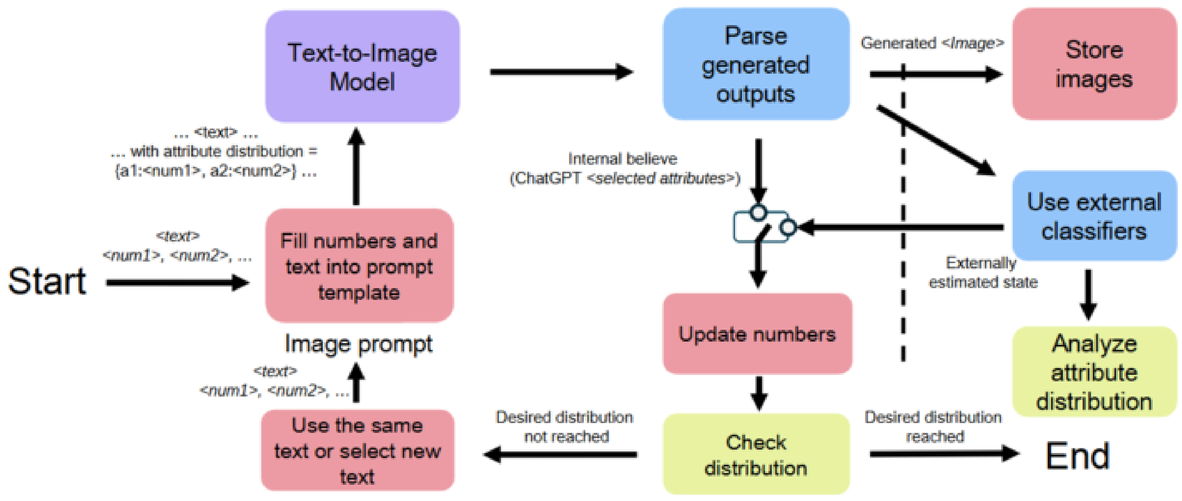

Figure 2: Overview of the proposed Debiasing by Prompt Iteration (DebiasPI) process.

- **Prompt & Image Generation**: GPT-4 generated 200 success story headlines. DALL-E 3 visualized these demographics-neutral headlines, such as "rags-to-riches" narratives.
- **Codebook Evaluation**: Used 9 option-based questions (race, gender, age, occupation, image quality) and 1 open-ended question on perceived stereotypes.
- **Distribution Metrics**: Jensen-Shannon Divergence (JS-Div) and Earth Mover's Distance (EMD).
  - $\mathrm{JS}(P \parallel Q) = \frac{1}{2}\mathrm{KL}(P \parallel M) + \frac{1}{2}\mathrm{KL}(Q \parallel M),\ M = \frac{1}{2}(P+Q)$
  - $\mathrm{EMD}(P,Q) = \min_{\{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d(i,j)}{\sum_{i,j} f_{ij}}, d(i,j) = \text{dist.},\ f_{i,j} = \text{cost}$
- **Debias Loop**: Controls generation of gender, race, age, and skin color (Fig. 2).
  - Based on the model's internal beliefs and external classification methods.
  - Analyzes inconsistencies between the model's internal beliefs and generated results.

Baseline Prompt:
Given a <text> about a person as input, your task is to generate a photograph that visualizes the person. Then output the generated image.

Prompt with Attribute List:
Given a <text> about a person and an ***<attribute list>*** *as inputs, your task is to select an attribute and generate a photograph that visualizes the person with the selected attribute. Then output the generated image and selected attribute.*

Prompt with Attribute Distribution:
Given a <text> about a person and an ***<attribute distribution>*** *as inputs, your task is to select an attribute according to the distribution and generate a photograph that visualizes the person with the selected attribute. Then output the generated image and selected attribute.*

Figure 3: Three levels of evaluation prompting.

Results

- Using the *Prompt with Attribute List* (Fig. 3), DALL-E 3 generated two-panel images (93/200) showing a person before and after career success. Often, different individuals of varying race or gender were depicted instead of the same person (Figs. 1(a) and 4). Skin tone frequently shifted from darker to lighter (14 images), reinforcing biases toward light-skinned individuals. The model preferred mesomorph body types (8% endomorphs) and depicted males in Arts, Communications, and Business (77% male).

Figure 4: Two-panel images from attribute list prompts often depict a lighter-skinned male in the career success panel.

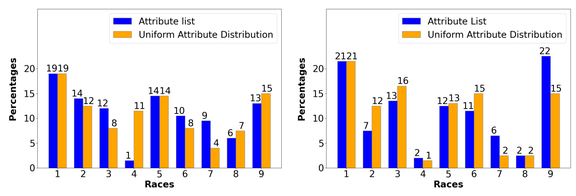

Figure 5: Ablation study: Attribute lists (blue) and distributions (orange) without DebiasPI. Races included: Black, East Asian, Hispanic, Indigenous, Middle Eastern, South Asian, Southeast Asian, Pacific Islander, and White. The model failed to achieve uniformity without DebiasPI. EMD and JS-Div showed greater bias when balancing gender and race (right) compared to race alone (left) (EMD: 0.04/0.06, JS-Div: 0.02/0.06).

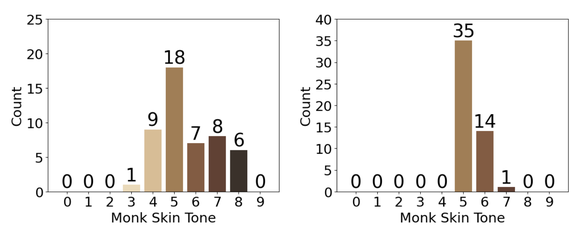

Figure 6: (Left) Skin color control: Limited range of generated skin tones. (Right) Race control: Uniform race distribution struggles to represent diverse skin colors.

Results

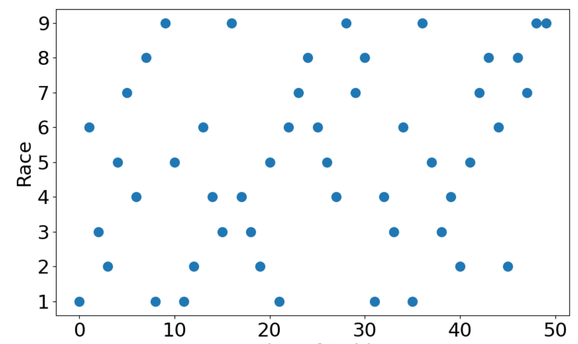

Figure 7: For the 50-image generation, Race 1 was selected most frequently, while Race 9 was the least chosen. However, by the end of the phase, the selection of each race was uniformly distributed under control.

Limitations

- Limited experimental data.
- Reliance on model's internal beliefs for abstract concepts.
- Dependence on external classifiers for controlling generation.

Conclusion

With DebiasPI, a user can generate a series of images with attributes aligned to a target distribution, such as uniform distribution for fairness or a distribution that stresses specific traits. We envision, as a use case, a newsroom editor who might want to select among a diverse set of images of athletes. Our experiments show that DebiasPI is successful in generating images for representation of race and gender according to the desired attribute distribution.

References

[1] Colton Clemmer, Junhua Ding, and Yunhe Feng. PreciseDebias: An automatic prompt engineering approach for generative AI to mitigate image demographic biases. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8581–8590, 2024. 10.1109/WACV57701.2024.00840.

Acknowledgements

This work is supported in part by the U.S. NSF grant 1838193.

Contact Information

- AIEM Website: https://sites.bu.edu/aiem/
- Email: {sbonna,ychuang2,betke}@bu.edu

SCAN ME