Words Matter: Reducing Social Bias in AI Image Generation via Prompt Design

1. Overview

Our project explores social biases embedded in AI-driven text-to-image generation (T2I) models—specifically how these models can produce biased outputs regarding gender, skin tone, and culture. These issues have drawn increasing attention as generative models like **Stable Diffusion**, **DALL·E 2**, and **Midjourney** become integrated into mainstream creative tools. While their outputs are often visually impressive and coherent, they also reflect societal stereotypes present in the training data, resulting in skewed portrayals when prompts are ambiguous or seemingly neutral (e.g., “a doctor” or “a teacher”).

Our work builds upon recent studies that examine the extent and form of these biases, but it focuses on a practical, low-effort solution: **prompt enhancement** using ethically guided interventions. Instead of proposing deep architectural changes to the models or training pipelines—which would be too complex for a semester-long course project—this research investigates how rewording the input prompts using inclusive or counter-stereotypical language can impact the diversity and fairness of generated images.

Our project will implement and test a small set of prompts under two conditions:

1. **Baseline prompts** that are general and unconstrained (e.g., “an image of a CEO”).

2. **Enhanced prompts** that include ethical cues or inclusive framing (e.g., “an image of a CEO, regardless of their gender or race”).

Images generated under both settings will be collected, analyzed, and compared across visual attributes such as gender appearance, skin tone, and cultural cues and evaluated by multiple benchmarks such as Inception Score and CLIP Score. This comparison will help determine whether such linguistic interventions lead to more equitable representations, and if so, in which types of prompts or use cases they are more effective.

2. Research Questions

This project aims to answer the following core questions:

- **RQ1:** What forms of social bias (gender, racial/skin tone, cultural) appear in outputs of text-to-image generation models when provided with neutral prompts?

- **RQ2:** Can prompt enhancement using simple, ethically framed language significantly reduce or balance the representation of social groups in these generated images?

- **RQ3:** How consistent are the effects of such prompt modifications across different types of prompts (e.g., occupations, traits, daily life scenarios)?

These questions are framed to be testable within a short time frame and align with recent literature calling for more prompt-level interventions as lightweight, scalable solutions for bias mitigation.

3. Value to User Community

This project is designed to provide immediate, practical insights for multiple communities working with generative AI:

For Developers:

They can benefit from understanding how prompt phrasing impacts output fairness, which may inform better user-facing guidance or UI design for generative tools. Rather than requiring model retraining or access to private datasets, prompt enhancement offers a **cost-effective** and **user-friendly** solution for mitigating bias at the point of use.

For Researchers:

This work contributes to a growing body of evidence on bias behavior in generative AI systems. While much prior work has focused on identifying the presence of bias and worked on old models, this project engages with mitigation strategies upon current trending models, adding value through its comparative analysis of prompt variants. It also explores whether such changes generalize across social categories and use cases.

For Content Creators and End Users:

As tools like Stable Diffusion and DALL·E are increasingly used by artists, marketers, designers, and educators, the ability to consciously steer outputs away from biased defaults without complex technical expertise is critical. This research highlights how to ask better questions so that users can receive fairer, more inclusive visual content.

Ultimately, the goal is to bridge the gap between theoretical concerns and real-world applications by showing how small changes in user input can lead to meaningful improvements in AI output. This democratizes fairness in AI by giving more control to users, even when the models themselves remain opaque or unchangeable.

4. Plan on pitch and demo

For the pitch, we will briefly introduce our topic, background, and experiment. In the demo, we would like to first illustrate a problematic scenario, then employ our prompt intervention technique and compare side by side using evaluation metrics. Finally, we analyze and summarize the effectiveness of our technique.

5. Delivery

If the experiments include any possible scripts, seed images, or benchmark samples, they will all be uploaded to the GitHub repository.
https://github.com/JunboHe-David/ai-image-generation-6156-final-project

6. Additional Links

To know more background information about evaluation benchmarks, the following links are helpful.

https://arxiv.org/pdf/2104.08718 CLIP score
https://arxiv.org/pdf/1801.01973 Inception score