**Mitigating Gender Bias in Text-to-Image Generation: A Fine-tuning method**

*Group members: Jiani Ji, Julia Zhu, Danny Yue, Wenshu Zhao, Felix Zhao*

# Introduction

Text-to-image models such as Stable Diffusion exhibit significant capabilities; however, they frequently reflect and amplify societal biases. An example of gender bias is illustrated as follows: Stable Diffusion generated predominantly female representations for the prompt “compassionate manager” (left), while exclusively male representations were produced for the prompt “manager” (right) (Jun, 2024).

A collage of people in different poses

AI-generated content may be incorrect.

# One big problem is representational bias, which happens when outputs favor or stereotype certain groups of people more than others. For example, starting Stable Diffusion with a neutral word like "a person" leads to mostly male images (about 65% in one study). Similarly, queries for high-status jobs like "doctor" or "CEO" also lead to images of men (Rahwan, 2025). On the other hand, prompts that use words that are usually associated with women, like "compassionate manager," tend to bring up images of women (see figure above). These kinds of biases don't just affect the gender of the person being portrayed; they can also affect the scene and things in the picture. Recent research showed that changing the gender word in a prompt changes more than just the subject's gender. It also changes background details, like the instruments or layouts, even if those details weren't given. These hidden connections show how unique the problem is: unlike classification bias, generative bias can show up in small visual clues and scene compositions, which makes it harder to spot and fix.

# Our motivation

Our motivation is clear. Images made by AI that are biased could reinforce harmful stereotypes and social inequality. Visual gender stereotypes can reinforce skewed ideas. For example, if most "doctor" pictures are of men and most "nurse" pictures are of women, it reinforces old gender roles. This shows biases that people already have and may also change the way people feel about it. Text-to-image tools are used by millions of people every day, so it's important to make sure they produce fair and varied images (Rahwan, 2025). . The problem is hard because it needs to be solved while still being realistic (showing how the world is distributed) and fair (not making prejudice worse). We use gender bias as an example of representational bias in this work and suggest a way to automatically find and fix this kind of bias in image creation. We want to make generative models more fair while keeping their creative integrity by carefully reviewing the model's outputs and adding targeted debiasing interventions.

# **Phase 1: Pre-processing & Image Generation**

To initiate our investigation into demographic bias in generative models, we first established a controlled environment for image generation using Stable Diffusion. The experiment was conducted on a cloud-based GPU instance using Google Colab, leveraging an NVIDIA A100-SXM4-40GB GPU, with CUDA 12.1 support. The environment was set up with essential libraries including diffusers, transformers, and torch, ensuring compatibility with the latest versions of Stable Diffusion pipelines.

We used the \*[Stable Diffusion v1.5 / v2.1] model hosted by Hugging Face, chosen for its balance between image quality and inference efficiency. The model was loaded in half-precision (fp16) to optimize GPU memory usage. For each prompt, we generated \_\_ images using a guidance scale of \_\_, inference steps of \_\_, and a fixed random seed for reproducibility. The resulting images were saved locally in PNG format with filenames corresponding to their prompts and index values for tracking.

Prompt selection was a critical component of our design. We curated a list of diverse and neutral base prompts that commonly reveal demographic bias, such as occupational roles (e.g., "a software engineer", "a nurse", "a pilot") and neutral social scenarios (e.g., "a person walking in a park", "a student studying at a desk"). A total of \_\_ unique prompts were used, each designed to be demographically ambiguous. The aim was to observe how the model implicitly filled in demographic features such as gender or race without being explicitly told.

The generated image set served as the baseline, against which subsequent bias mitigation strategies were evaluated. The outputs of this phase were stored in a structured directory with corresponding metadata, allowing for systematic downstream processing and annotation in later phases of the pipeline.

# **Phase 2: Demographic Detection using FairFace**

# After generating images using Stable Diffusion, we moved to the demographic annotation phase using the FairFace model, a deep learning tool designed to classify race and gender across diverse human faces. This step was critical for analyzing the demographic distribution of the outputs and detecting potential biases introduced by the generative model.

# The generated images were first passed through a face detection module to isolate valid facial regions. FairFace was then applied to these cropped face regions, returning predictions for gender (male, female) and race (White, Black, East Asian, Southeast Asian, Indian, Middle Eastern, Latino-Hispanic, and others). These predictions were stored in structured JSON and CSV formats, enabling downstream statistical and visualization tasks. To ensure traceability, each prediction was associated with the image filename, generation prompt, and model configuration.

# Inherent Bias in AI: Why GenAI Still Reinforces Stereotypes

# Some challenges emerged during this phase. Artistic rendering styles—especially when faces were stylized, abstract, or partially obscured—occasionally led to false negatives or incorrect demographic predictions. To mitigate this, we filtered out low-confidence detections and optionally used ensemble voting where multiple face crops or augmentations were evaluated. In total, \_\_% of generated images contained at least one recognizable face, and among those, \_\_% yielded confident demographic predictions.

# This phase served not only as a diagnostic layer for bias detection but also as a feedback mechanism for guiding bias mitigation strategies (Phase 3). The demographic distributions obtained here were later used to compute representation imbalance, define bias scores, and conduct before-and-after comparisons of mitigation techniques.

# A grey background with text overlay AI-generated content may be incorrect.**Phase 3: Bias Optimization via Prompt Engineering and Output Filtering**

To mitigate demographic bias in image generation, we implemented a two-pronged strategy: prompt engineering and post-generation output filtering. This phase focused on guiding the Stable Diffusion model toward generating more demographically balanced results while preserving image quality and semantic alignment with the prompt.

In the prompt engineering subphase, we refined the original prompt set to include explicit demographic cues aimed at enforcing diversity. For instance, a generic prompt such as "a software engineer" was expanded to include variations like "a Black female software engineer", "an Asian male software engineer", and "a white nonbinary software engineer". By introducing these controlled demographic terms, we aimed to produce a broader and more inclusive representation in the output space. A total of \_\_ modified prompts were generated from the base set of \_\_ prompts.

In parallel, we applied output filtering as a reactive approach. Using DeepFace and FairFace results from the previous phase, we tagged each image with its predicted gender and race attributes. Then, we introduced a filtering pipeline that either selected a balanced subset of outputs or discarded images that overrepresented dominant demographic groups. For example, if over 70% of generated doctors were white males, the filter retained a proportional number of white male outputs while prioritizing retention of underrepresented demographics.

These two methods—prompt control and output filtering—were combined to create a Bias-Optimized Image Set. This set aimed to reflect a more equitable demographic spread for each concept while maintaining high-quality and semantically consistent imagery. Additionally, metadata such as inference time and filtering logs were stored for later comparison in the evaluation phase.

# **Phase 4: Evaluation**

To assess the effectiveness of our bias mitigation pipeline, we conducted a multi-faceted evaluation across both quantitative metrics and qualitative observations. The evaluation phase focused on five key dimensions: (1) realism, (2) stereotype presence, (3) demographic appropriateness, (4) response time, and (5) semantic fidelity using CLIP-based embedding analysis. Result as picture:

Realism was assessed using a human rating system and optionally an image realism classifier. A group of annotators rated each image on a Likert scale from 1 (unrealistic) to 5 (highly realistic). For automated comparison, we also computed the Fréchet Inception Distance (FID) against a curated dataset of real human portraits. The average realism score for the direct approach was \_\_, while the RAG-enhanced and bias-optimized sets scored \_\_ and \_\_, respectively.

To capture stereotype presence, we defined a set of profession-related tropes (e.g., depicting nurses as women, engineers as men) and manually reviewed images for stereotypical portrayal. We computed a Stereotype Rate as the proportion of images that reinforced such tropes. The baseline generation showed a stereotype rate of \_\_%, which reduced to \_\_% in the bias-optimized pipeline.

Demographic appropriateness was measured by comparing the predicted race and gender distribution of generated images (from FairFace) to reference demographics (\_\_) derived from external datasets or real-world statistics. Using Kullback–Leibler (KL) divergence and Earth Mover’s Distance (EMD), we quantified the deviation of image distributions from target distributions. The KL divergence decreased from \_\_ (baseline) to \_\_ (after optimization), indicating improved demographic alignment.

A grey background with text overlay

AI-generated content may be incorrect.Response time was tracked throughout the pipeline to ensure that added fairness mechanisms did not introduce significant latency. Average generation time per prompt increased from \_\_ seconds in the baseline to \_\_ seconds in the full pipeline, primarily due to the filtering and reranking steps.

Lastly, to ensure that semantic integrity of the prompt was preserved, we used CLIP embeddings to compare cosine similarity between the generated image and the original text prompt. The mean cosine similarity remained consistent (\_\_ baseline, \_\_ post-optimization), confirming that demographic balancing did not distort semantic alignment.

This comprehensive evaluation demonstrates that our extended pipeline effectively reduces bias while maintaining or improving overall output quality and relevance.

# **Future Work**

Building on this foundational work, several promising directions can be explored. First, automated prompt augmentation using demographic balancing rules or language models could replace manual prompt engineering, leading to more scalable and adaptive generation strategies. We also propose integrating reinforcement learning with human feedback (RLHF) or adversarial training to fine-tune Stable Diffusion models directly with fairness objectives in mind, beyond filtering alone.

A grey background with text overlay

AI-generated content may be incorrect.Additionally, enhancing demographic detection models like FairFace with confidence calibration or ensembling multiple facial analysis models could improve classification robustness, especially for non-photorealistic images. A potential extension of the evaluation pipeline is the use of user-in-the-loop frameworks, where real-time feedback guides the selection or filtering of generated outputs.

Long-term, developing a benchmark dataset and fairness evaluation suite for generative models—similar to ImageNet or COCO, but annotated for bias-related features—would provide the field with a much-needed standard for comparison. Finally, while our work focused on race and gender, expanding the pipeline to consider age, disability, attire, and cultural symbols would ensure a more holistic approach to fair image generation.

# References:

1. Jun, Y. (2024, April 8). A brief overview of gender bias in AI. The Gradient. https://thegradient.pub/gender-bias-in-ai/
2. AlDahoul, N., Rahwan, T., & Zaki, Y. (2025). AI-generated faces influence gender stereotypes and racial homogenization. Scientific Reports, 15, Article 14449. https://doi.org/10.1038/s41598-025-99623-3