

Abstract

Methods and Systems in Text-to-Image Diffusion Models for Fairness

The present invention provides methods and systems for reducing biases in text-to-image diffusion models, particularly biases related to gender, race, and their intersections in occupational prompts. The invention introduces a fairness framework based on distributional alignment, comprising two core technical solutions: (1) a distributional alignment loss that adjusts the output of the model toward user-defined target distributions, and (2) an adjusted direct finetuning (adjusted DFT) of the model's sampling process using an adjusted gradient to optimize losses based on generated images. These techniques reduce bias while supporting diverse perspectives on fairness, such as age-controlled debiasing across multiple concepts. The method's scalability allows for debiasing multiple prompts simultaneously, improving the inclusivity of diffusion model outputs across varied demographics.

Fig. 4