Open-set Bias Detection in Text-to-Image Generative Models

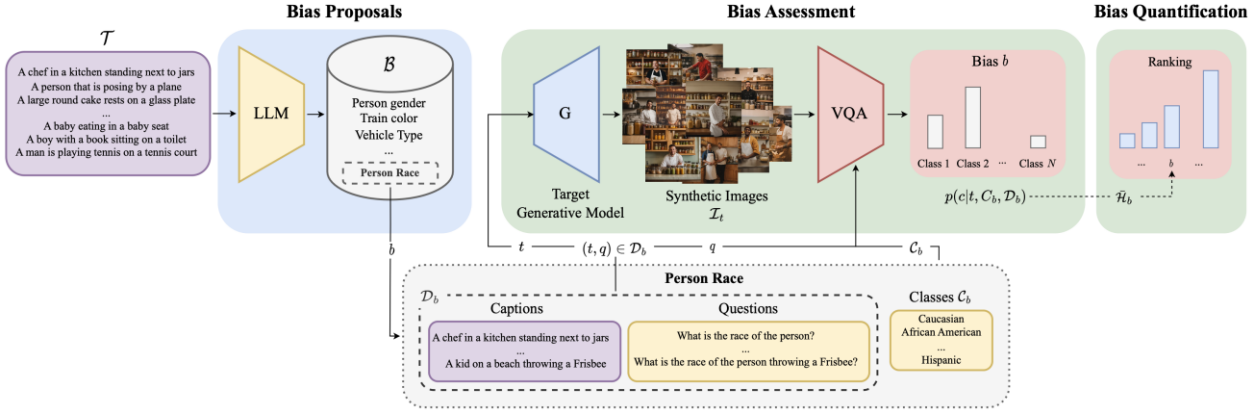
What has been done earlier?

What is the problem?

- Detection of biases in generative models that may not be present in the training data but emerge in the outputs.

Early work focused on identifying and quantifying bias in AI systems, especially in natural language processing (NLP) and image classification.

Existing works focus on detecting closed sets of biases defined a priori, limiting the studies to well-known concepts .



Siddhant Srivastav,B421050

Open-set Bias Detection in Text-to-Image Generative Models

What are the remaining challenges? What novel solution proposed by the authors to solve the problem?

Remaining Challenges :

- how context influences bias in generated images, ensuring that the model interprets nuanced meanings correctly.
- Necessity to deeply investigate their (Gen AI Models) safety and fairness to not disseminate and perpetuate any kind of biases.

Novel Solution Proposed :

- A new pipeline that identifies and quantifies the severity of biases agnostically, without access to any precompiled set.
- A Large Language Model (LLM) to propose biases given a set of captions.
- A Vision Question Answering model recognizes the presence and extent of the previously proposed biases

Siddhant Srivastav , B421050