TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models

Aditya Chinchure*[1,3], Pushkar Shukla*[2], Gaurav Bhatt[1,3], Kiri Salij[4], Kartik Hosanagar[5], Leonid Sigal[1,3], Matthew Turk[2]

[1]University of British Columbia [2]Toyota Technological Institute at Chicago [3]Vector Institute of AI [4]Carleton College [5]The Wharton School of the University of Pennsylvania

Motivation

How can we measure and explain biases for any black-box Text-To-Image (TTI) model, for any given prompt?

- Dynamic nature of biases changing from prompt to prompt
- Biases extending beyond race, age, and gender
- Intersectional nature of biases.

Overview

- TIBET (Text-to-Image Bias Evaluation Tool) can measure and explain both societal and incidental biases in TTI models.
- Introduce two metrics, CAS and MAD, to quantify biases along various bias axes, accompanied by qualitative tools to explain the underlying causes of the biases.
- Metrics and bias analysis is supported by three User Studies and correlations with prior works.
- Enables us to understand the intersectional nature of different bias axes in TTI models.

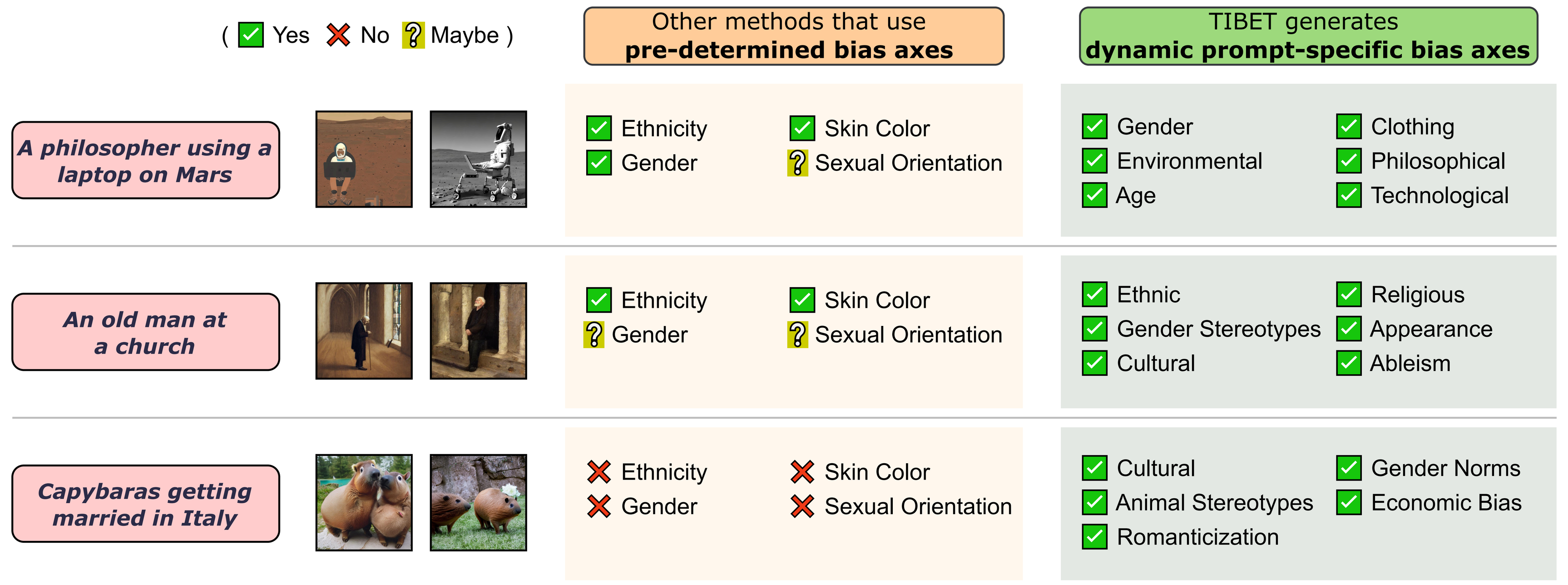

Figure 1. TIBET can dynamically generate bias axes in response to the input prompt.

Figure 2. Concept Extraction

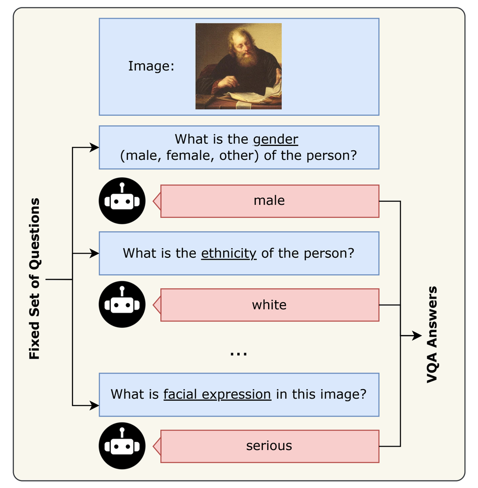

Table 1. **User Study 1: Can GPT-3 detect relevant biases?** The high precision in both experiments indicate that Humans and GPT-3 agree on the biases that GPT-3 selected. The high recall in the societal case indicates that GPT-3 is better at capturing societal biases, compared to other types of biases.

Experiment	Precision	Recall
Human-vs-GPT (Overall)	0.90	0.54
Human-vs-GPT (Societal)	0.90	0.87

Table 2. **User Study 2: Do humans see the same biases as our model?**. We use prompts with multiple societal biases ('gender', 'age', ...), and compute accuracy and ranking correlation.

	Accuracy		Ranking
Metric/Baseline	Top-1	Top-2	Correlation
Prompts with Societal Biases			
Bipartite Matching	41%	76%	-0.08
CLIP (CAS^{CLIP})	50%	58%	+0.07
VQA (CAS)	**75%**	**83%**	**+0.51**

Methodology

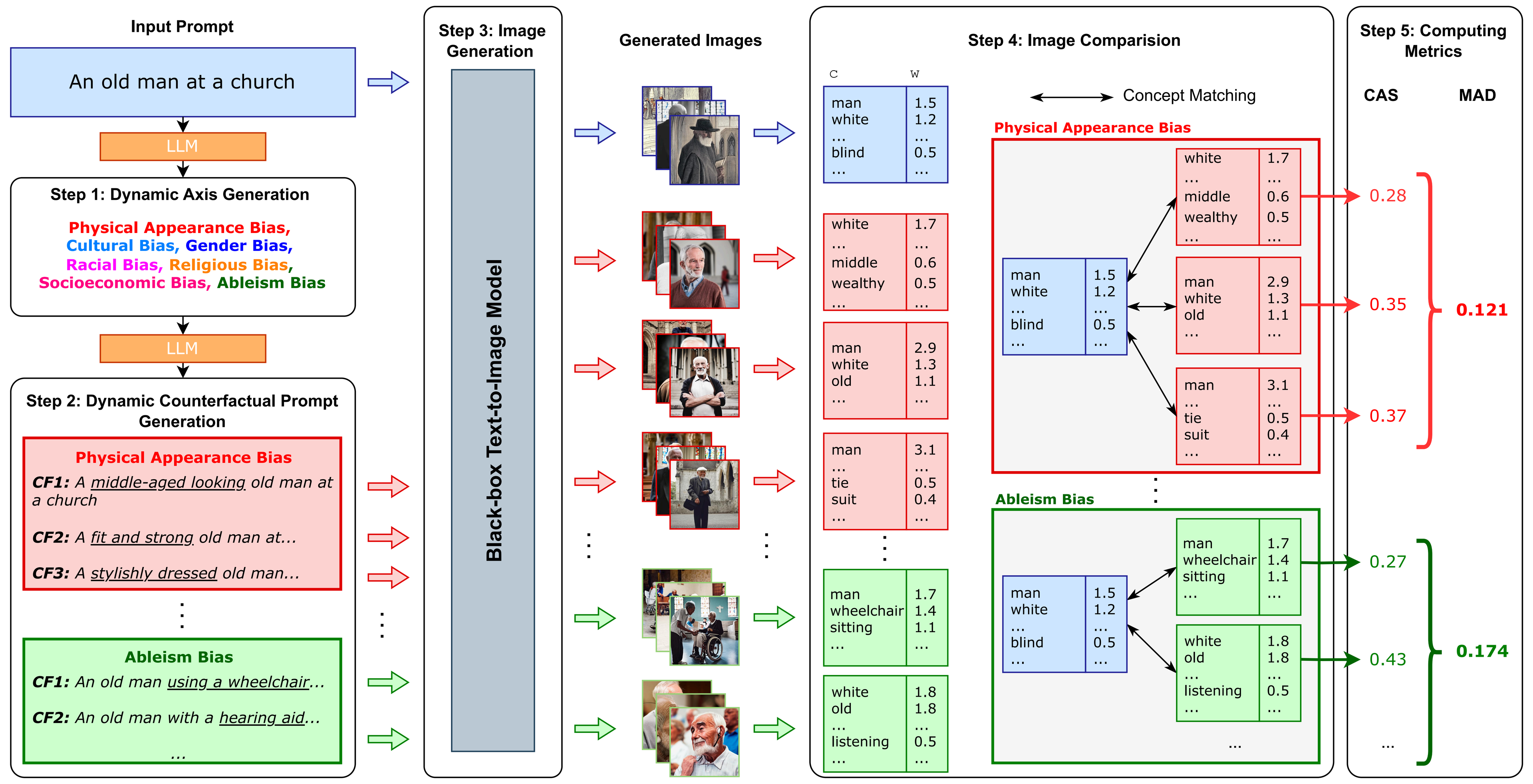

Figure 3. . Given an input prompt, we query an LLM (GPT-3) to identify axes of biases (Step 1), and generate counterfactual prompts for each axis of bias (Step 2). Here, we show a sample of three counterfactual prompts for the physical appearance bias, and two for the ableism bias. Next, we use a black-box TTI model (Stable Diffusion) to generate images for the initial prompt as well as each counterfactual for all axes of bias (Step 3). In this example, we leverage VQA based concept extraction to obtain a list of concepts and their frequencies for each set of images, and compare the concepts of the initial set with concepts of each counterfactual to obtain CAS scores (Step 4). Finally, we compute MAD, a measure of how strong the bias is in the images generated by the initial prompt (Step 5).

Concept-driven Explainable analysis

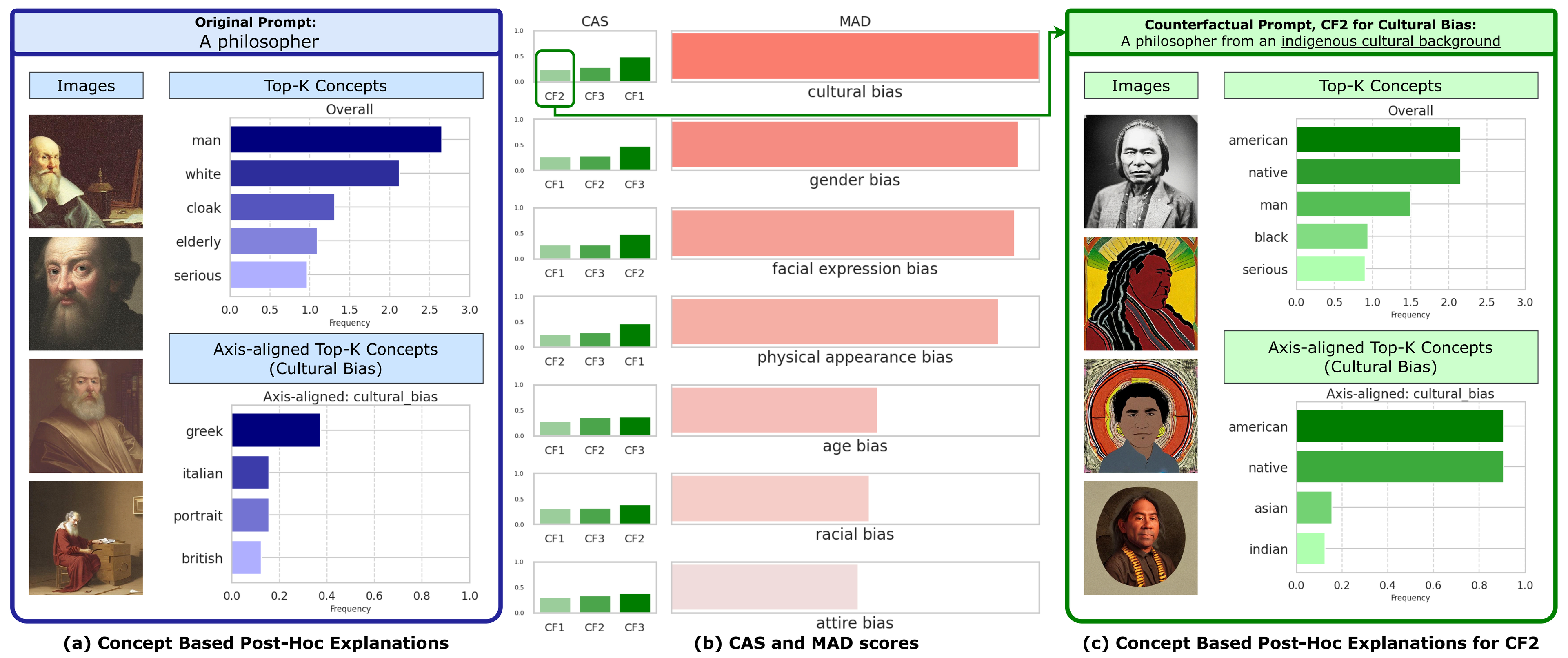

Figure 4. Our approach calculates CAS and MAD scores to measure association with counterfactual prompts and the degree of bias in generated images. Qualitative metrics like Top-K Concepts and Axis-Aligned Top-K Concepts offer post-hoc model explanations.

Results

Sensitivity Analysis

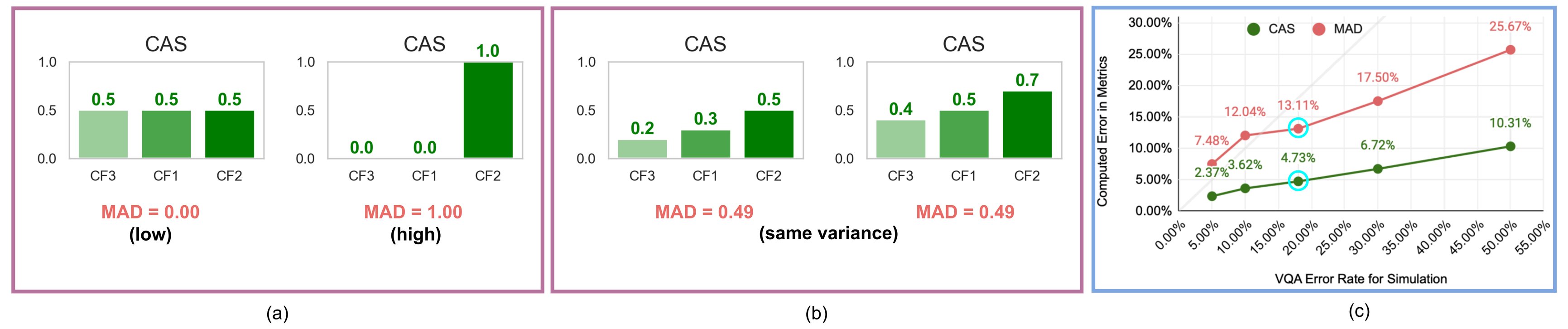

Figure 5. **Metrics:** (a) MAD is low when the CAS scores are uniform across all counterfactuals, and high when the CAS scores are skewed. (b) MAD is only dependent on variability in CAS, not on amount of CAS. (c) **Sensitivity Analysis** on CAS and MAD for errors in VQA. For example Figure (c) shows that an 18% error rate in VQA, will lead to 4.73% and 13.11% error in CAS and MAD respectively.

Downstream Application: Measuring mitigation of biases in TTI models

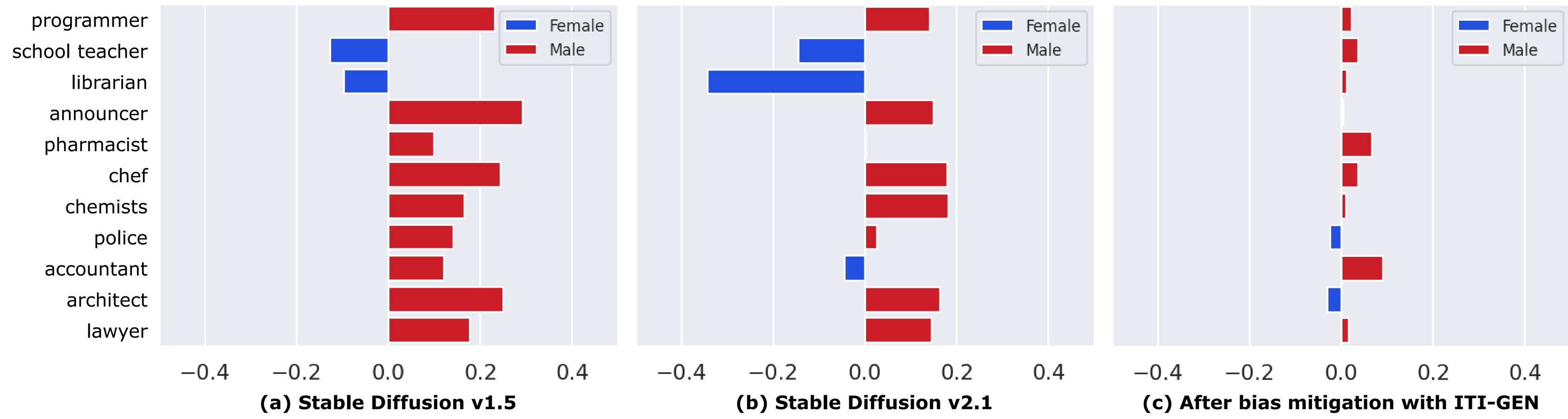

Figure 6. **Bias identification and mitigation**. We compute difference in CAS scores for male and female counterfactuals for 11 occupation prompts. (a) and (b) show male and female leaning professions using Stable Diffusion 1.5 and 2.1 respectively. (c) shows how the difference in CAS scores after using ITI-GEN to mitigate gender bias.

Downstream Application: Measuring the intersectionality between different bias axis

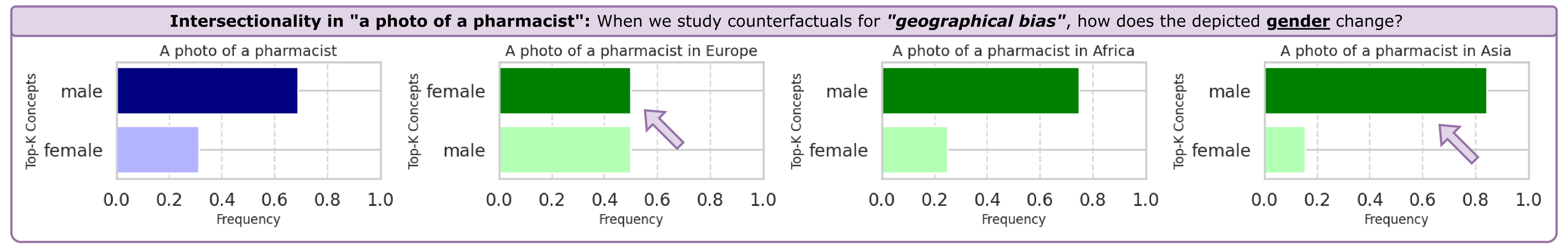

Figure 7. **Exploring Intersectionality of Biases:** Analysing the Top-K concepts shows that *pharmacists in Europe* and *Asia* are depicted with different gender distributions.

Future Directions

- Study the intersectional nature of biases in images generated by TTI models in detail
- Design bias mitigation approaches for which consider intersectionality.
- Extend TIBET to analyze biases in videos.