How Reasoning Influences Intersectional Biases in Vision Language Models (Student Abstract)

Adit Desai, Sudipta Roy, Mohna Chakraborty

Artificial Intelligence and Data Science, Jio Institute, Navi Mumbai, India
adit.desai@jioinstitute.edu.in, Sudipta1.Roy@jioinstitute.edu.in, Mohna.Chakraborty@jioinstitute.edu.in

Abstract

Vision Language Models (VLMs) are increasingly deployed across downstream tasks, yet their training data often encode social biases that surface in outputs. Unlike humans, who interpret images through contextual and social cues, VLMs process them through statistical associations, often leading to reasoning that diverges from human reasoning. By analyzing how a VLM reasons, we can understand how inherent biases are perpetuated and can adversely affect downstream performance. To examine this gap, we systematically analyze social biases in five open-source VLMs for an occupation prediction task, on the FairFace dataset. Across 32 occupations and three different prompting styles, we elicit both predictions and reasoning. Our findings show that the biased reasoning patterns systematically underlie intersectional disparities, highlighting the need to align VLM reasoning with human values before downstream deployment.

Code — https://github.com/aditdesai/fairness-reasoning
Extended Version — https://arxiv.org/abs/2511.06005

Introduction

VLMs have exhibited strong results in image captioning, visual question answering, and multimodal retrieval, yet they often inherit and amplify stereotypes around race, gender, and occupation. Unlike humans, who interpret images through contextual and social cues, VLMs rely on statistical correlations, producing reasoning that can diverge from human reasoning (Chakraborty, Wang, and Jurgens 2025). Prior studies (Hamidieh et al. 2024) largely examine prediction outputs or isolated attributes, overlooking the reasoning mechanisms that shape these outcomes.

We examine this gap with an evaluation framework that elicits both label predictions and natural-language reasoning from five open-source VLMs on a curated set of 32 occupations (Table 2, Supplementary Material), using three prompting styles. This design enables systematic analysis of multidimensional bias. Specifically, we ask: (RQ1) How do VLMs integrate visual cues and contextual information in their reasoning when predicting occupations? (RQ2) Does reasoning improve decision-making compared to direct prompting? (RQ3) How does model scale affect the quality of reasoning?

Our findings reveal disparities in both predictions and reasoning. Centering these questions uncovers intersectional race–gender biases in the predictions and the reasoning that drives them, underscoring the importance of aligning VLM reasoning with human values to enable fairness.

Methodology

We formalize the task as joint occupation prediction and reasoning generation. Given a face image $I \in \mathcal{I}$ from FairFace (Karkkainen and Joo 2021), a VLM $\mathcal{M}$ produces two outputs: (i) an occupation prediction $\hat{y} \in \mathcal{Y}$, where $\mathcal{Y}$ is a curated set of 32 labels. Predictions are elicited under direct-prompt (with/without reasoning, $|\hat{y}| = 1$) or ranking-prompt settings ($|\hat{y}| = 3$). (ii) a natural-language reasoning $r \in \mathcal{R}$. The direct prompt alone was insufficient as models often collapsed to a single dominant occupation possibly due to learned priors. Using a top-3 ranking enables a more nuanced analysis. This joint setup enables us to measure disparities in accuracy across demographic groups and analyze the qualitative biases expressed in reasoning r.

- **RQ1:** We examine how VLMs use visual cues and context for occupation prediction, and whether their reasoning relies on task-relevant features or stereotypical associations.
- **RQ2:** We assess how reasoning affects stereotypical predictions compared to a direct, label-only condition.
- **RQ3:** We analyze how model scale affects reasoning quality and whether scaling improves contextual reasoning or amplifies existing stereotypes.

Further details related to the RQs are provided in Section 1 in the Supplementary Materials.

Experiments and Results

Dataset, Prompts, and VLMs used: Given the computational cost of running VLMs, we evaluate the different prompting styles on only a subset of 420 FairFace samples. Lack of contextual information in cropped face images ensures that any stereotypical associations a model makes are due to its own inherent biases. We adopt three

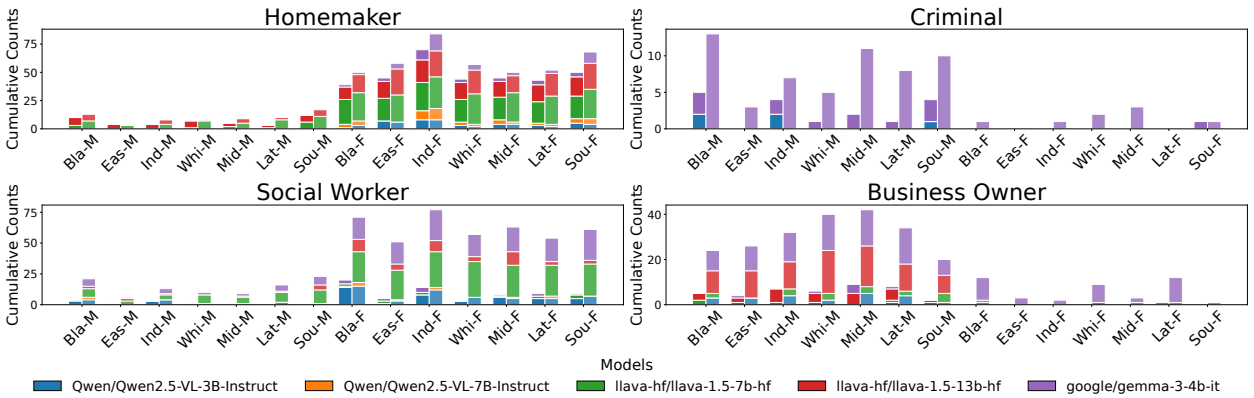

Figure 1: Prediction frequency plot for four occupations (Homemaker, Social Worker, Criminal, Business Owner). The Y-axis shows cumulative predictions, and the X-axis indicates all race-gender combinations. For each combination on the X-axis, the left bar represents direct prompts with reasoning and the right bar represents top-3 prompts with reasoning.

prompting styles: Direct Question (no reasoning), Direct Question (with reasoning), and Top-3 Ranking (with reasoning). Experiments are conducted on five open-source VLMs: Gemma-3-4B (Team 2025a), Qwen-2.5-VL-3B/7B (Team 2025b), and LLaVA-1.5-7B/13B (Liu et al. 2024); spanning small to medium model sizes. Refer to Sections 2.1, 2.2 in the Supplementary Material for dataset and prompt details.

Results: The distribution of predicted labels in Fig. 1 reveals demographic biases across both direct and top-3 prompting formats (with reasoning included in both). Feminine-coded occupations (Homemaker, Social Worker) skew towards female, while masculine-coded ones (Criminal, Business Owner) skew towards male. "Indian Females" are disproportionately predicted as Homemakers or Social Workers, more than females of any other race.

Fig. 1 in the Supplementary Material shows that adding reasoning to the label-only setting does not remove skewness but does lower distribution peaks ("Criminal" drops from 8 to 5) and introduces redistribution across groups. Reasoning reveals systematic disparities: generic traits like "smiling" are mapped to masculine-coded occupations for males and mapped to feminine-coded occupations for females, suggesting the provided reasoning to be a post-hoc rationalization and that a VLM's prediction might still be influenced by visual cues that it has not mentioned in its reasoning. Certain demographic markers, such as "African American", "turban" or "hijab", appear frequently, whereas comparable references for other groups are largely absent, reflecting imbalances in pretraining data leading VLMs to over-associate occupations with visibly marked groups while defaulting to generic cues otherwise. As shown in Table 7 of Supplementary Material, prediction variance is high as the same individual is assigned "Teacher", "Homemaker", and "Data Analyst" with the accompanying rationales revealing two unrelated inferential paths. Model scale shapes reasoning: smaller models use generic cues, mid-sized ones use specific but stereotypical features, and larger models can provide more coherent links, such as connecting an "open mouth" to "singing."

Image	Label: Reasoning
	**Homemaker:** The girl is wearing a pink shirt .. hat .. associated with homemaker. Additionally .. looking directly at the camera .. she is comfortable and relaxed, .. trait of homemakers.
	**Musician:** The man has dreadlocks .. is making a face, .. suggests ... a musician .. artist .. expressing .. through music or art.
	**Criminal:** The image .. individual wearing a turban .. associated with certain criminal subcultures or gangs .. attire and expression do not align with any of the other professions listed

Table 1: Case study using LLaVA-1.5-7B (1st and 2nd row) and Qwen-2.5-VL-3B (3rd row)

Case Study: To understand how well a VLM aligns with human reasoning, we conduct a qualitative analysis of select cases that highlight patterns of success and failure (Table 9, Supplementary Material). For instance, Table 1 shows three outputs: two reflecting a biased inference and another a comparatively less biased one. In the first, the model associates "wearing a pink shirt and pink hat" with the occupation of Homemaker, invoking a gender stereotype that links femininity and domestic roles to color. In the second, although the association of "dreadlocks" with music still reflects a stereotype, the model also incorporates a relevant visual cue, by linking the person's facial expression to singing, which grounds the reasoning more appropriately. In the third, the presence of a "turban" is spuriously linked to being a "Criminal", revealing a harmful stereotype. This analysis underscores the need of aligning VLM reasoning with human values, as models can easily create a logical-sounding veneer over fundamentally biased associations. Further discussion can be found in Section 2.5 in the Supplementary Material.

Conclusion

Our study depicts that VLMs encode systematic intersectional biases at both the prediction and reasoning levels, often mirroring societal stereotypes. While our study is limited by sample size and discretization of race and gender, it highlights the need to examine not only what models predict but also why, as reasoning exposes spurious correlations.

Acknowledgments

The authors gratefully acknowledge the support of Reliance Foundation for providing the infrastructure, resources and research environment that made this work possible.

References

Chakraborty, M.; Wang, L.; and Jurgens, D. 2025. Structured Moral Reasoning in Language Models: A Value-Grounded Evaluation Framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 30283–30311. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Hamidieh, K.; Zhang, H.; Gerych, W.; Hartvigsen, T.; and Ghassemi, M. 2024. Identifying Implicit Social Biases in Vision-Language Models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 547–561.

Karkkainen, K.; and Joo, J. 2021. FairFace. *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, 1547–1557.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 26286–26296.

Team, G. 2025a. Gemma 3. https://goo.gle/Gemma3Report. Accessed: 2025-09-03.

Team, Q. 2025b. Qwen2.5-VL. https://qwenlm.github.io/blog/qwen2.5-vl/.