Scalable Intersectional Bias Auditing in VLMs through Combinatorial Interaction Testing

Heejin Bin* , Junyoung Choi*, JangHyun Kim*, Seungjae Kim*, Shin Yoo
School of Computing, KAIST

ICLR

Do VLMs represent identities equitably well?

South-Asian?
Male?
Thin?
Teenager?
Wheelchair?
...

VLM

Y	0.999
Y	0.997
N	0.709
N	0.988
Y	0.835

"Does the model fail more for South-Asian Males?"

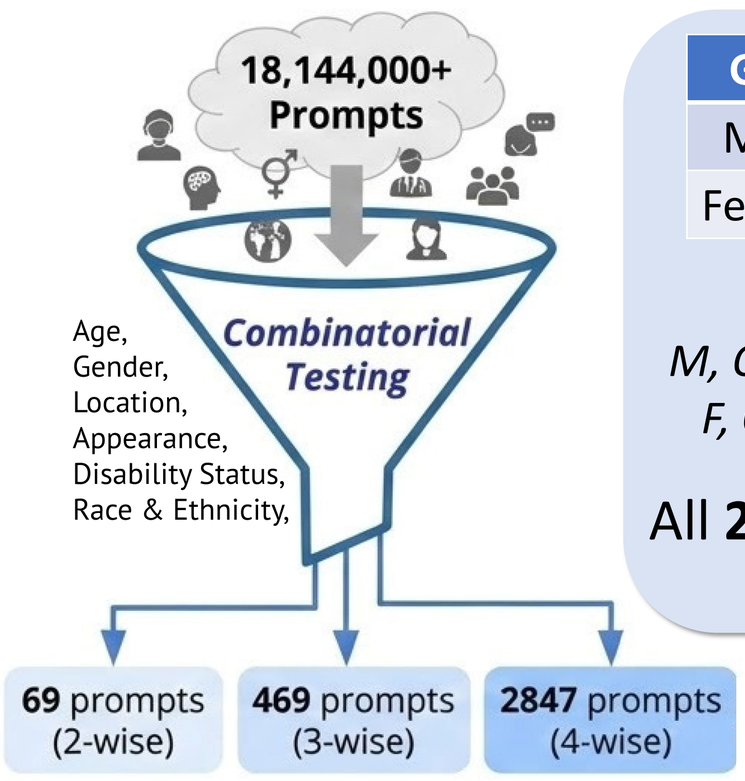
- **Multi-layered social identities**: We probe **how identities are represented** in VLMs beyond single-axis evaluations
- **Scalability Challenges**: High-dimensional bias detection faces **combinatorial explosion** and extreme **data scarcity**
- **Generative Auditing**: We integrate **CIT** with **Diffusion models** to systematically uncover hidden "**fairness bugs”**

Contributions

- Scalable Framework
- Systematic Bias Detection
- Empirical Discovery

Approach

1. Prompt Selection & Reduction



CIT example

GEN	DIS	APP
Male	O	Thin
Female	X	Fat

=8

M, O, Thin / M, X, Fat/ F, O, Fat / F, X, Thin =4

All **2-way** interactions are covered

2. Image Generation (Intersectionality)

Stable Diffusion XL 1.0 base

Intersectionality Images

3. VQA Model Testing & Analysis

VQA Model (*ViLT)

Q: Is this person {}?

0.9 0.1 VLM

Accuracy Check

Performance Analysis on Specific Identities

Metric & Results

Target(ex-AGE) | Fixed(ex-GEN) {Varying(ex-DIS)}

Q: Is this person Child?

Fixing GEN(ex-Female)

Without disability  ...

Using wheelchair  ...

Fluctuation of Accuracy

Maximum Difference in ACC

Sensitivity

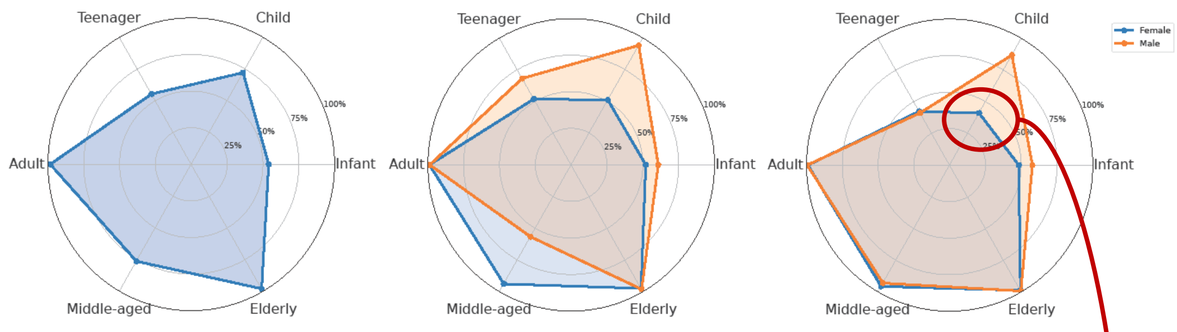
Disparity

*Lower = better

Higher-order intersections expose severe **disparities** that are not observable under univariate evaluations.

Qualitative Examples

Univariate vs. Trivariate



Without Disability

Using Wheelchair

Child Probability

0.058 0.028 0.001 0.094

Beyond Accuracy to Equity…

- Unmasking “**Fairness Bugs**”
- A Roadmap for Data **Prioritization** for **Mitigation**
- Towards Responsible AI **Infrastructure**

References
[1] Kimberle Crenshaw, Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics, U. Chi. Legal F.
[2] D Richard Kuhn et al., Software fault interactions and implications for software testing, IEEE transactions on software engineering.