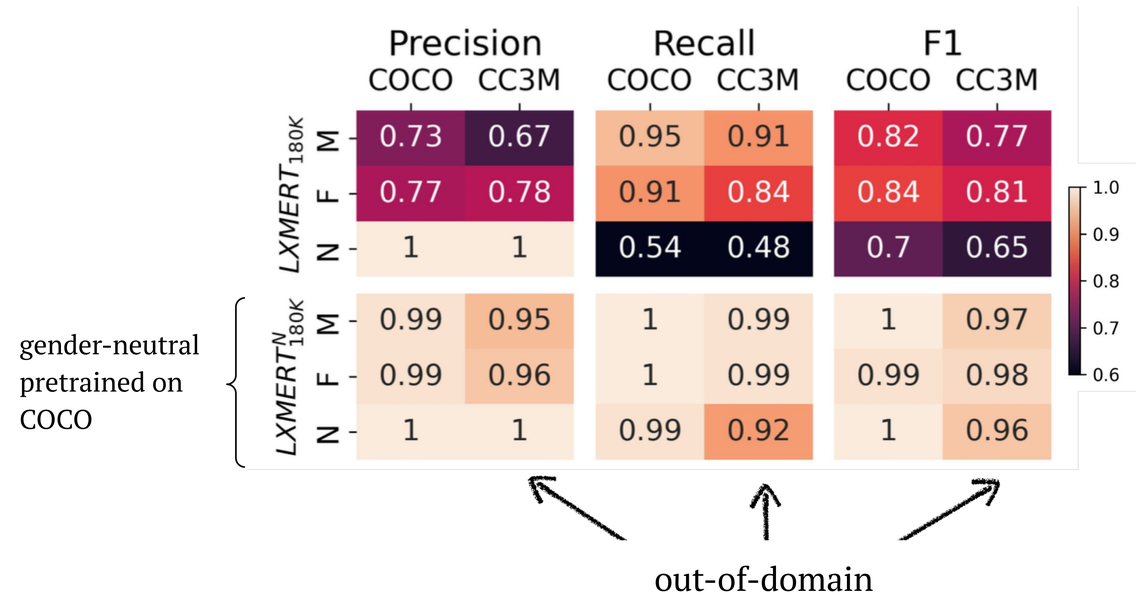
Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models

Laura Cabello Emanuele Bugliarello Stephanie Brandl Desmond Elliott
University of Copenhagen
lcp@di.ku.dk

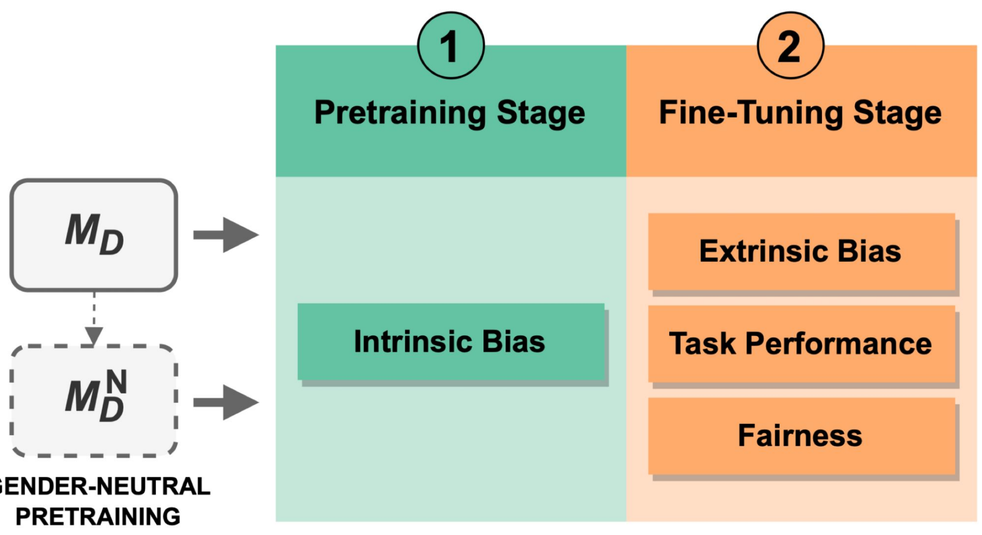
Contribution

- We present a comprehensive analysis of **gender bias amplification and fairness** (group disparity) of encoder-only and encoder-decoder V&L models
- We present a simple way to promote fairness in VLMs: extra **pretraining steps on unbiased (gender-neutral) data**
  - reduces fine-tuning variance and group disparity on VQAv2 and retrieval tasks
  - does *not* compromise task performance

Intrinsic Bias



Overview



Intrinsic, Extrinsic Bias Amplification

$$\text{BiasAmp}_{A\to T} = \frac{1}{|A||T|}\sum_{\substack{a\in A\\ t\in T}} y_{at}\Delta_{at} - (1-y_{at})\Delta_{at}$$

Wang and Russakovsky (2021)

Fairness

Groups: 👩 👨 🧑

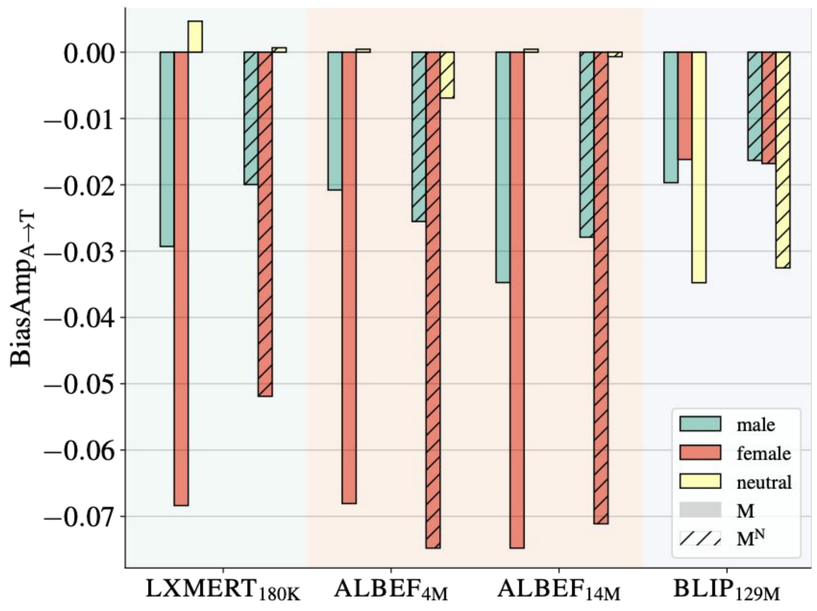
Measure: group **performance disparity** (performance gap)

Gender-neutral data

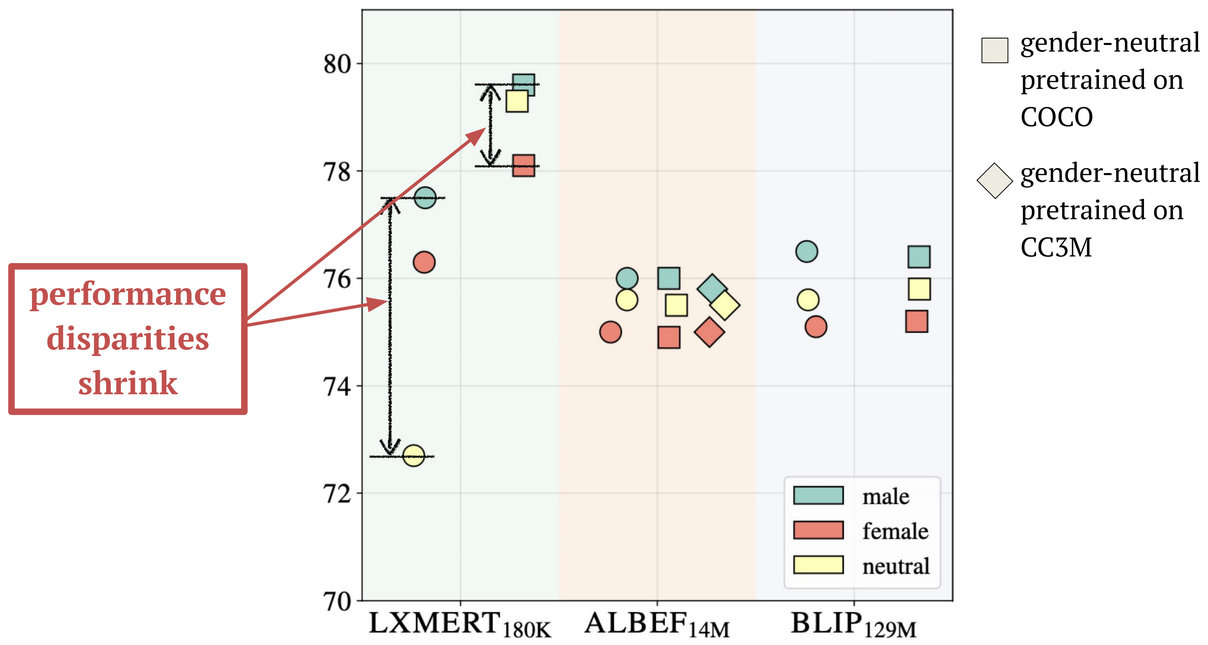
We swap gendered terms for gender-neutral terms on pretraining data from COCO and Conceptual Captions (CC3M)
e.g., 🤰 mother → parent□, 👧 girl → child 🧒

Extrinsic Bias (VQA)

Does the model predict more often the word X when "man" appears in the question?



Task Performance (VQA)



Conclusions

- **Intrinsic bias** can reinforce harmful biases, but these **may not impact the treatment of groups** (or individuals) on downstream tasks
- Bias in a model and its empirical fairness (group disparities) are in fact **independent matters**
- Continued pretraining on gender-neutral data **reduces group disparities** on VQAv2 and retrieval tasks **without performance penalty**