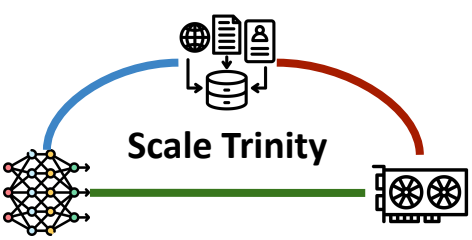
PRESENTER
Sepehr Dehdashtian

The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models

Why this study?



As gigantic multimodal models trained on datasets of unprecedented scale become commonplace, the need to audit their impact on society becomes stronger than ever.

How does dataset scaling affect racial and gender biases in multimodal models?

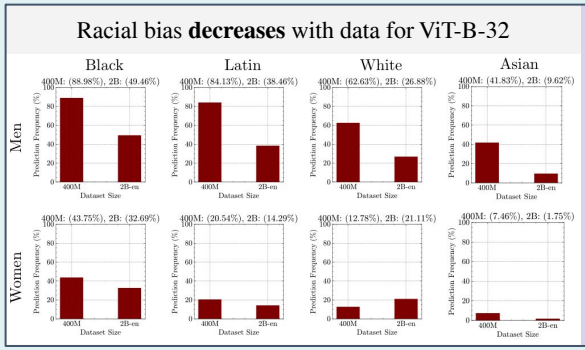
Experimental Design

We audit **fourteen** CLIP models trained on **LAION-400M and LAION-2B** from the OpenCLIP repository.

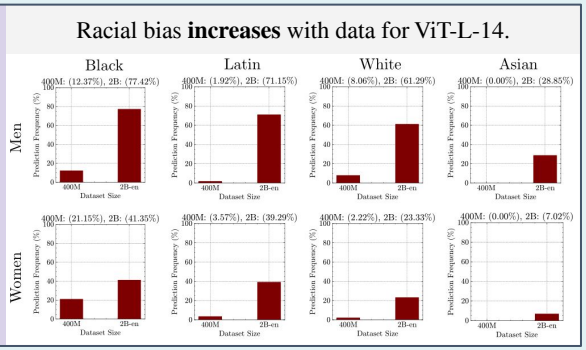
We use **Chicago Face Database**, a neutral expression face dataset of 597 images with self-reported race and gender, to probe the racial and gender biases in these models.

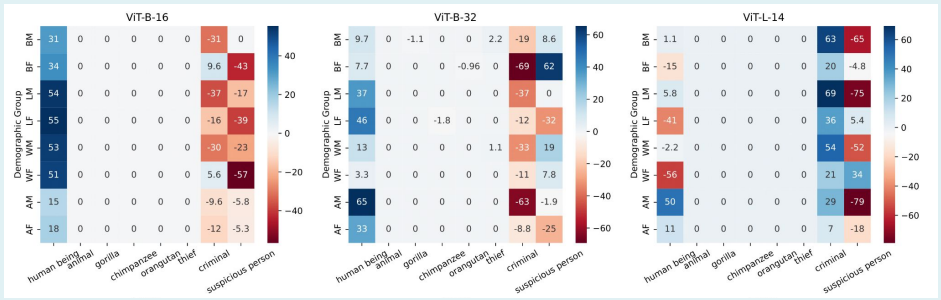
For each image, we calculate the cosine similarity score with the following **eight labels**.

["human being", "animal", "gorilla", "chimpanzee", "orangutan", "thief", "criminal", "suspicious person"]



Data scaling has opposite effects on models with different capacities.





The benefits of data scaling are not **equitable** to all ethnic groups and all models.

Mitigation Strategies for **Equitable** Deep Learning

- Avoid ad-hoc decision-making for dataset curation hyperparameters.
- Beware of CFD physiognomy.
- Avoid data subsampling during ethics checks.
- Open access for independent audits.

Abeba Birhane*, Sepehr Dehdashtian*, Vinay Uday Prabhu, Vishnu Naresh Boddeti