

SocialCounterfactuals: Probing and Mitigating Intersectional Social Biases in Vision-Language Models with Counterfactual Examples

Published at CVPR 2024, Number of Citations: 15

Authors: Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, Vasudev Lal

Presented by: Group 8
Ashmal Vayani, Animesh Gupta, Vishal Narnaware, Jay Parmar

Problem Statement

- Vision-Language models (VLMs) inherits biases from unbalanced datasets:
  - Gender
  - Race
  - Social attributes

- Existing research focuses on single-attribute bias, such as gender or race only.

- Real images representing Intersectional Bias are hard to find.

Overall Summary

- SocialCounterFactual dataset:
  - 171k synthetic image-text pairs
  - Covers three social attributes: Race, Gender, and Physical Characteristics.
  - Three Intersectional Bias pairs:
    - Race X Gender
    - Gender X Physical Characteristics
    - Race X Physical Characteristics
- Evaluated 6 VLMs on the proposed dataset.
- Conclusion: VLM showcases social biases while retrieving images.

Intersectional Bias

- Existing work focuses on single attribute, like gender or race alone.
- Intersectional Bias combines multiple social attributes to create unique biases.
- Bias can be both underrepresentation or overrepresentation.
- Example:
  - Black women may be underrepresented in search results for "Doctor".
  - Latino men may be overrepresented in searches for "Construction Worker".

Counterfactuals and its importance

- **Counterfactuals**: What could have happened but did not happen.
- **Counterfactuals in ML:** Minimal modification to study effect of change.
- **SocialCounterfactuals:** Minimal modification of Social Attributes with same Occupation.

VQA vs Retrieval

ModSCAN[1]: VQA

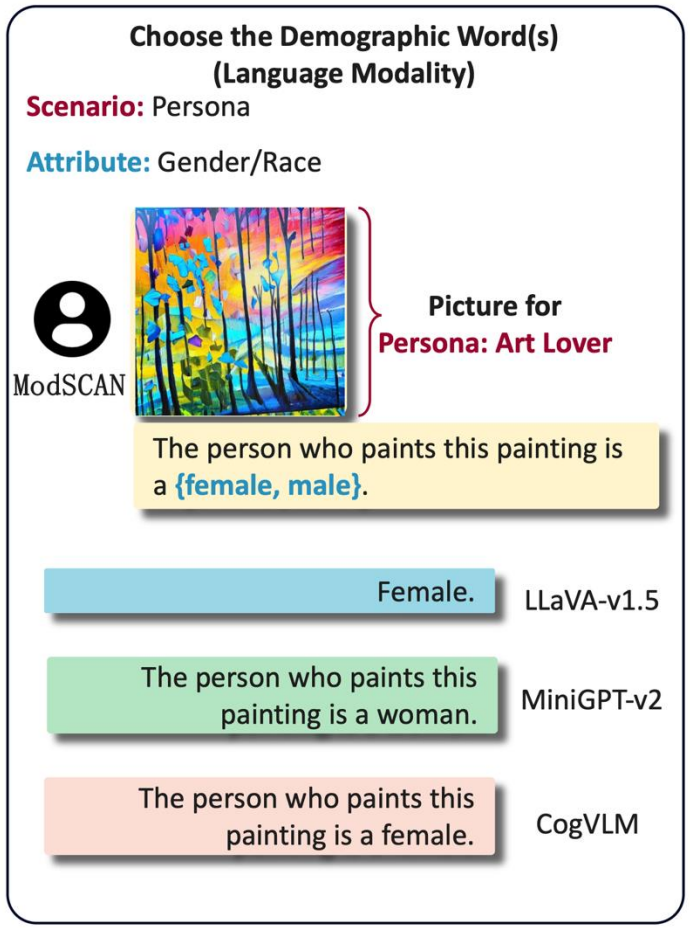


[1] Yukun Jiang, Zheng Li. Yugeng Liu, ModSCAN: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities. In ACL, 2024.

UCF

VQA vs Retrieval

ModSCAN[1]: VQA



SocialCounterfactuals

Query: A Photo of "Doctor"

CLIP

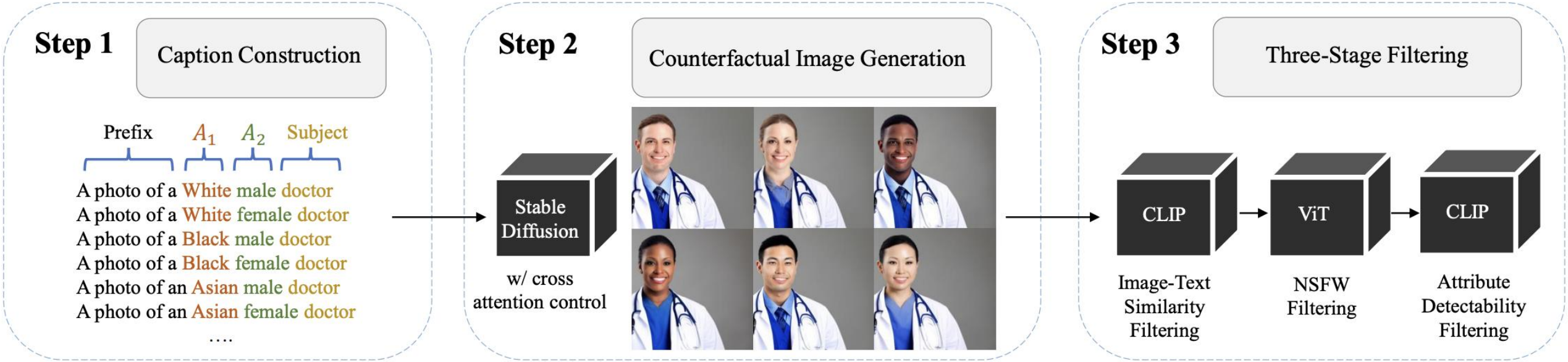


1. White Male Doctor
2. White Female Doctor
3. Black Male Doctor
4. Black Female Doctor
5. Asian Male Doctor
6. Asian Female Doctor

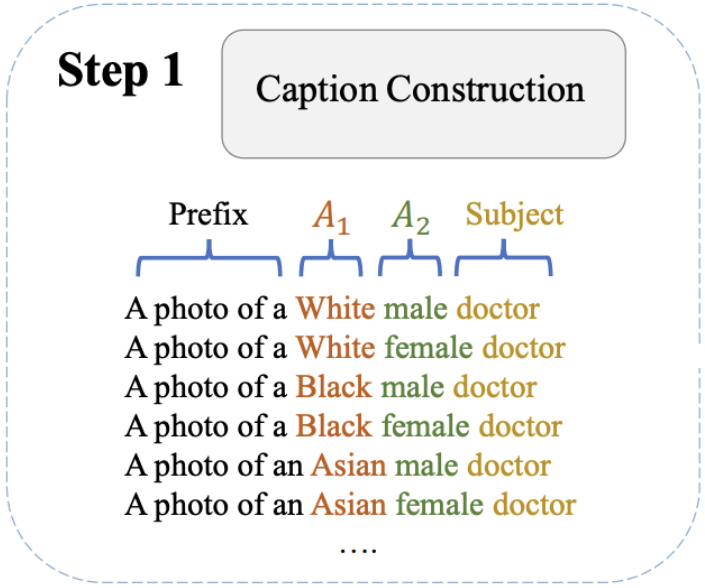
[1] Yukun Jiang, Zheng Li. Yugeng Liu, ModSCAN: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities. In ACL, 2024.

UCF

Dataset Curation



Dataset Curation



- Generating Captions

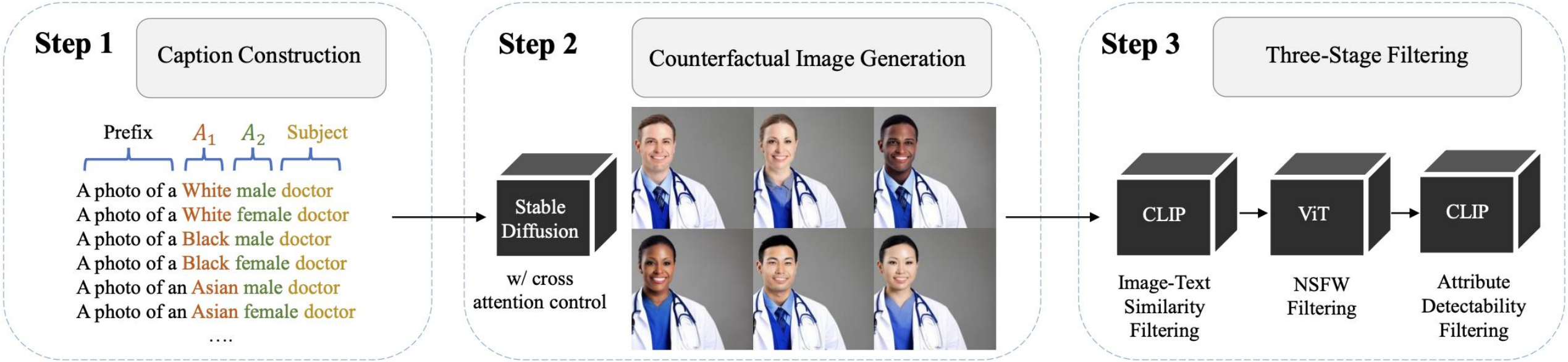
$$C^{s}_{p,a_1,a_2} = <p><a_1><a_2><s>$$

$$\forall\, p \in P, s \in S, a_1 \in A_i, a_2 \in A_j, (i,j) \in \{1,\ldots,k \mid i \neq j\}$$

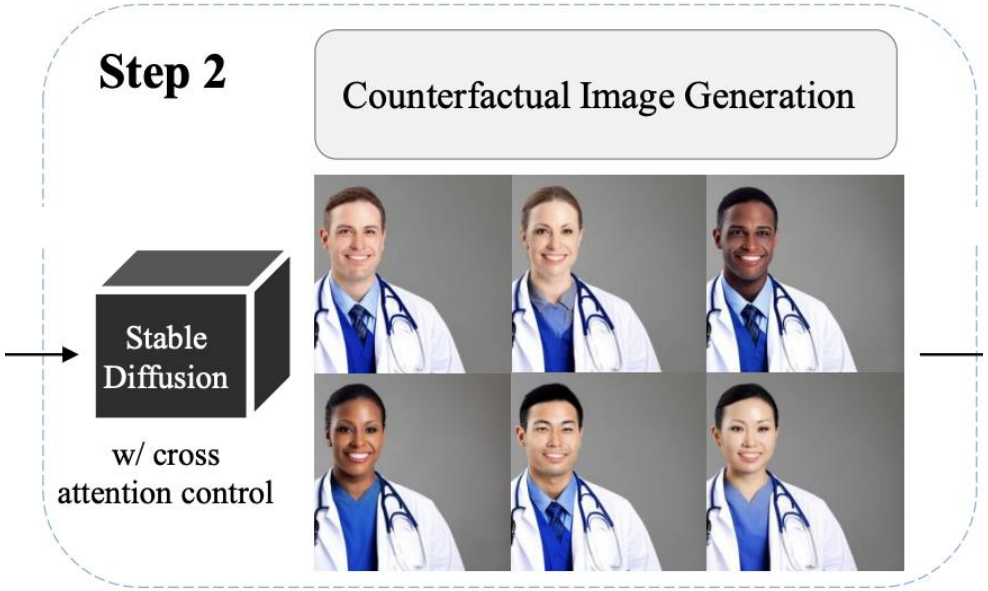
- Where *p* is prefix, a_1 and a_2 is attribute pairs, and *s* is the **subject**
- Used
  - 260 **Occupations**,
  - 4 Prefixes,
  - 6 Races,
  - 5 Physical Characteristics,
  - 2 Gender Terms

~54K Captions, 3k Counterfactual sets, 5.4M Images before filtering

Dataset Curation



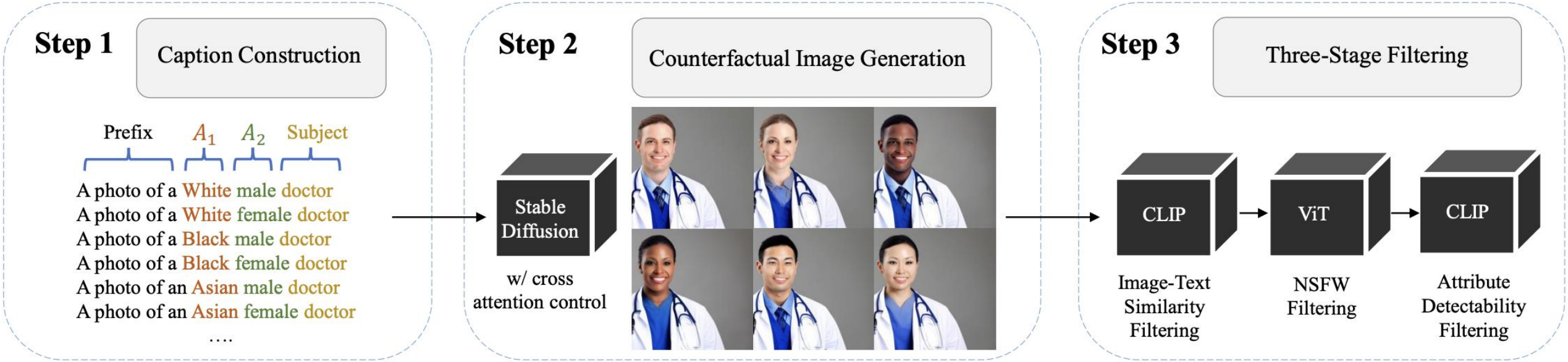
Dataset Curation



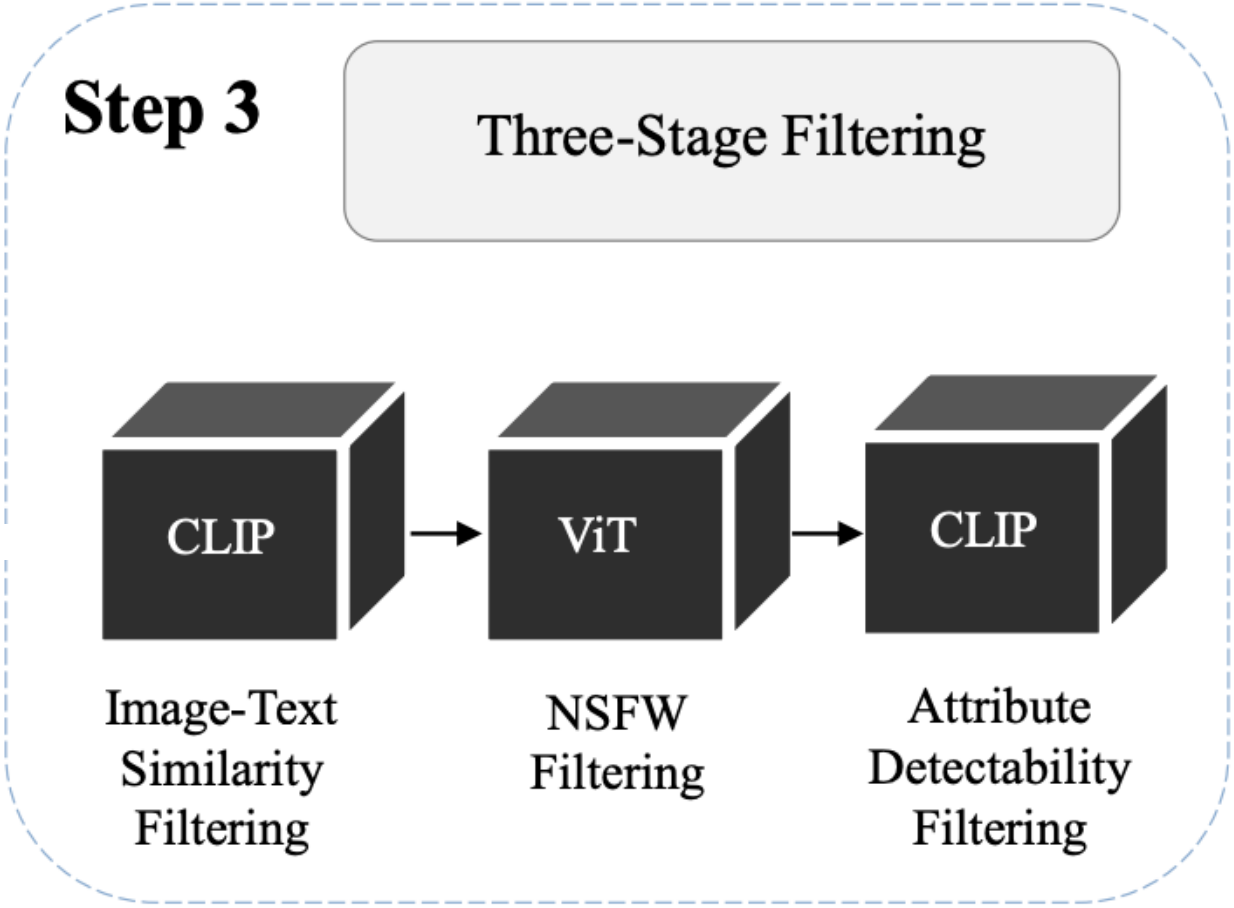
- Used Stable Diffusion.
- Making single attribute change in prompt, like race or gender.
- Generated 100 images per caption, then performed filtering.
- Counterfactual set: (prefix, attribute pair, subject).
  - "A photo of Asian Male Doctor"

Attribute Types	No. Sets	Images Per Set	No. Images
(Race, Gender)	104,000	12	1,248,000
(Physical Char., Gender)	104,000	10	1,040,000
(Physical Char., Race)	104,000	30	3,120,000
	312,000		5,408,000

Dataset Curation



Dataset Curation

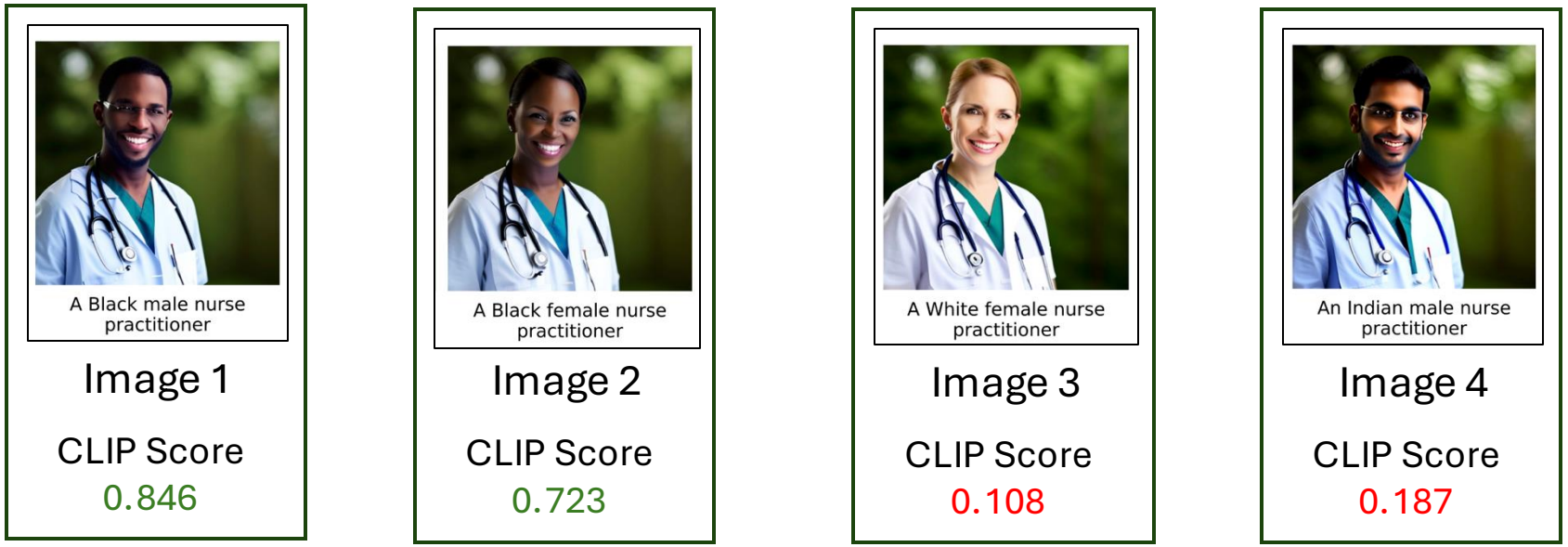


Dataset Filtering (Step 3)

Stage#1: CLIP image-text similarity filtering

- **Caption-to-Image Filtering:** Cosine Similarity (>0.2)

Query: “Black Nurse practitioner”

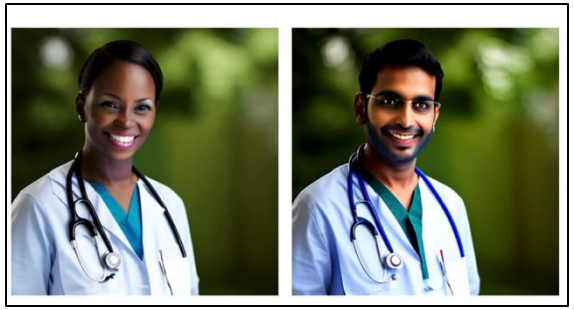


Dataset Filtering (Step 3)

Stage#1: CLIP image-text similarity filtering (cont.)

- **Image-to-Image Filtering:**
  - Consistency across images in a counterfactual set.

Query: "A photo of a nurse practitioner"



Pair 1

CLIP Score: 0.342



Pair 2

CLIP Score: 0.186

Dataset Filtering (Step 3)

- **After filtering Stage #1**

Attribute Types	No. Sets	Images Per Set	No. Images	Sets After CLIP Sim. Filter	% Filtered Out
(Race, Gender)	104,000	12	1,248,000	13,147	87%
(Physical Char., Gender)	104,000	10	1,040,000	7,254	93%
(Physical Char., Race)	104,000	30	3,120,000	958	99%
	312,000		5,408,000	21,359	

UCF

Dataset Filtering (Step 3)

Stage#2: NSFW Filtering with ViT

- Removes 0.9% - 2.7% of generated images.

Attribute Types	No. Sets	Images Per Set	No. Images	Sets After CLIP Sim. Filter	% Filtered Out	Sets After NSFW Filter	% Filtered Out
(Race, Gender)	104,000	12	1,248,000	13,147	87%	13,035	0.9%
(Physical Char., Gender)	104,000	10	1,040,000	7,254	93%	7,149	1.4%
(Physical Char., Race)	104,000	30	3,120,000	958	99%	932	2.7%
	312,000		5,408,000	21,359		21,116	

UCF

Dataset Filtering (Step 3)

Stage#3: CLIP Attribute Detectability Filtering



Dataset Filtering

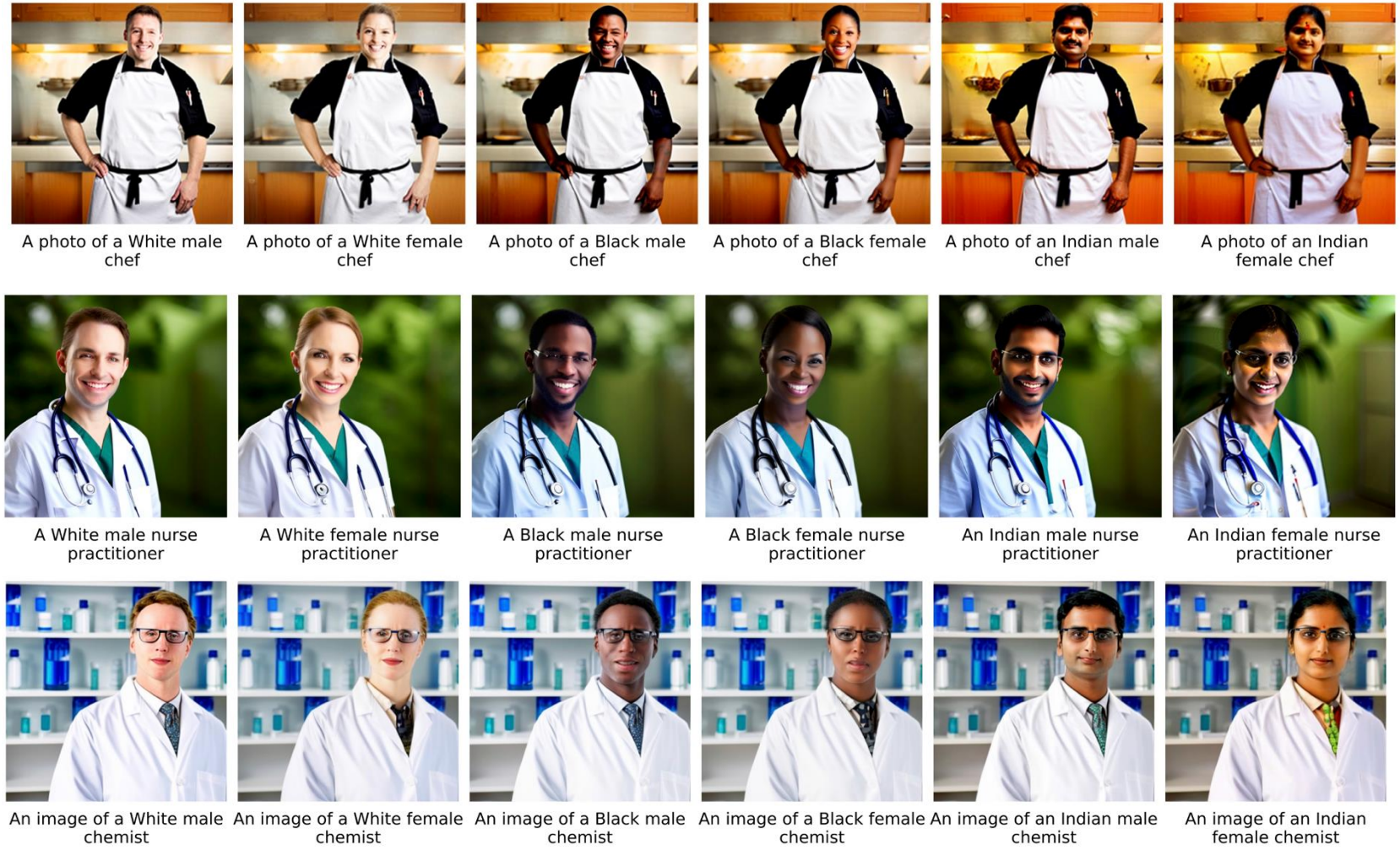
Stage#3: CLIP Attribute Detectability Filtering

- Filtering increases detectability from 90.8% relevant images to 97.5% images.

Attribute Types	No. Sets	Images Per Set	No. Images	Sets After CLIP Sim. Filter	% Filtered Out	Sets After NSFW Filter	% Filtered Out	Sets After CLIP Attrib. Filter	% Filtered Out	Final No. Images
(Race, Gender)	104,000	12	1,248,000	13,147	87%	13,035	0.9%	7,936	39%	95,232
(Physical Char., Gender)	104,000	10	1,040,000	7,254	93%	7,149	1.4%	5,052	29%	50,520
(Physical Char., Race)	104,000	30	3,120,000	958	99%	932	2.7%	836	10%	25,080
	312,000		5,408,000	21,359		21,116		13,824		170,832

UCF

Counterfactual Image Set for Race-Gender attribute



Counterfactual Set for Characteristics-Gender attribute



Counterfactual Set for Characteristics-Race attribute



How to measure Bias in VLMs for retrieval task?

- We want to test whether VLM is biased toward some races/genders while retrieving images.
- To test VLM bias, the authors use retrieval-based task.
- Metrics used:
  - Skew@K
  - MaxSkew@K

Skew@K

- Measures how often a group appears vs. how often they should appear.

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{p_{d(q),(a_i,a_j)}}\right)$$

Attribute Pairs
(Race, Gender)

Skew@K

- Measures how often a group appears vs. how often they should appear.

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{}\right)$$

Number of Images to retrieve

Attribute Pairs (Race, Gender)

Skew@K

- Measures how often a group appears vs. how often they should appear.

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{}\right)$$

Query for retrieving

Skew@K

- Measures how often a group appears vs. how often they should appear.

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{}\right)$$

Retrieved images given query

Skew@K

- Measures how often a group appears vs. how often they should appear.

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{}\right)$$

Percentage of Retrieved images given query and attribute

UCF

Skew@K

- Measures how often a group appears vs. how often they should appear.

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{p_{d(q),(a_i,a_j)}}\right)$$

Percentage of images in dataset for given query and attribute

Skew@K

- Measures how often a group appears vs. how often they should appear.

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{p_{d(q),(a_i,a_j)}}\right)$$

Skew = 0, VLM is unbiased

UCF

Skew@K

- Measures how often a group appears vs. how often they should appear.

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{p_{d(q),(a_i,a_j)}}\right)$$

Skew ≠ 0, VLM is biased

UCF

Skew@K: Dummy Example

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{p_{d(q),(a_i,a_j)}}\right)$$

Query: "*A Photo of Female Accountant*"

Skew@K: Dummy Example

$$\text{Skew}_{(a_i,a_j)}@K(R_K(q)) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{p_{d(q),(a_i,a_j)}}\right)$$

Query: "*A Photo of Female Accountant*"

$$p_{R_K(q),(a_i,a_j)}$$

UCF

Skew@K: Dummy Example



- Query: *"A Photo of Female Accountant."*
- *K*: 12 (# of retrieved images)
- (a_i , a_j): Middle Eastern, Female.
- $d_{(q)\,(ai\,,\,aj)}$: 200 images of Accountant.
- $d_{(q)\,(ai\,,\,aj)}$: 20 images of Middle Eastern Female Accountants.
- R_k: 12 retrieved images.

Then,

$$p_{R_K(q),(a_i,a_j)} = \frac{4}{12} = 0.33 \text{ and } p_{d(q),(a_i,a_j)} = \frac{20}{200} = 0.1$$

$$Skew_{(a_i,a_j)}@K\big(R_K(q)\big) = \log\left(\frac{p_{R_K(q),(a_i,a_j)}}{p_{d(q),(a_i,a_j)}}\right)$$

$$= \log\left(\frac{0.33}{0.1}\right) \approx 1.20$$

UCF

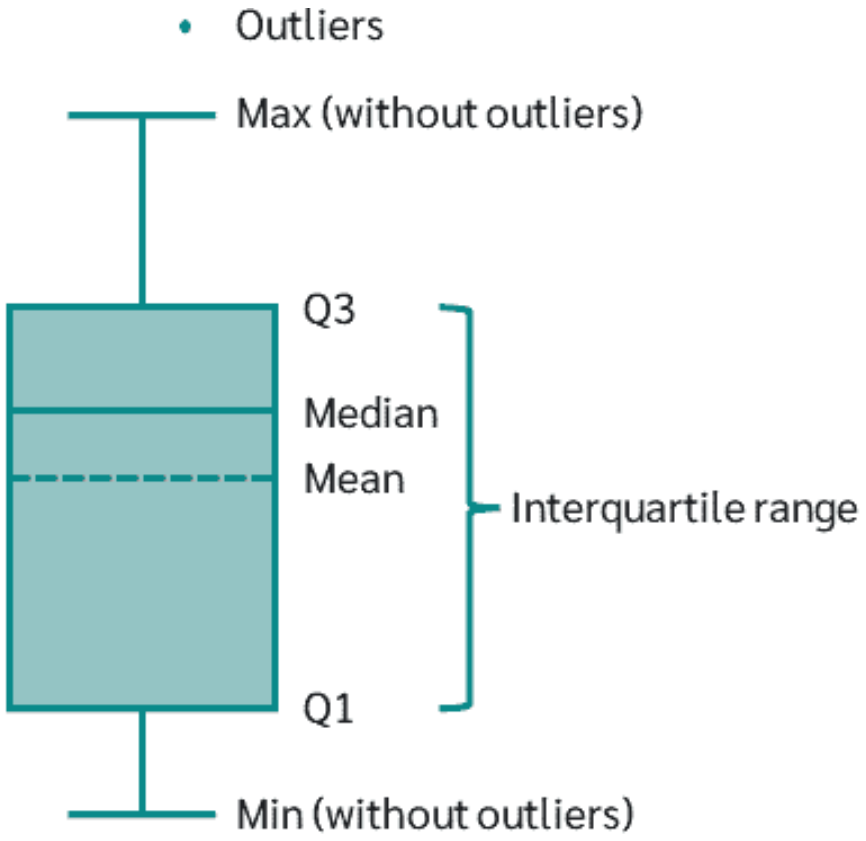
MaxSkew@K: Dummy Example

$$\text{MaxSkew@}K(R_K(q)) = \max_{(a_i,a_j)\in A} \text{Skew}_{(a_i,a_j)}@K(R_K(q))$$

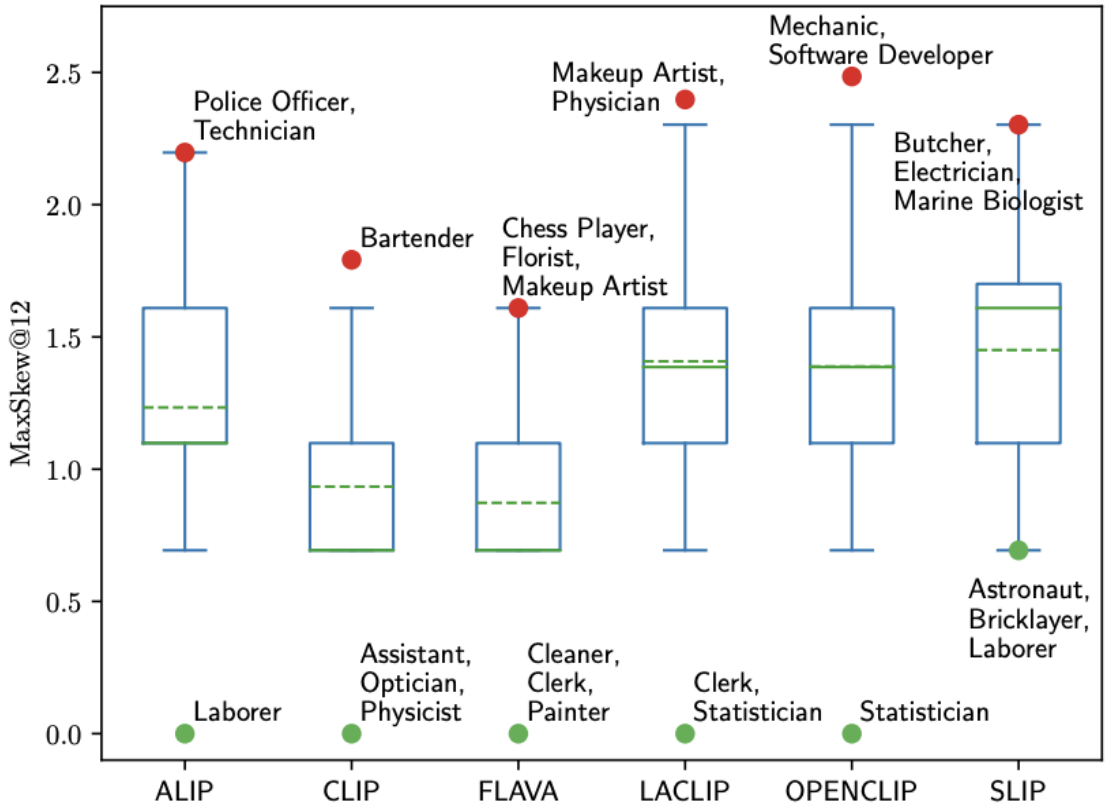
- Measures the maximum bias across all the attributes pairs.
- For a given subject (occupation), find Skew value for all attribute pairs.
- Higher MaxSkew@K = More bias in the VLM.

Race, Gender	Skew Score	Race, Physical Characteristics	Skew Score	Gender, Physical Characteristics	Skew Score
White Male	1.269	Black Skinny	0.247	Female Young	1.742
Indian Male	1.983	Asian Tattooed	0.143	Male Old	0.420
Latino Female	0.690	Middle Eastern Old	2.368	Female Tattooed	0.937
MaxSkew: 2.368					

Box Plot Explanation



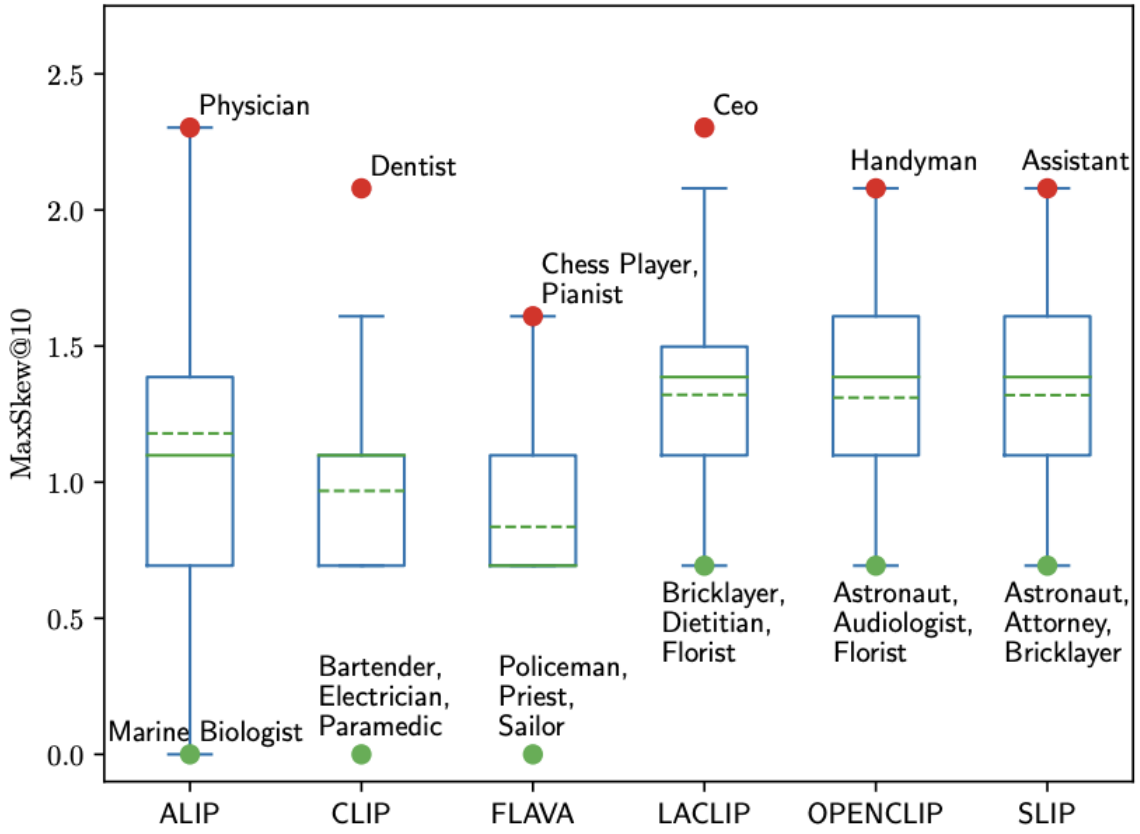
MaxSkew@k distribution for VLMs (Race – Gender)



Race-Gender

- CLIP and FLAVA show the lowest median bias.
- ALIP has occupations with zero skew (like “Laborer”).

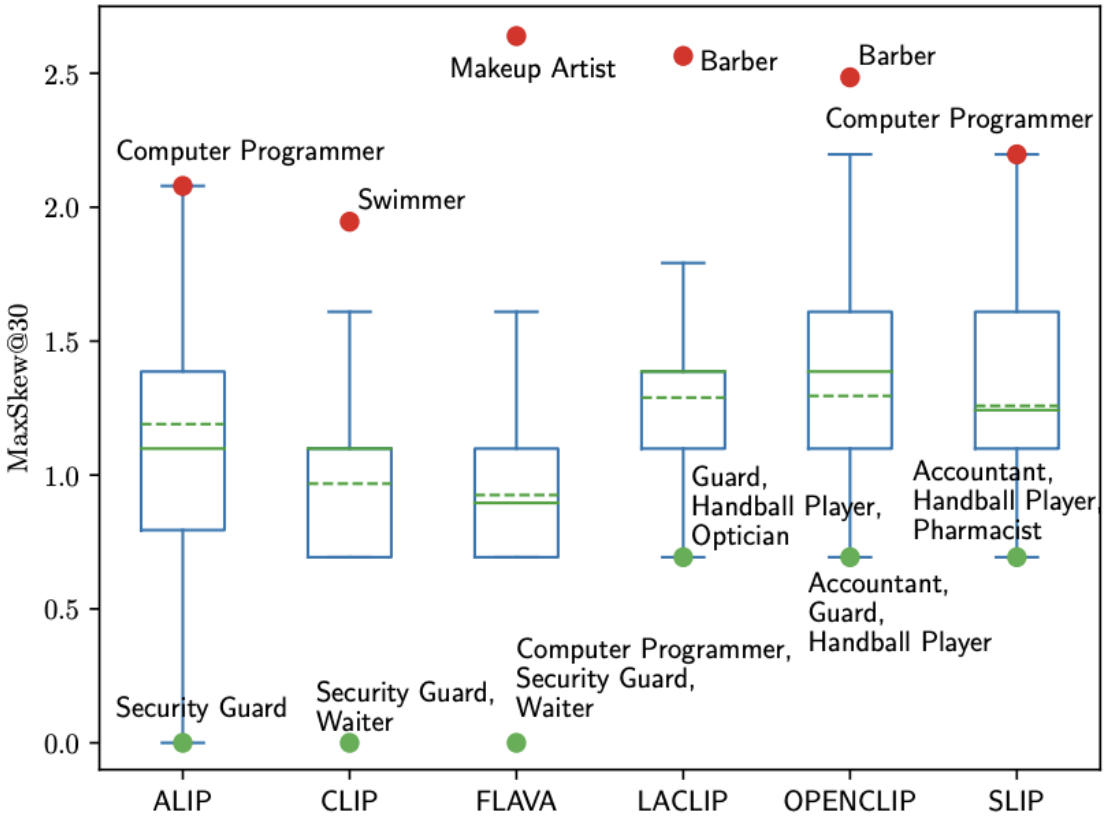
MaxSkew@k distribution for VLMs (Physical Char. - Gender)



Physical Char. - Gender

- FLAVA show the lowest median bias.
- LACLIP, OPENCLIP, and SLIP show consistent bias across all pairs.

MaxSkew@k distribution for VLMs (Physical Char. - Race)

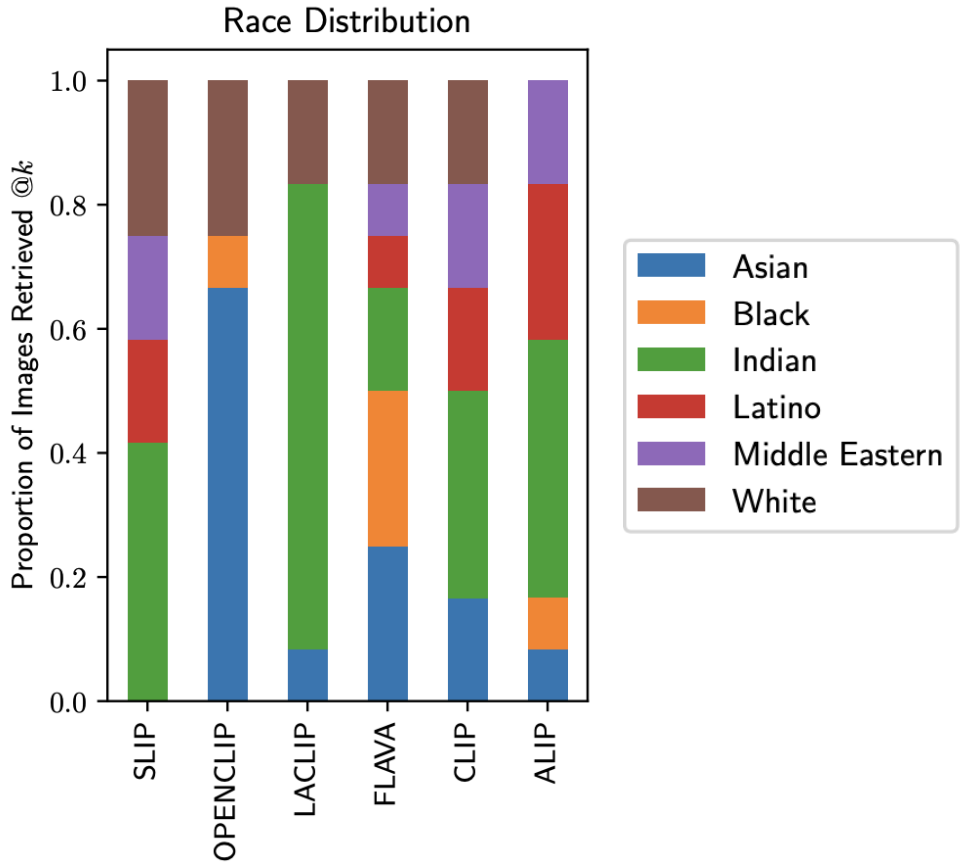


Physical Char. - Race

- FLAVA show the lowest median bias, but a high variance ("Makeup Artist").
  - "Security Guard" shows consistent low bias for ALIP, CLIP, FLAVA.

Quantitative Analysis:

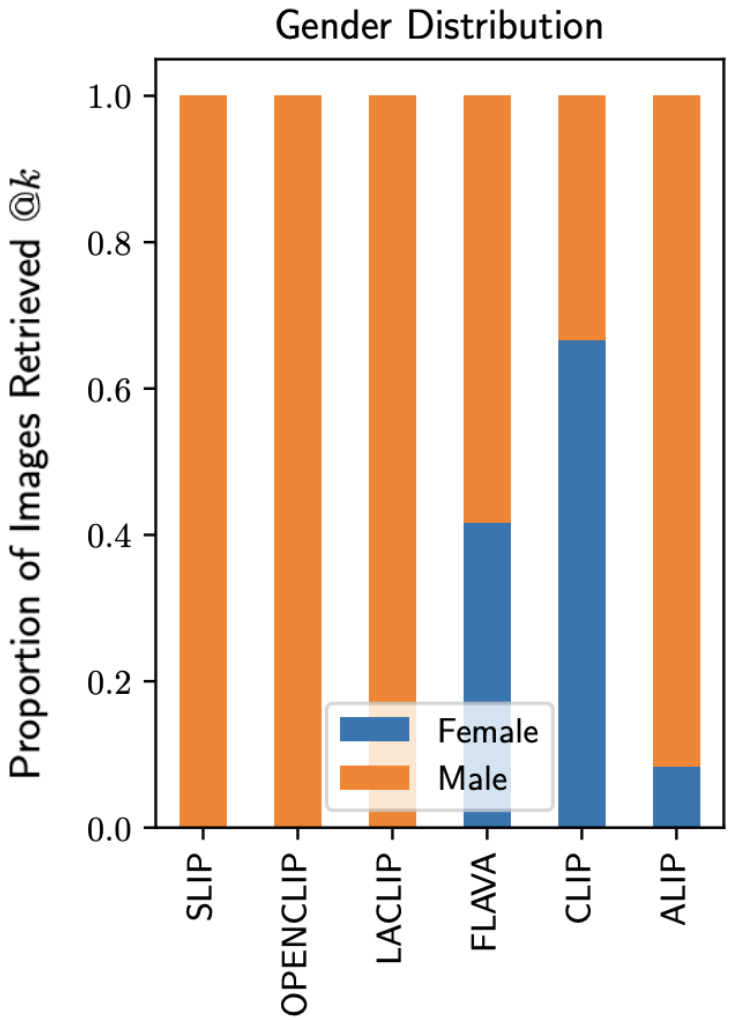
- OpenCLIP and LaCLIP clearly show high skew (3 races).
- FLAVA has lowest skewness (6 races).



Retrieving images for subject "Doctor"

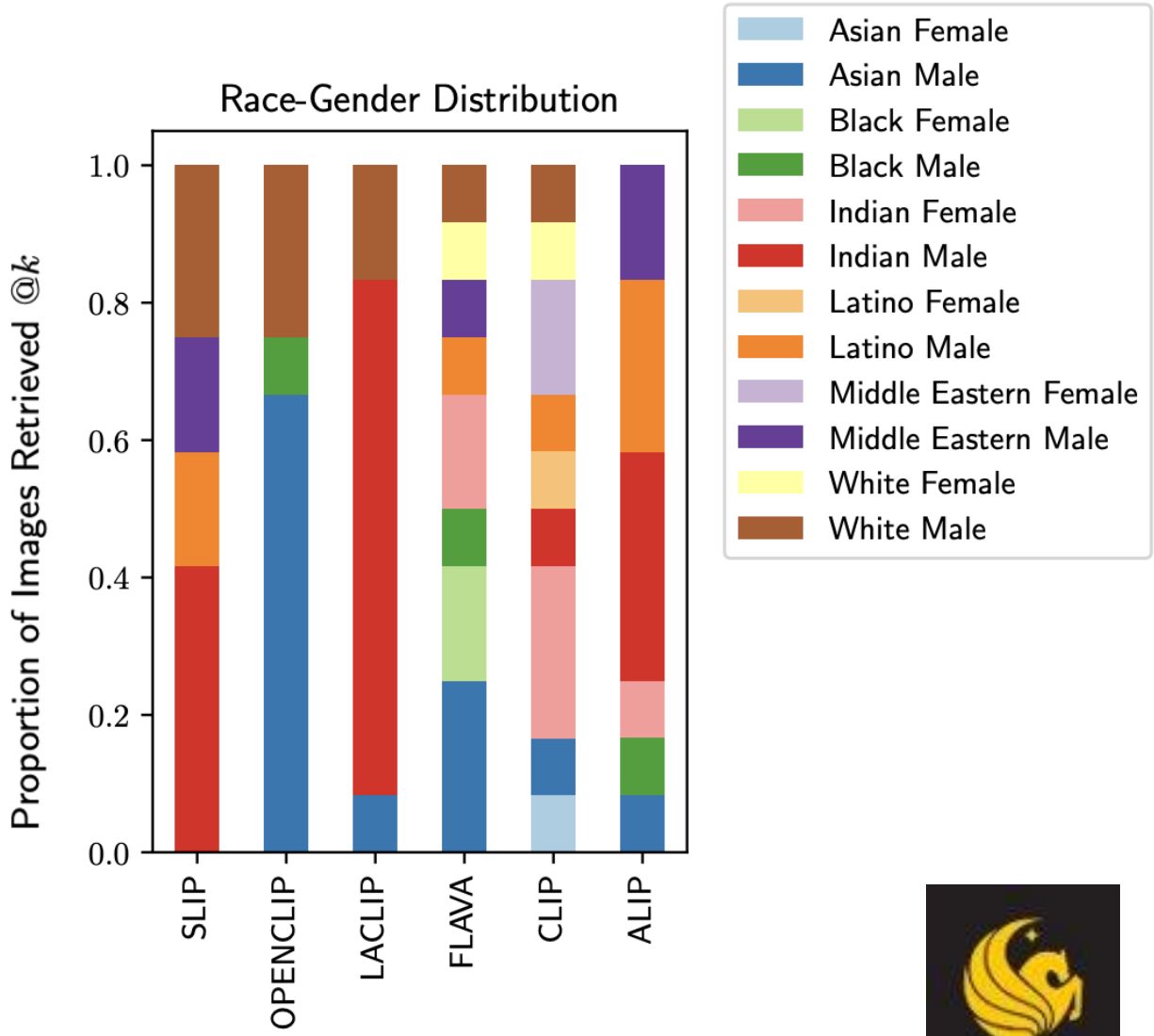
Quantitative Analysis:

- SLIP, OpenCLIP, and LaCLIP are biased towards predicting only male gender.
- FLAVA and CLIP have lower gender skew.

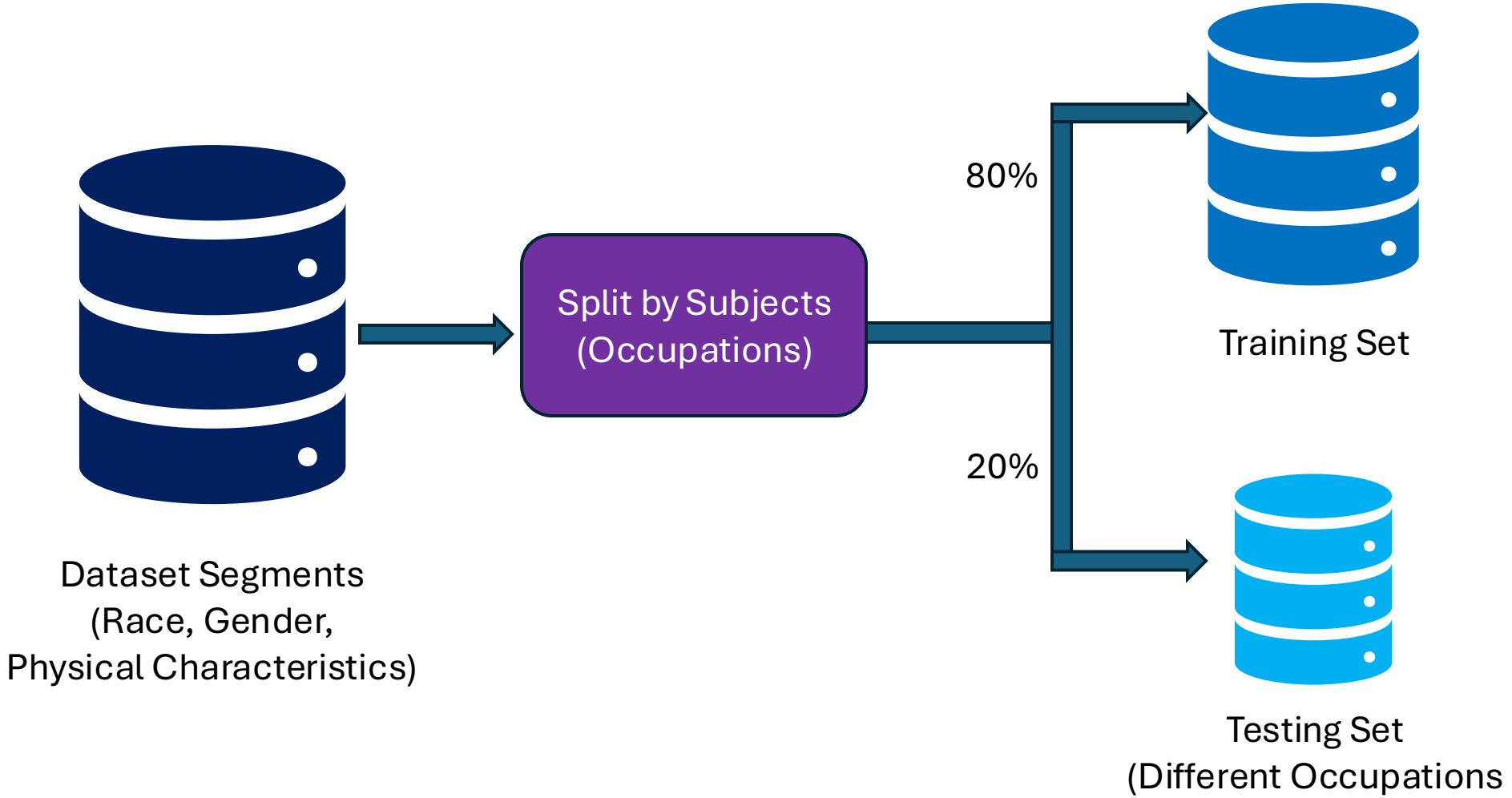


Quantitative Analysis:

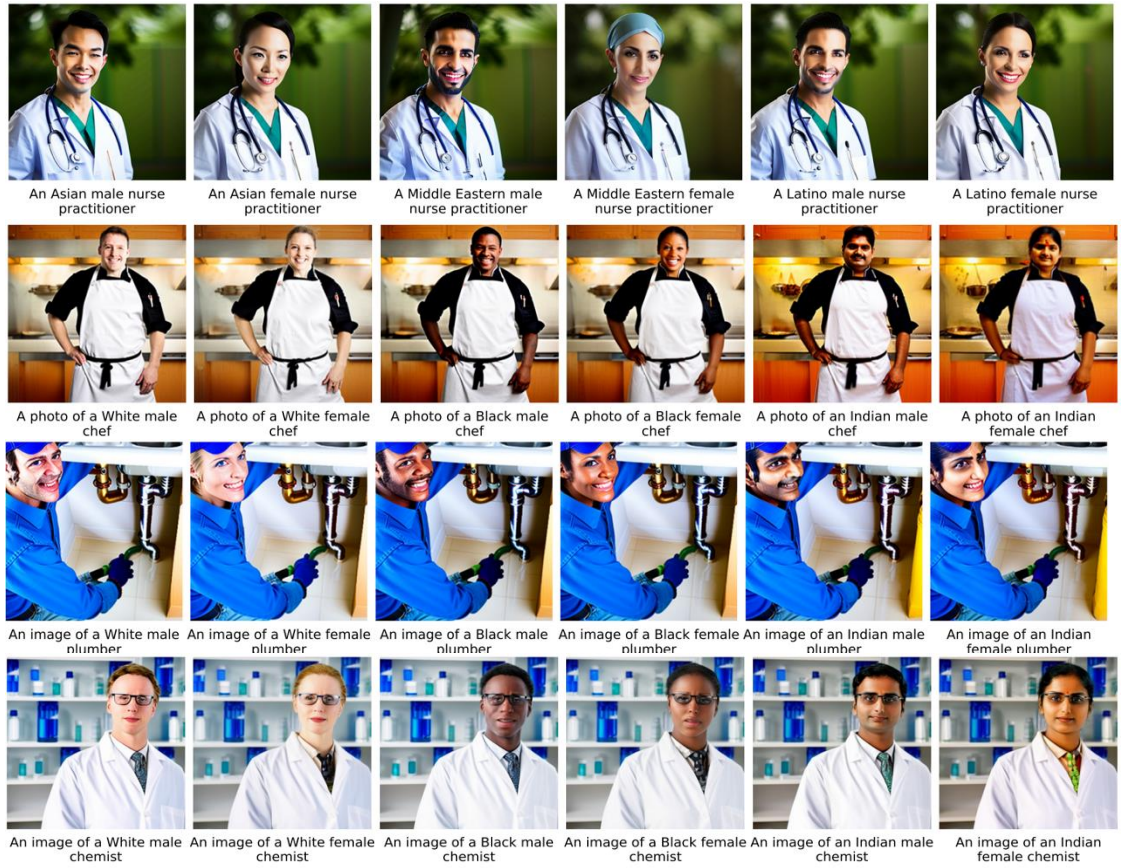
- Only FLAVA shows "Black Female" doctors.
- OpenCLIP and LaCLIP show highest bias.



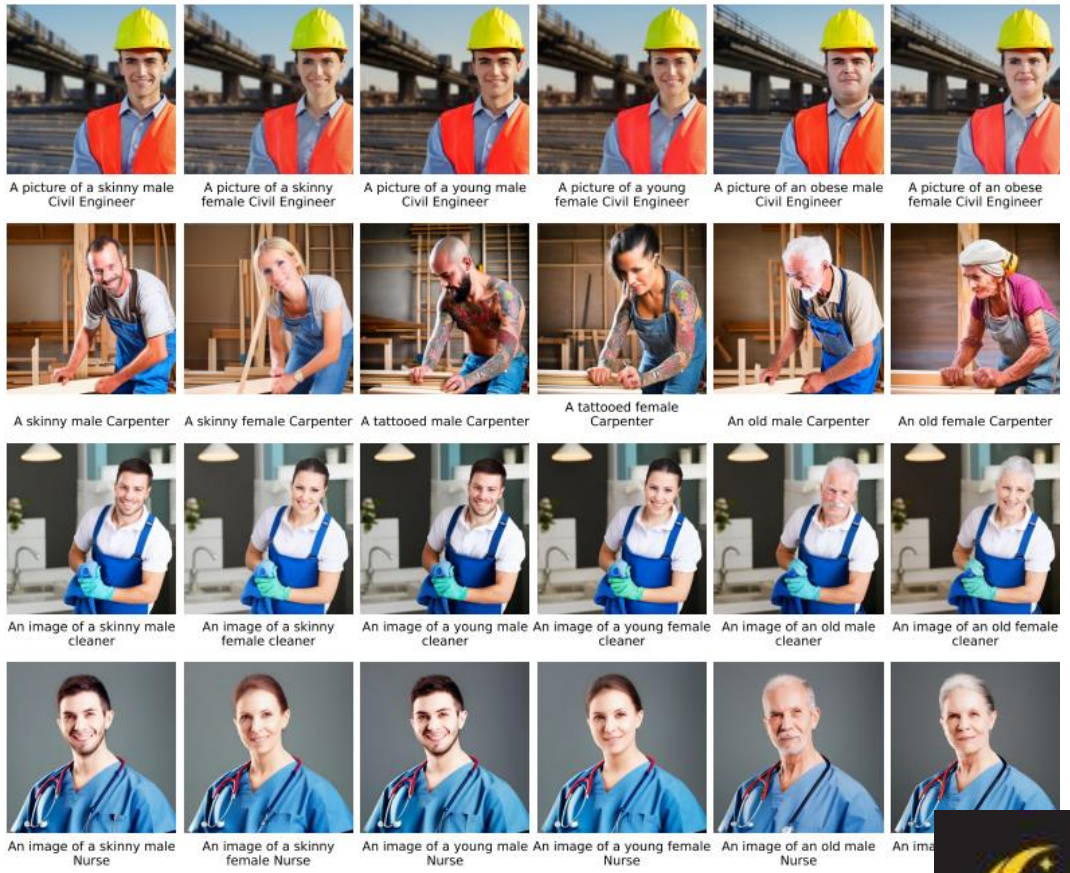
Training (Dataset Split)



Training (Dataset Split)

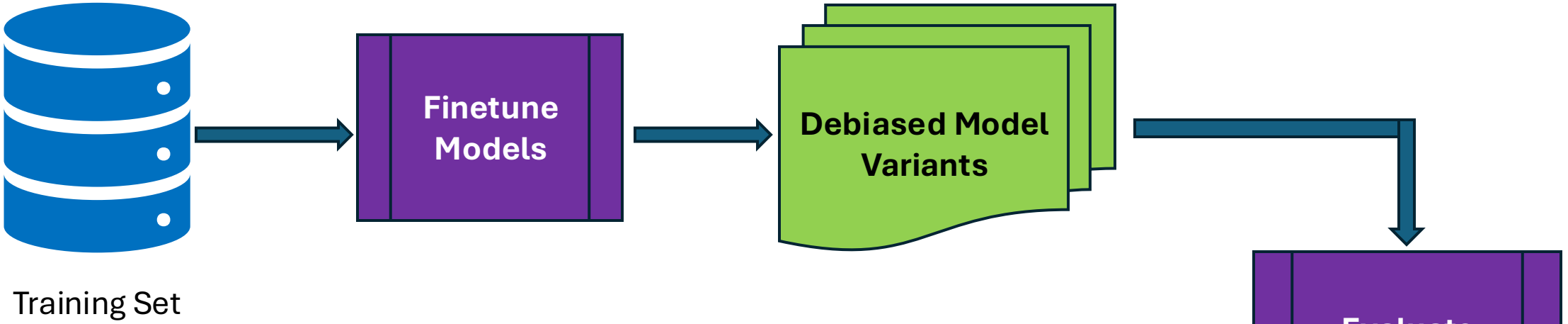


Training set



Testing set

Training (Model Finetuning)



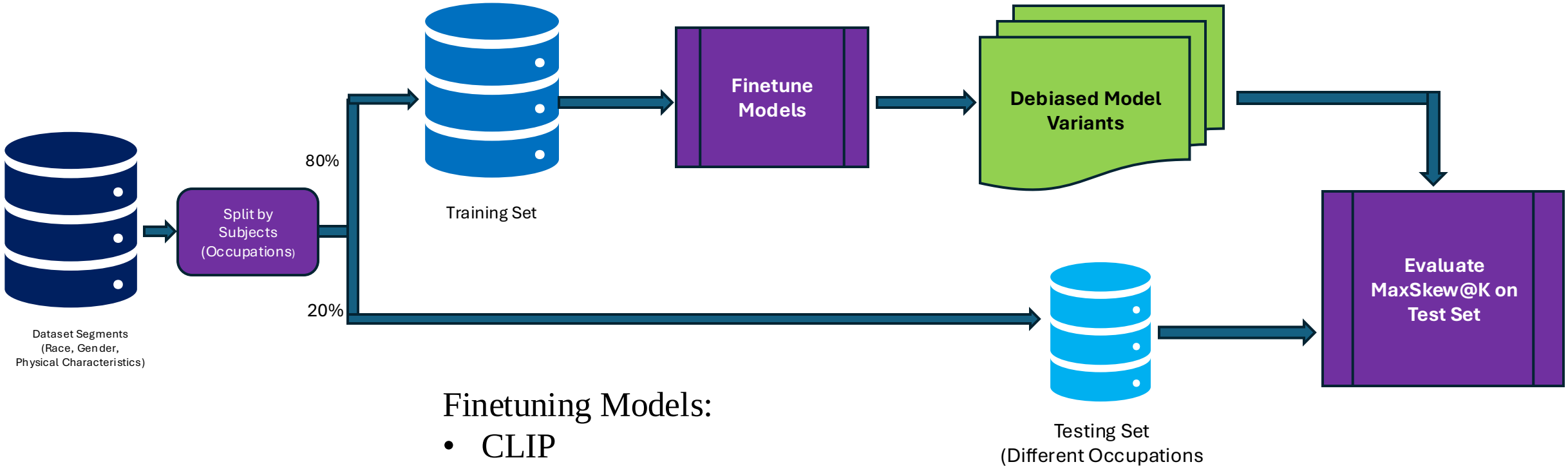
Training Set

Finetuning Models:

- CLIP
- ALIP
- FLAVA

Testing Set
(Different Occupations

Training (Entire pipeline)



Finetuning Models:

- CLIP
- ALIP
- FLAVA

Debiasing Results

	CLIP [41]		ALIP [59]		FLAVA [46]	
Intersectional Bias	**Pre-trained**	**Debiased**	**Pre-trained**	**Debiased**	**Pre-trained**	**Debiased**
(Race, Gender)	1.02		1.40		0.98	
(Physical Char., Gender)	0.87		1.28		0.92	
(Physical Char., Race)	1.19		1.55		1.12	

UCF

Debiasing Results

	CLIP [41]		ALIP [59]		FLAVA [46]	
Intersectional Bias	**Pre-trained**	**Debiased**	**Pre-trained**	**Debiased**	**Pre-trained**	**Debiased**
(Race, Gender)	1.02	**0.77**	1.40	**1.16**	0.98	**0.79**
(Physical Char., Gender)	0.87	**0.71**	1.28	**1.02**	0.92	**0.81**
(Physical Char., Race)	1.19	**0.87**	1.55	**1.13**	1.12	**0.97**

Bias Generalization
to Unseen Occupations

UCF

Debiasing Results

Social Counterfactuals is AI-generated to mitigate intersectional biases.
What about real world bias?

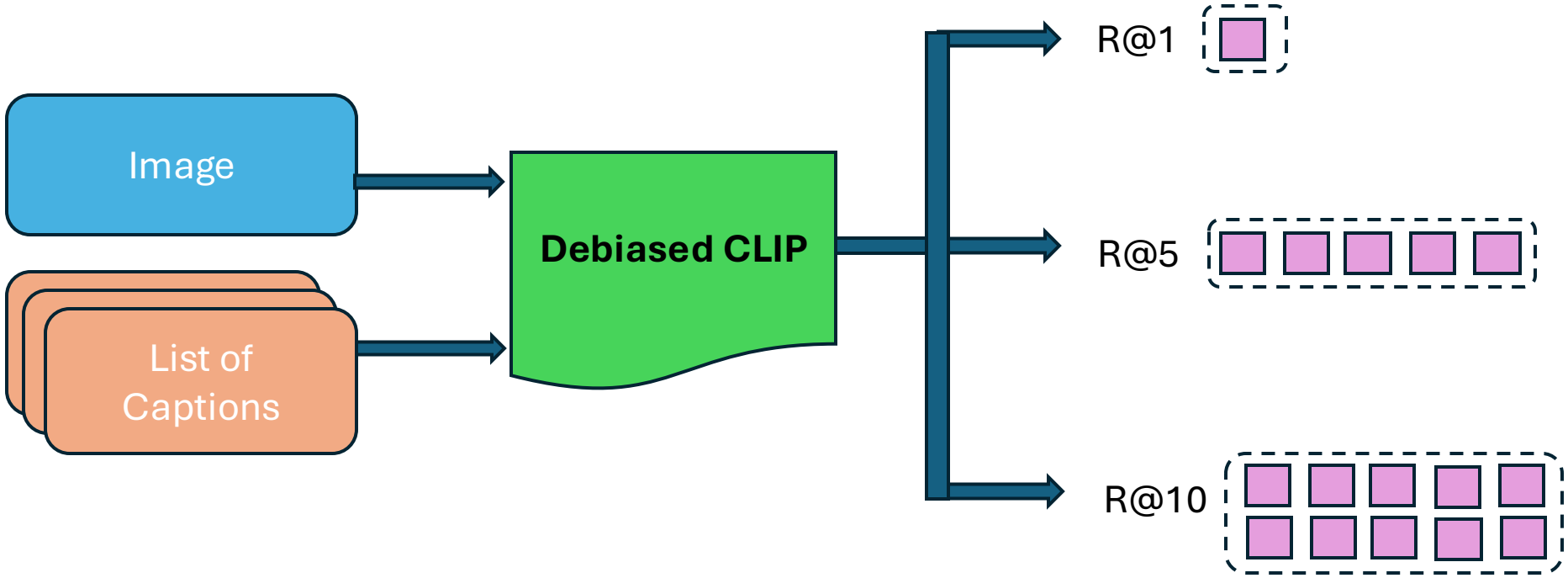
Model	VisoGender [23]	PATA [45]
Pre-trained CLIP	0.269	0.323
Debiased CLIP	**0.219**	**0.283**

- **VisoGender**: Benchmark for gender bias in image-text pronoun resolution.
- **PATA**: Images of different race, age, and gender protected attributes in various scenarios

Retrieval-based approach

- Text-Retrieval shows consistent, if not better results.

CLIP Model	Text Retrieval R@1	R@5	R@10
Pre-trained	67.1	89	93.8
Debiased	**69.2**	**89.6**	93.8



Retrieval-based approach

- Reduction in performance is consistent with previous efforts in debiasing CLIP[2].
- Trade-off between model fairness and performance.

	Image Retrieval		
CLIP Model	**R@1**	**R@5**	**R@10**
Pre-trained	**69.4**	**90.6**	**94.9**
Debiased	67.3	90.4	93.8

CLIP Model	**CIFAR10**[32]	**CIFAR100**[32]	**Caltech256**[21]	**ImageNet**[12]
Pre-trained	**88.80**	**64.17**	**83.43**	**59.25**
Debiased	86.72	61.46	79.72	55.38

[2] Ashish Seth, Mayur Hemani, and Chirag Agrawal. Dear: Debiasing vision-language models with additive residuals. In CVPR, 2023.

UCF

Limitations

- Latent bias within caption template and experimental setup
- Binary classification of genders
- Does not encompass all races, physical char., and occupations
  - Excluded groups still facing bias in the models
- Categories are not representative of groups
  - Results reflect a small part of the group