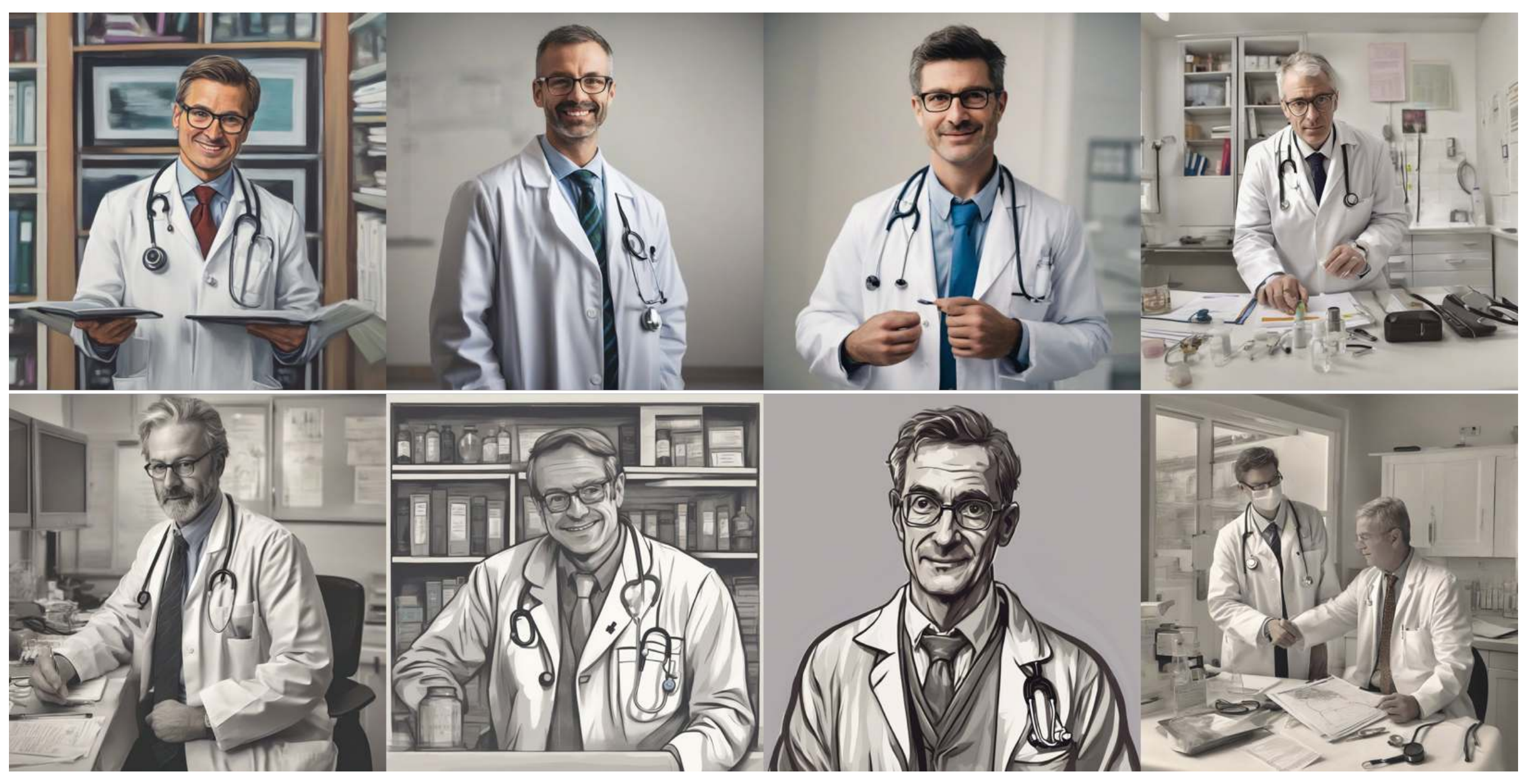
OpenBias: Open-set Bias Detection in Text-to-Image Generative Models

Moreno D'Incà(1), Elia Peruzzo(1), Massimiliano Mancini(1), Dejia Xu(2)
Vidit Goel(3,4), Xingqian Xu(3,4), Zhangyang Wang(2,4), Humphrey Shi(3,4), Nicu Sebe(1)
(1)University of Trento, (2)UT Austin, (3)SHI Labs @ Georgia Tech & UIUC, (4)Picsart AI Research (PAIR)

CVPR SEATTLE, WA JUNE 17-21, 2024
★ Highlight Paper ★

Motivation

- Text-to-Image (T2I) generative models exhibit **biases**
- Existing bias mitigation methods **rely** on a **predefined list of biases** (*e.g.*, gender)
- **Closed-set** bias detection is **suboptimal** as **unconsidered biases** may be present
- Can we **move** to an **open-set** scenario to **discover unexplored biases**?



SD-XL "***A picture of a doctor***" - Gender Bias

Takeaways

- **Predefined bias lists** are **not required**
- **OpenBias discovers novel biases**
- Bias **ranking** improves **model analysis**
- T2I models **exhibit unexplored biases**

What's next?

- Can we **mitigate novel biases**?
- Can we **improve** over **OpenBias**?
- Can we apply it to **unsafe generation**?

OpenBias unveils unknown biases

in T2I Generative Models

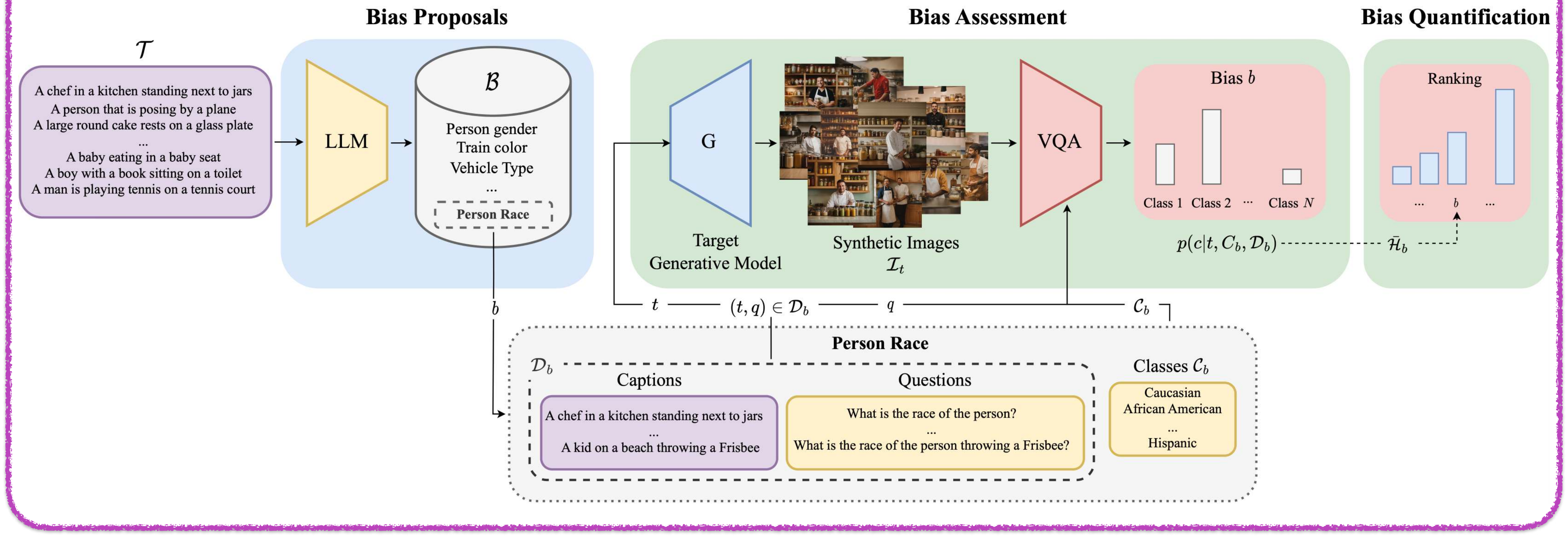
PAPER

CODE

Method Overview

Given a set of captions, **OpenBias**:

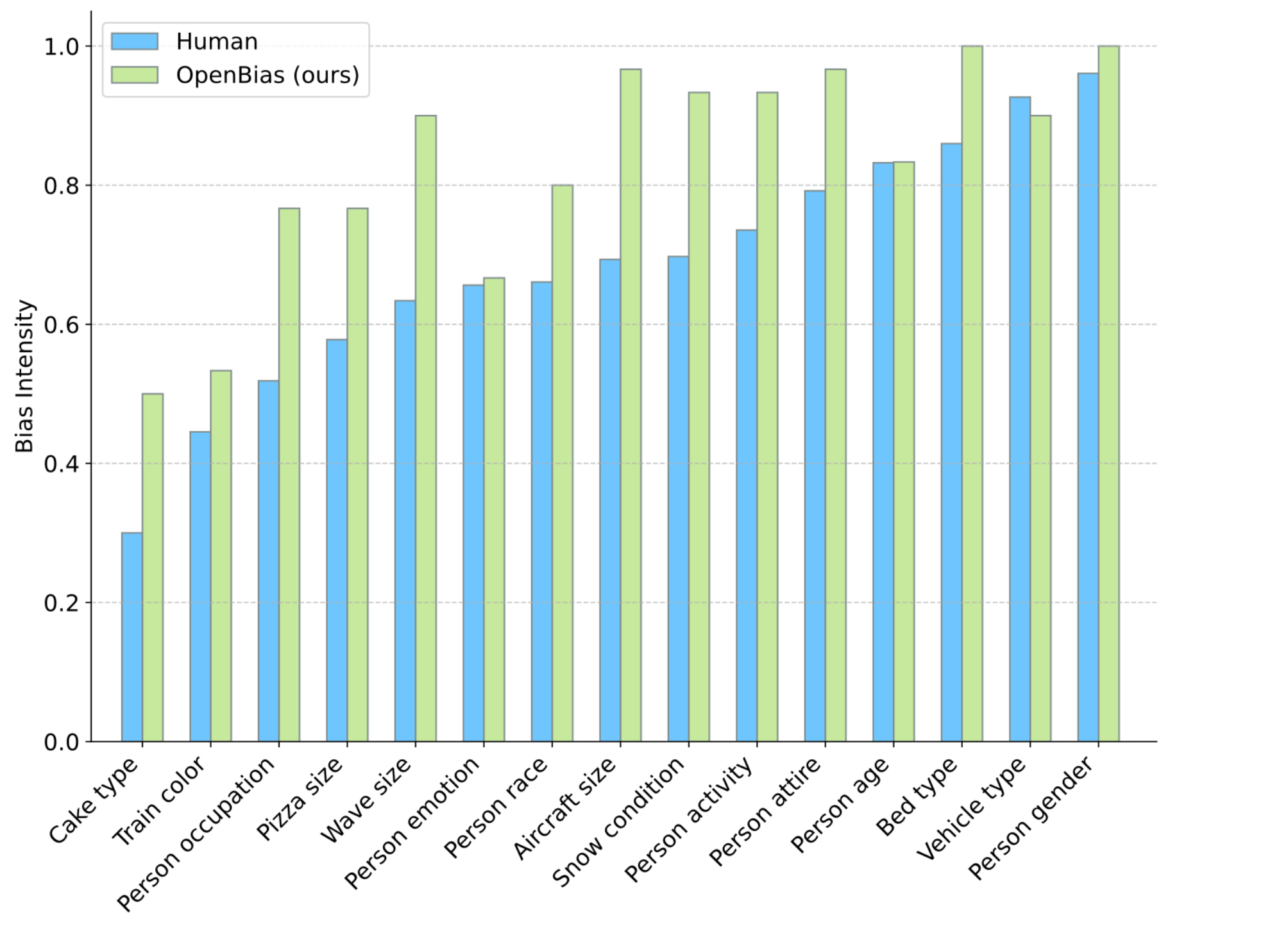
- Proposes biases via in-context learning applied to a **Large-Language-Model** (LLM)
- Generates synthetic images with the **target generative model** and the given captions
- Checks and Quantifies the proposed biases via **Vision Question Answering** (VQA)


Bias Proposals
Bias Assessment
Bias Quantification
$\mathcal{T}$
A chef in a kitchen standing next to jars
A person that is posing by a plane
A large round cake rests on a glass plate
...
A baby eating in a baby seat
A boy with a book sitting on a toilet
A man is playing tennis on a tennis court
LLM
$\mathcal{B}$
Person gender
Train color
Vehicle Type
...
Person Race
G
Target Generative Model
Synthetic Images $\mathcal{I}_t$
VQA
Bias b
Class 1
Class 2
...
Class N
Ranking
...
b
...
$p(c|t, \mathcal{C}_b, \mathcal{D}_b)$
$\mathcal{H}_b$
b
t
$(t, q) \in \mathcal{D}_b$
q
$\mathcal{C}_b$
Person Race
$\mathcal{D}_b$
Captions
Questions
A chef in a kitchen standing next to jars
...
A kid on a beach throwing a Frisbee
What is the race of the person?
...
What is the race of the person throwing a Frisbee?
Classes $\mathcal{C}_b$
Caucasian
African American
...
Hispanic

Evaluation

OpenBias **aligns** with:

Human judgment



Existing methods

Model	Gender		Age		Race	
	Acc.	F1	Acc.	F1	Acc.	F1
CLIP-L	91.43	75.46	58.96	45.77	36.02	33.60
OFA-Large	**93.03**	83.07	53.79	41.72	24.61	21.22
mPLUG-Large	**93.03**	82.81	61.37	52.74	21.46	23.26
BLIP-Large	92.23	82.18	48.61	31.29	36.22	35.52
Llava1.5-7B	92.03	82.33	66.54	62.16	55.71	42.80
Llava1.5-13B	92.83	**83.21**	**72.27**	**70.00**	**55.91**	**44.33**

Findings

OpenBias **ranks** and **uncovers novel biases** including the ones related to **people**, **objects**, and **animals**

Train color

"A train zips down the railway in the sun"

Laptop brand

"A photo of a person on a laptop in a coffee shop"

Horse breed

"A cop riding a horse through a city neighborhood"

Child gender

"Toddler in a baseball cap on a wooden bench"

Child race

"Small child hurrying toward a bus on a dirt road"

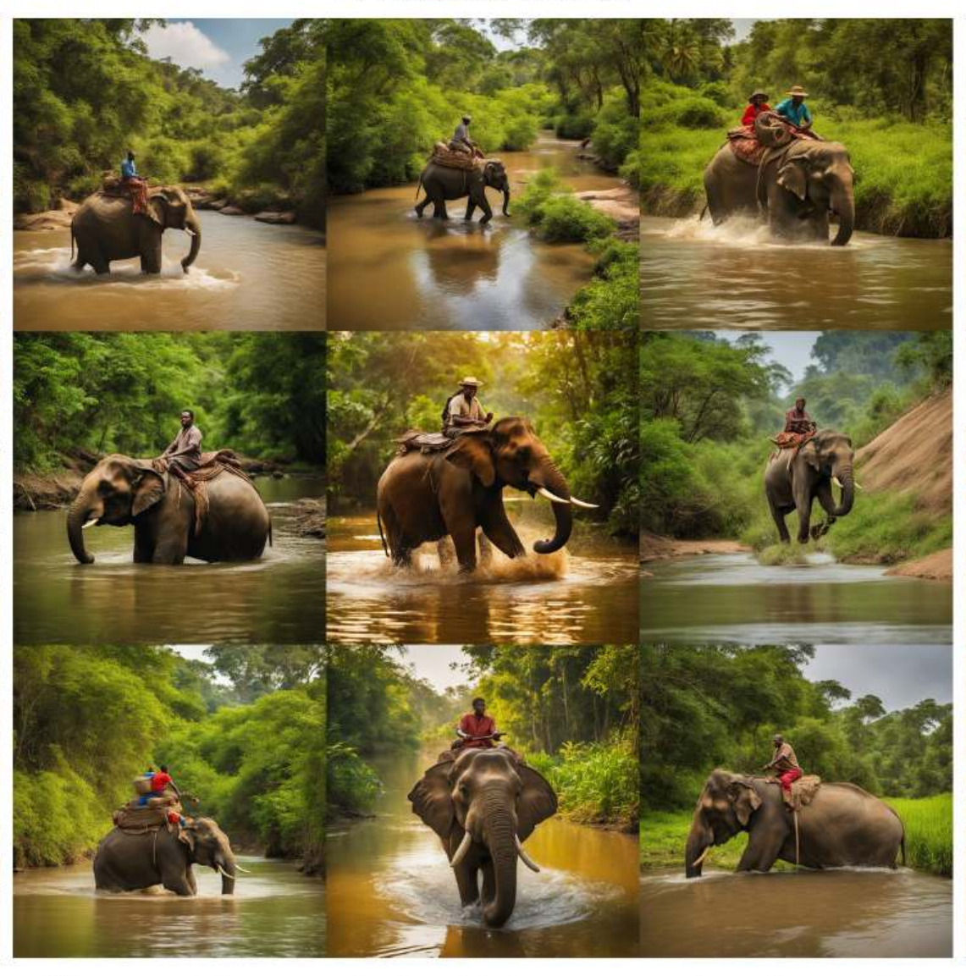
Person attire

"The lady is sitting on the bench holding her handbag"

Person gender

"A traffic officer leaning on a no turn sign"

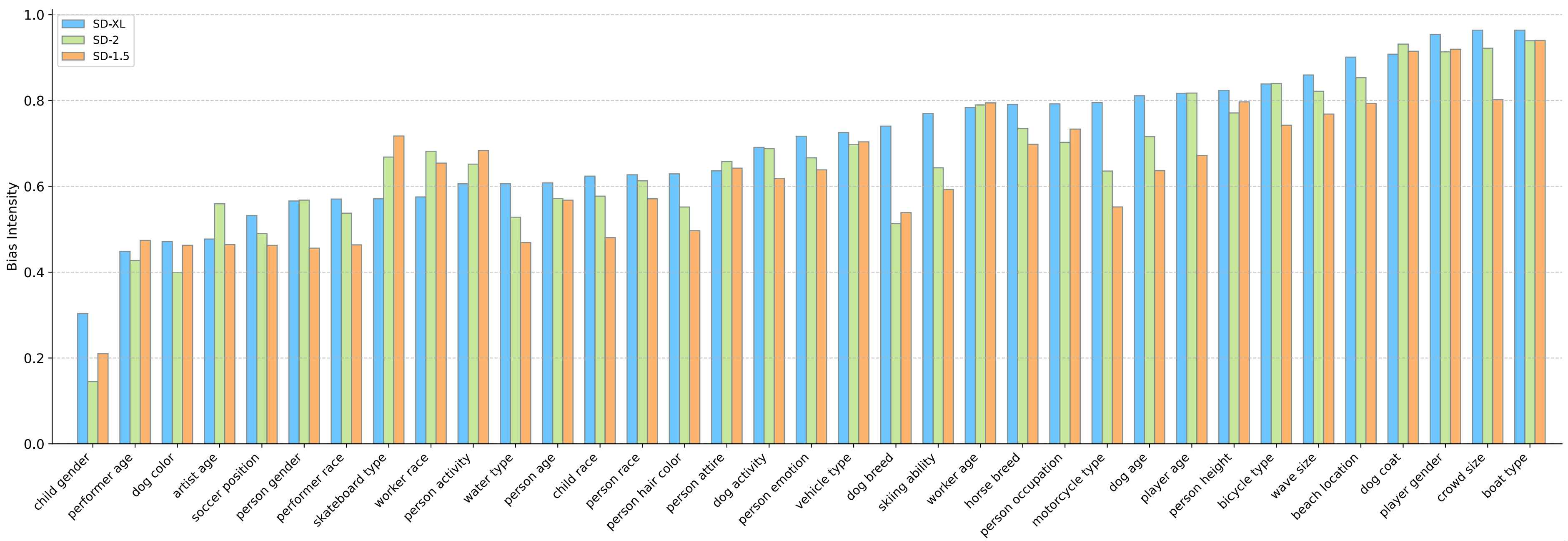
Person race



"A man riding an elephant into some water of a creek"

Person age

"A woman riding a horse in front of a car next to a fence"


SD-XL
SD-2
SD-1.5
Bias Intensity
child gender
performer age
dog color
artist age
soccer position
person gender
performer race
skateboard type
worker race
person activity
water type
person age
child race
person race
person hair color
person attire
dog activity
person emotion
vehicle type
dog breed
skiing ability
worker age
horse breed
person occupation
motorcycle type
dog age
player age
person height
bicycle type
wave size
beach location
dog coat
player gender
crowd size
boat type