FHIBE Bias Evaluation Report

Task: Person Localization

Model: faster_rcnn_2025June30

Dataset version: DEC24

Reported generated on June 30, 2025

Task Description

The person localization task evaluates how well a model can predict person bounding boxes in the FHIBE images. It is evaluated on the full body FHIBE dataset. Each image contains one or two people, with one ground truth bounding box per person.

Evaluation Dataset

The model was evaluated using the downsampled FHIBE dataset. This dataset consists of 10319 images comprising one or two people in a background setting. Each image is resized such that the larger dimension has 2048 pixels, while maintaining the aspect ratio of the original, higher resolution image.

Accompanying the dataset is a rich set of metadata. The metadata include self-reported demographic information, such as age, ancestry, and gender pronouns of image subjects, as well as annotator-provided metadata for more objective physical attributes such as pose and subject actions. The metadata also characterize the scene via lighting and weather descriptions. Drawn annotations are also provided for person bounding boxes, face bounding boxes, full body keypoints, person segmentation masks, and face segmentation masks.

The metadata specific to this task include the person bounding box annotations.

The following attributes were used to aggregate results in this report. For descriptions and definitions of these attributes, see the supplement.

Attributes: pronoun, age, ancestry, apparent_skin_color

Evaluation Results

Here we report the results for each metric that was evaluated.

Metric 1. AR_IOU: Average recall over intersection over union.

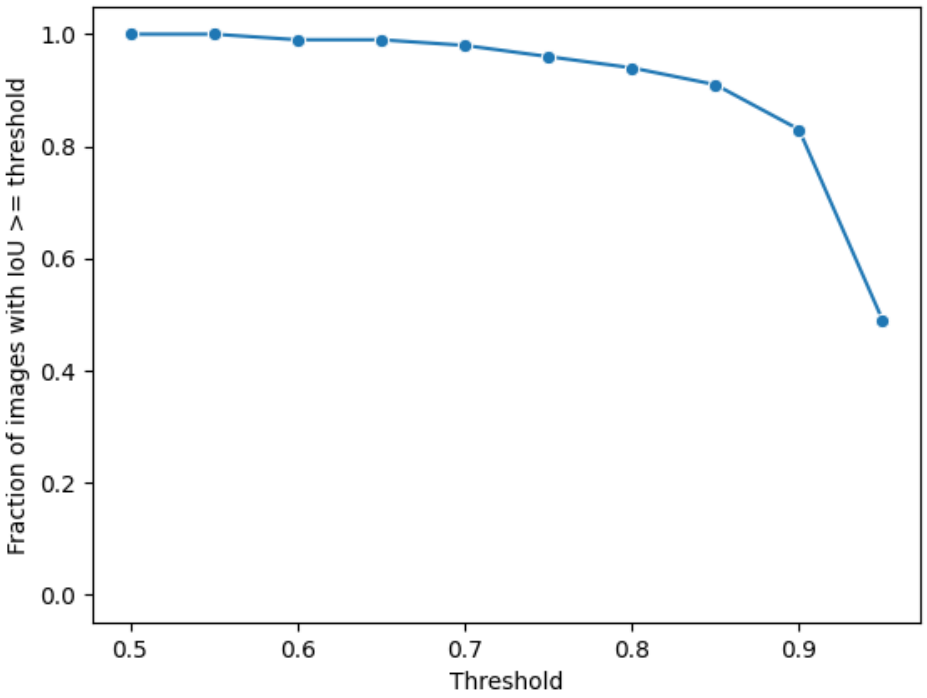
Description:

For each ground truth bounding box, the best IoU out of all predicted bounding boxes is obtained. At each value in a list of IoU thresholds between 0 and 1, each image is given a value of 1 (correct) or 0 (incorrect) based on whether IoU > threshold. Using these binary outcomes, the recall is calculated. The average recall over all thresholds is reported.

Thresholds: 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95

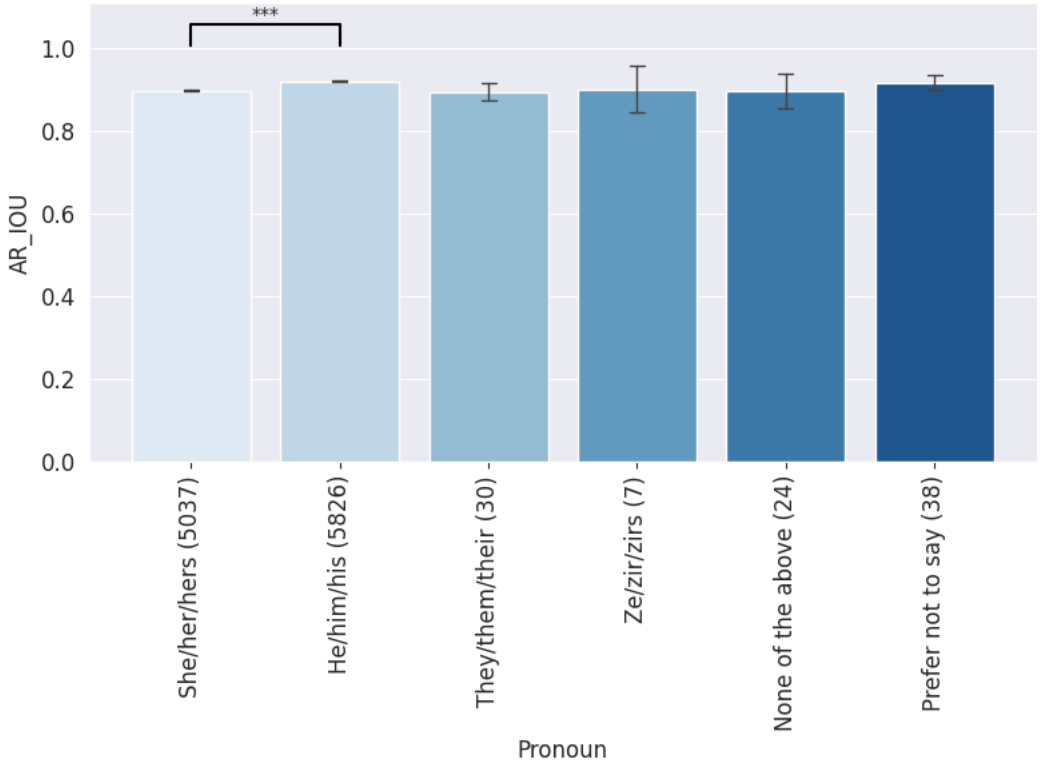
Results:

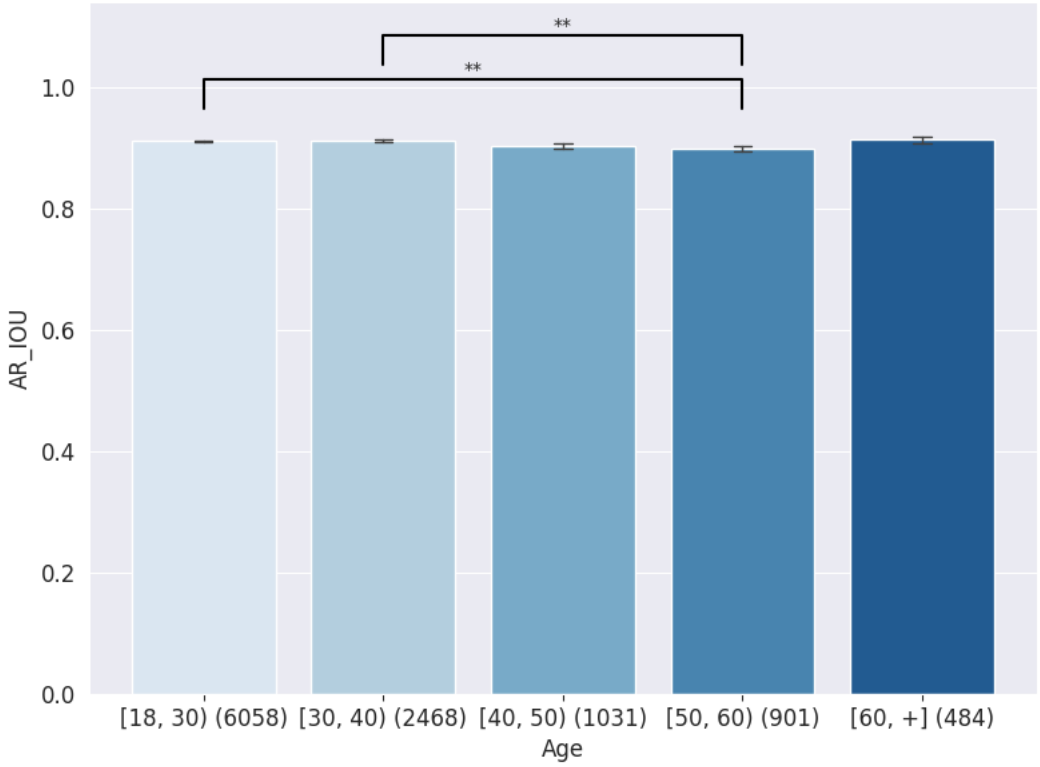
IoU vs. threshold curve calculated over all images

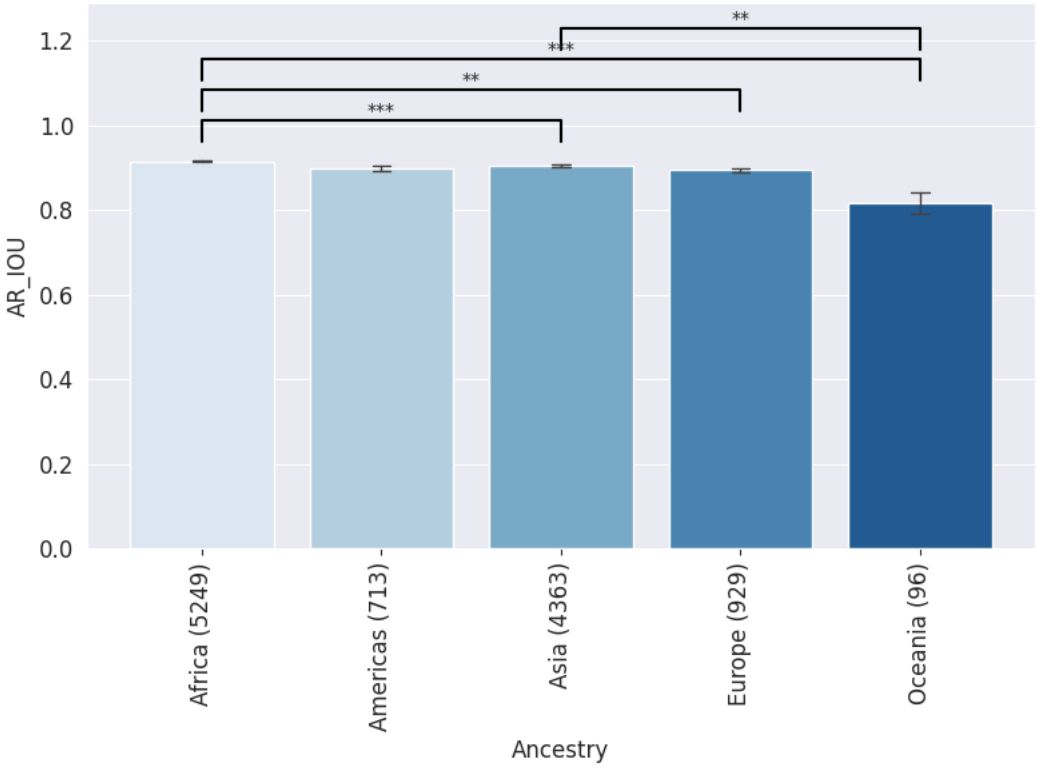


Metric performance aggregated in demographic groups

Bar plots show mean and 68% confidence intervals. Group sizes are included in parentheses in the x-axis labels. Significant differences are indicated with brackets and asterisks, where * indicates p<0.05/m, ** indicates p<0.01/m, and *** indicates p<0.001/m, where p is the p-value and and m is the number of attribute pairs tested for significance (Bonferroni correction). Groups with size < 20 are not included in significance calculations.


***
1.0
0.8
0.6
0.4
0.2
0.0
AR_IOU
She/her/hers (5037)
He/him/his (5826)
They/them/their (30)
Ze/zir/zirs (7)
None of the above (24)
Prefer not to say (38)
Pronoun


**
**
1.0
0.8
0.6
0.4
0.2
0.0
AR_IOU
[18, 30) (6058)
[30, 40) (2468)
[40, 50) (1031)
[50, 60) (901)
[60, +] (484)
Age


**
***
**
***
1.2
1.0
0.8
0.6
0.4
0.2
0.0
AR_IOU
Africa (5249)
Americas (713)
Asia (4363)
Europe (929)
Oceania (96)
Ancestry

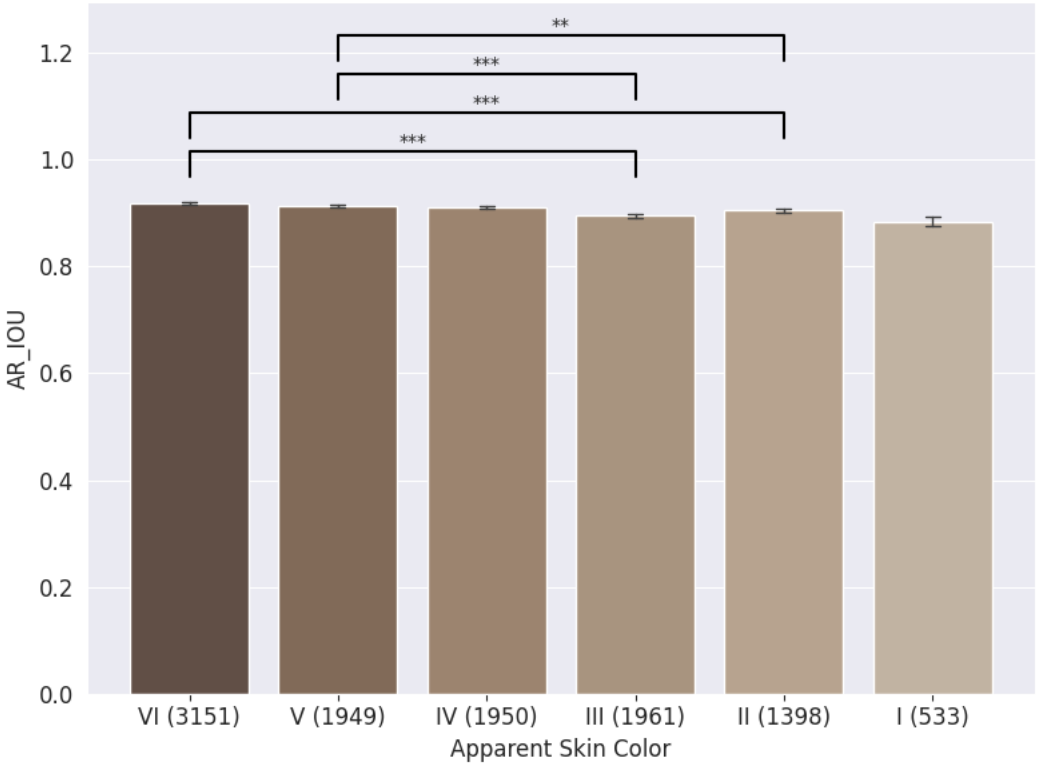

**
***
***
***
1.2
1.0
0.8
0.6
0.4
0.2
0.0
AR_IOU
VI (3151)
V (1949)
IV (1950)
III (1961)
II (1398)
I (533)
Apparent Skin Color

Table of highest disparity in metric performance across all attribute intersections

Disparity is defined as 1 - (AVG(worst group) / AVG(best group))

Only statistically significant results are shown. Significance is determined via p < alpha/m, where p is the p value, alpha=0.05, and m is the number of attribute pair for a single attribute (or attribute combination).

Attributes with highest disparities in AR_IOU performance							
Attribute(s)	**Worst group**	**Worst mean/median (class size)**	**Best group**	**Best mean/median (class size)**	**Mean/median disparity**	**P value**	**Effect size**
• Pronoun	• She/her/hers	0.90/0.90 (5037)	• He/him/his	0.92/1.00 (5826)	0.02 / 0.10	< 0.001	9.35
• Ancestry	• Asia	0.91/0.90 (4363)	• Africa	0.92/1.00 (5249)	0.01 / 0.10	< 0.001	3.58
• Ancestry	• Europe	0.90/0.90 (929)	• Africa	0.92/1.00 (5249)	0.02 / 0.10	< 0.001	3.27
• Ancestry	• Oceania	0.82/0.90 (96)	• Africa	0.92/1.00 (5249)	0.11 / 0.10	< 0.001	3.51
• Apparent Skin Color	• III	0.90/0.90 (1961)	• VI	0.92/1.00 (3151)	0.03 / 0.10	< 0.001	4.50
• Apparent Skin Color	• II	0.91/0.90 (1398)	• VI	0.92/1.00 (3151)	0.02 / 0.10	< 0.001	3.88
• Apparent Skin Color	• III	0.90/0.90 (1961)	• V	0.91/1.00 (1949)	0.02 / 0.10	< 0.001	3.39
• Apparent Skin Color	• II	0.91/0.90 (1398)	• V	0.91/1.00 (1949)	0.01 / 0.10	0.001	2.94
• Age	• [50, 60)	0.90/0.90 (901)	• [18, 30)	0.91/0.90 (6058)	0.01 / 0.00	< 0.001	3.19
• Age	• [50, 60)	0.90/0.90 (901)	• [30, 40)	0.91/0.90 (2468)	0.01 / 0.00	< 0.001	3.18
• Ancestry	• Oceania	0.82/0.90 (96)	• Asia	0.91/0.90 (4363)	0.10 / 0.00	0.002	2.91

Supplement

In this section, we describe the attributes that were used to aggregate results in this report. The following attributes were grouped at the continent level: ancestry.

- Gender pronouns: self-reported. Multiple selections are allowed, except when 'None of the above' or 'Prefer not to say' is selected. Subjects select from the following options:
  'She/her/hers', 'He/him/his', 'They/them/their', 'Ze/zir/zirs', 'None of the above', 'Prefer not to say'
- Age: self-reported age between 0 and 130. We aggregate results in five age bins:
  [18, 30), [30, 40), [40, 50), [50, 60), [60, +]
- Ancestry: self-reported. Subjects select from a list of 27 geographic regions. Multiple selections are allowed.
- Apparent skin color: self-reported at time of image capture. Subjects select a single value from the six-point Fitzpatrick skin type scale:
  I (Very light), II (Light), III (Intermediate), IV (Tan), V (Brown), VI (Dark)