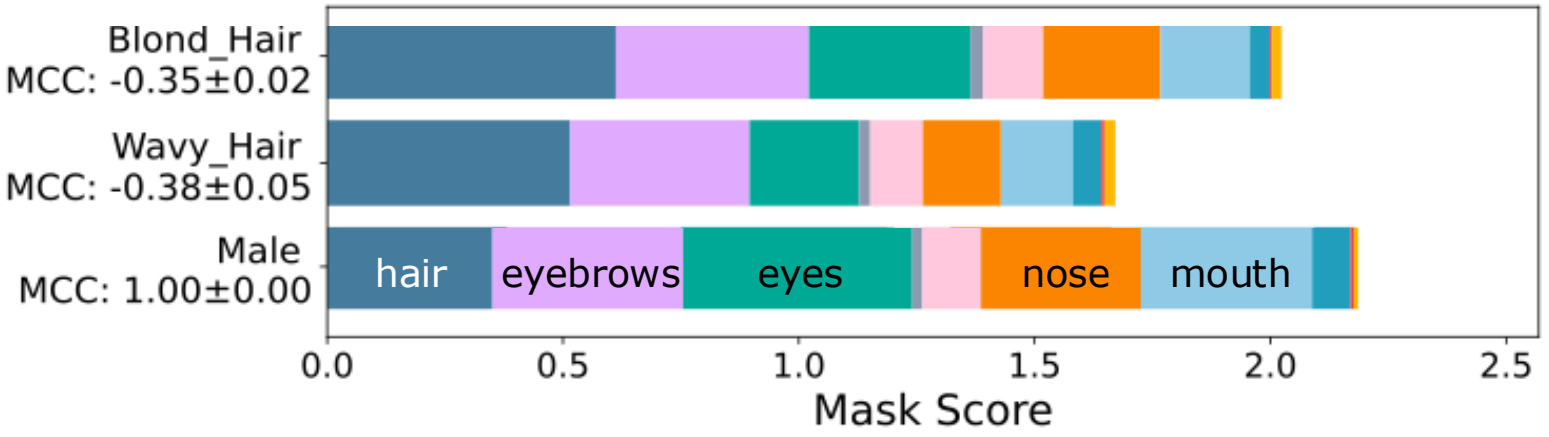
Attention IoU: Examining Biases in CelebA using Attention Maps

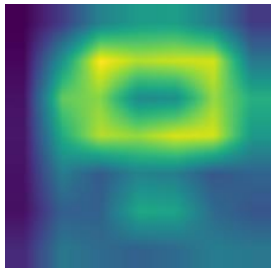
Aaron Serianni, Tyler Zhu, Olga Russakovsky, Vikram V. Ramaswamy

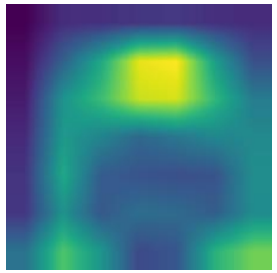
Princeton University

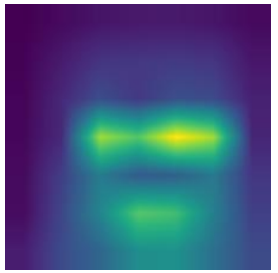
Overview

- Biases exist everywhere in computer vision
- Existing metrics focus on dataset distributions or model predictions
- Introduce Attention-IoU metric, which uses attention maps to reveal biases within a model's internal representation
- Analyze CelebA dataset, finding distinct ways bias are represented in models




Blond_Hair

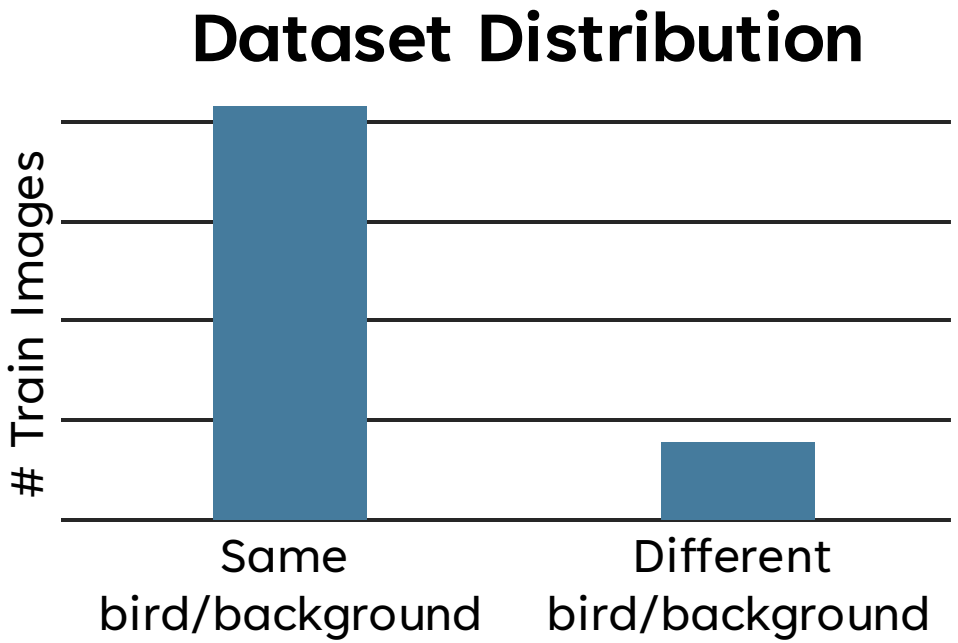

Wavy_Hair


Male


Average Image

Example: Waterbirds dataset

Landbird on water background



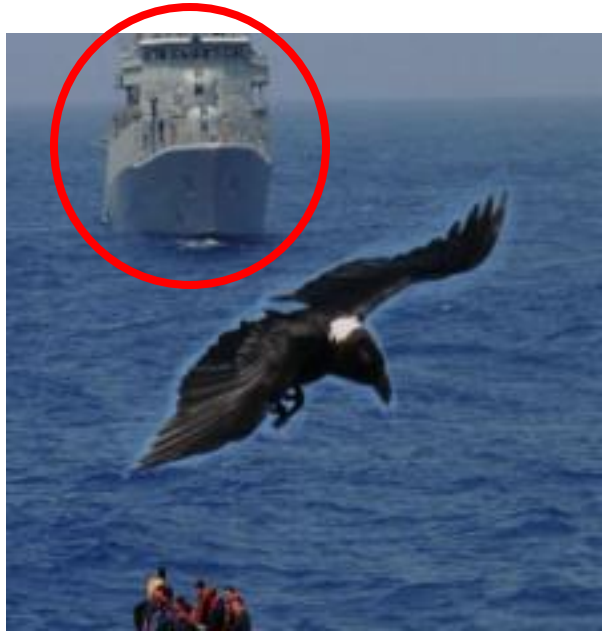
Model Output

Worst Group Accuracy (WGA) – 35%

S. Sagawa, P.W. Koh, T.B. Hashimoto, and P. Liang. Distributionally Robust Neural Networks. *ICML*, 2020.

Many ways a model can be biased

Background Bias



Object Bias

Depiction Bias

Many ways a model can be biased

Background Bias

Object Bias

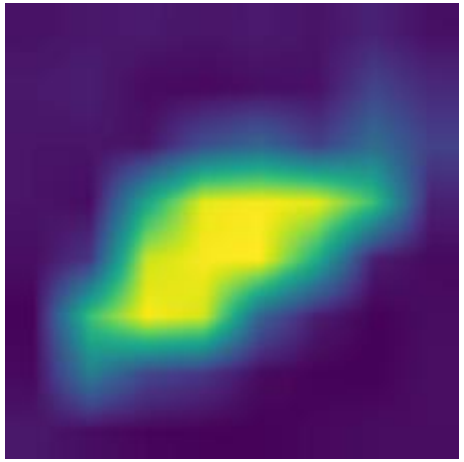
Depiction Bias

Solution: Use attention maps!

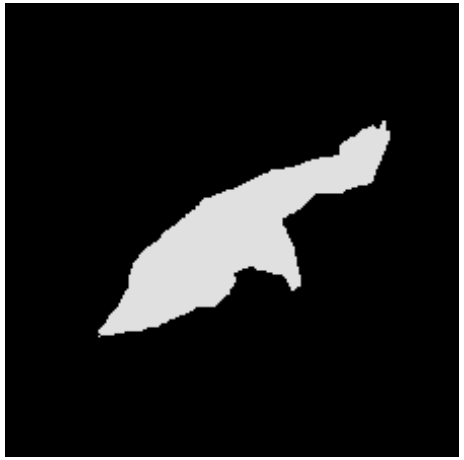
R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra.
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *IJCV*, 2020.

Goal for metric

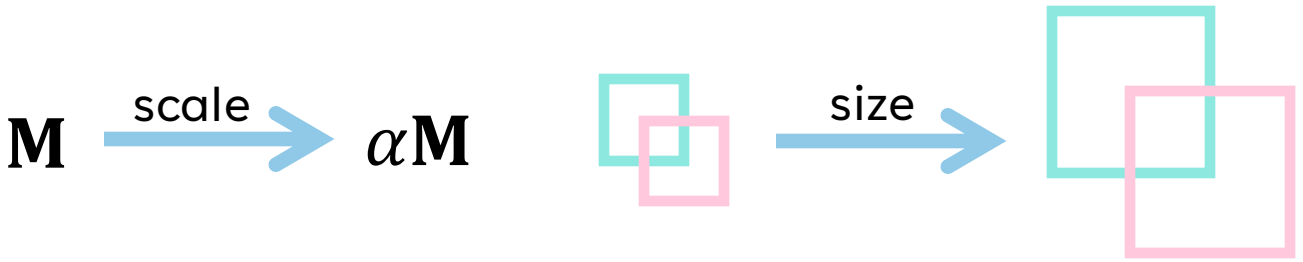
- Create a *quantifiable* metric that uses attention maps to identify biases
- By comparing attention maps with each other, or ground-truth masks, identifies where the model is attending towards
- Needs to be *scale-invariant* and *size-invariant*



Attention Map



Ground-truth Mask



Attention-IoU Metric

$$\mathcal{B}_{\text{A-IoU}}(\mathbf{M}_1, \mathbf{M}_2) = \frac{\langle \widehat{\mathbf{M}}_1, \widehat{\mathbf{M}}_2 \rangle_F}{\left\| \frac{\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}_2}{2} \right\|_F^2} = \frac{\sum_{i,j} (\widehat{\mathbf{M}}_1)_{ij} \cdot (\widehat{\mathbf{M}}_2)_{ij}}{\sum_{ij} \left(\frac{\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}_2}{2} \right)_{ij}^2}$$

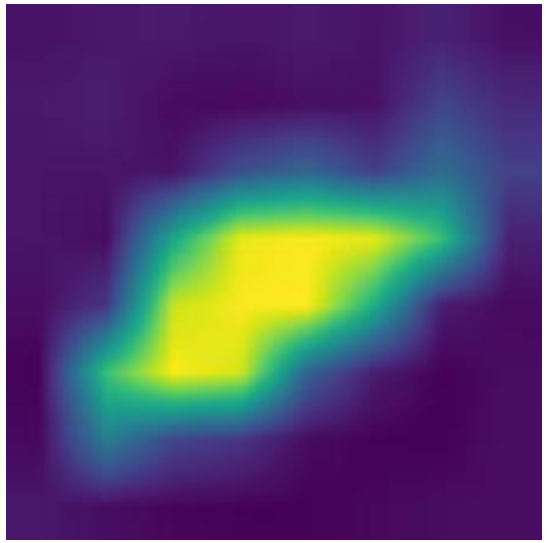
$\mathbf{M_1}, \mathbf{M_2}$ — attention maps/feature mask

$\widehat{\mathbf{M}}_i = \frac{\mathbf{M}_i}{||\mathbf{M}_i||_1}$ — normalized map

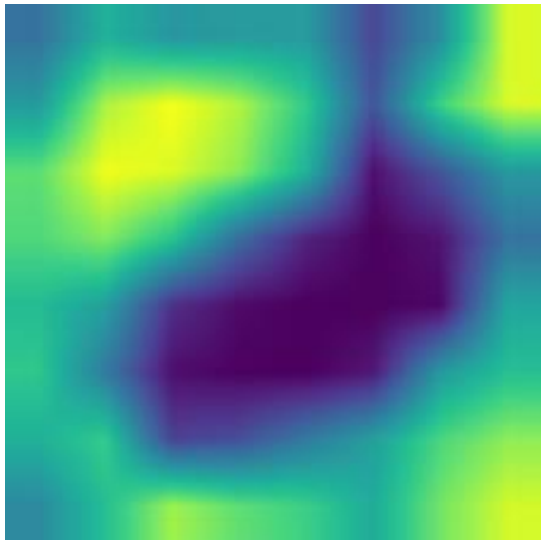
$\langle \cdot, \cdot \rangle_F$, $||\cdot||_F$ — Frobenius norm

Heatmap and Mask Scores

$$\text{Attention-IoU}_{\text{Heatmap}}(t,p) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_{\text{A-IoU}}(\text{GradCAM}_t(\mathbf{x}_i), \text{GradCAM}_p(\mathbf{x}_i)).$$



$\text{GradCAM}_{\text{bird}}(\mathbf{x}_i)$

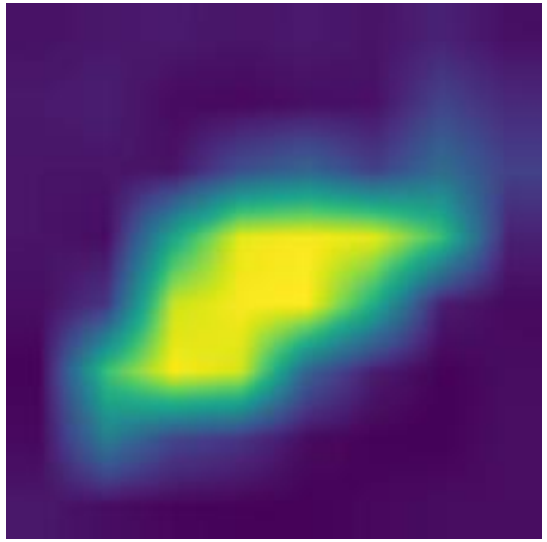


$\text{GradCAM}_{\text{background}}(\mathbf{x}_i)$

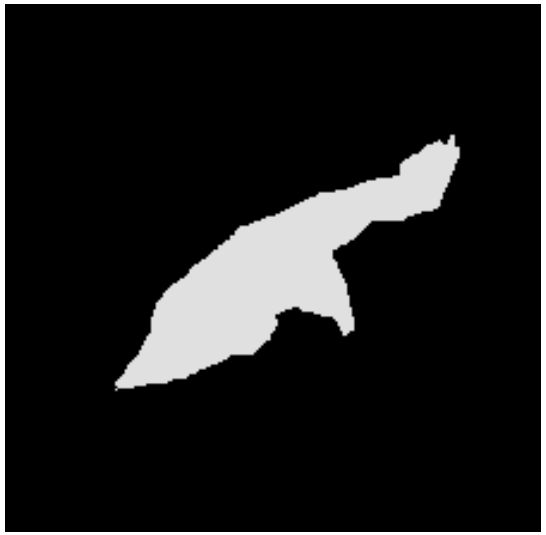
Heatmap and Mask Scores

$$\text{Attention-IoU}_{\text{Heatmap}}(t,p) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_{\text{A-IoU}}(\text{GradCAM}_t(\mathbf{x}_i), \text{GradCAM}_p(\mathbf{x}_i)).$$

$$\text{Attention-IoU}_{\text{Mask}}(t,f) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_{\text{A-IoU}}(\text{GradCAM}_t(\mathbf{x}_i), \text{interp}(\text{mask}_f(\mathbf{x}_i))).$$



$\text{GradCAM}_{\text{bird}}(\mathbf{x}_i)$



$\text{mask}_{\text{bird}}(\mathbf{x}_i)$

CelebA Dataset

- Annotated dataset of celebrity faces
- Widely used for facial recognition, image generation, and fairness

40 Facial Attributes

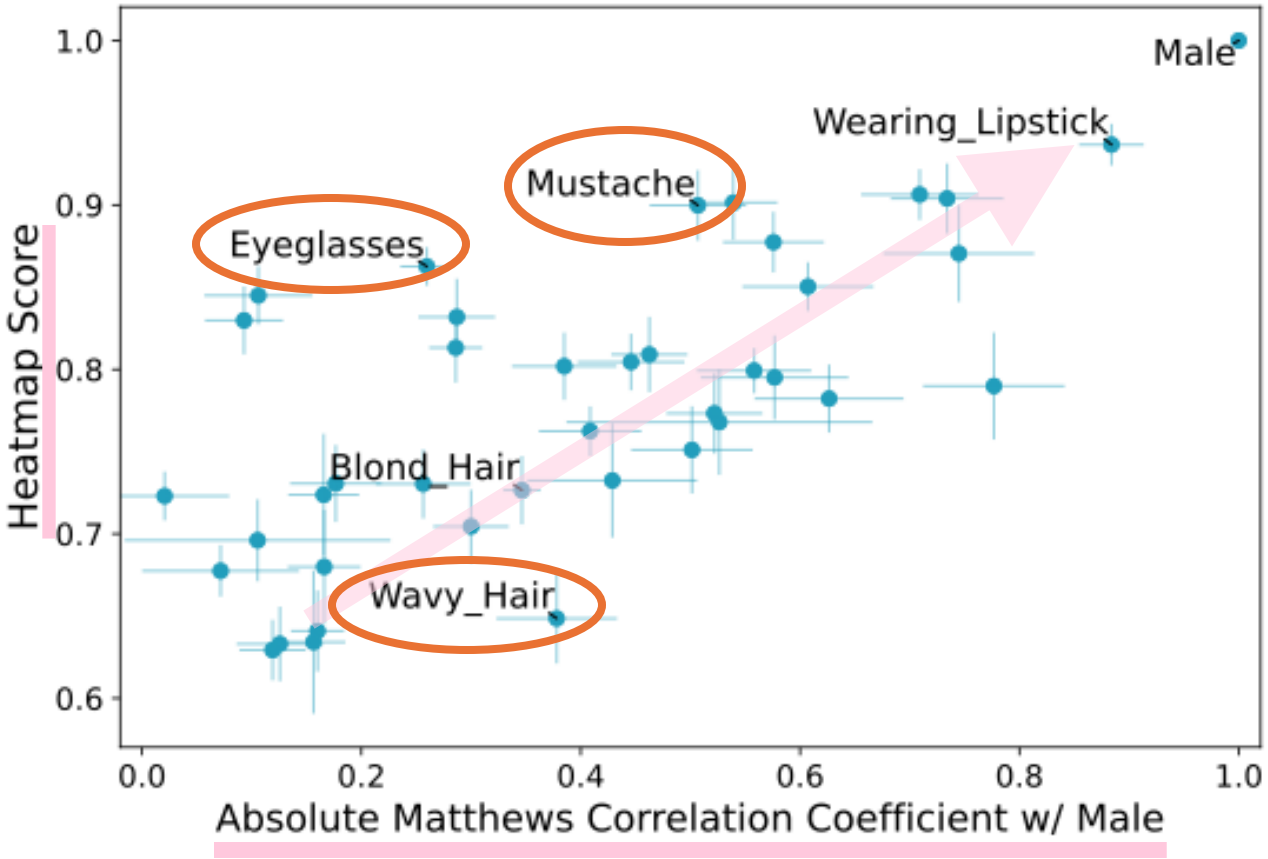
Eyeglasses, Smiling
Black_Hair, not **Male**

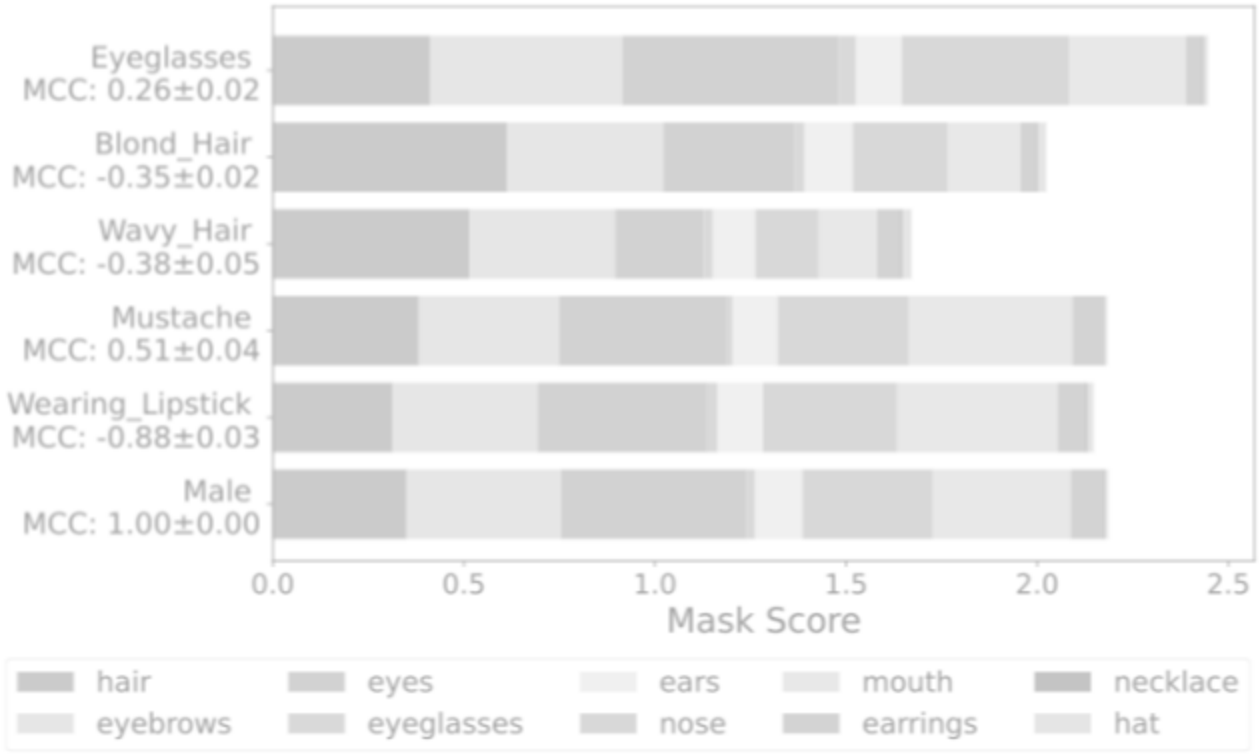
19 Segmentation Masks

eyes, eyebrows
mouth, hair, nose

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. *ICCV*, 2015.

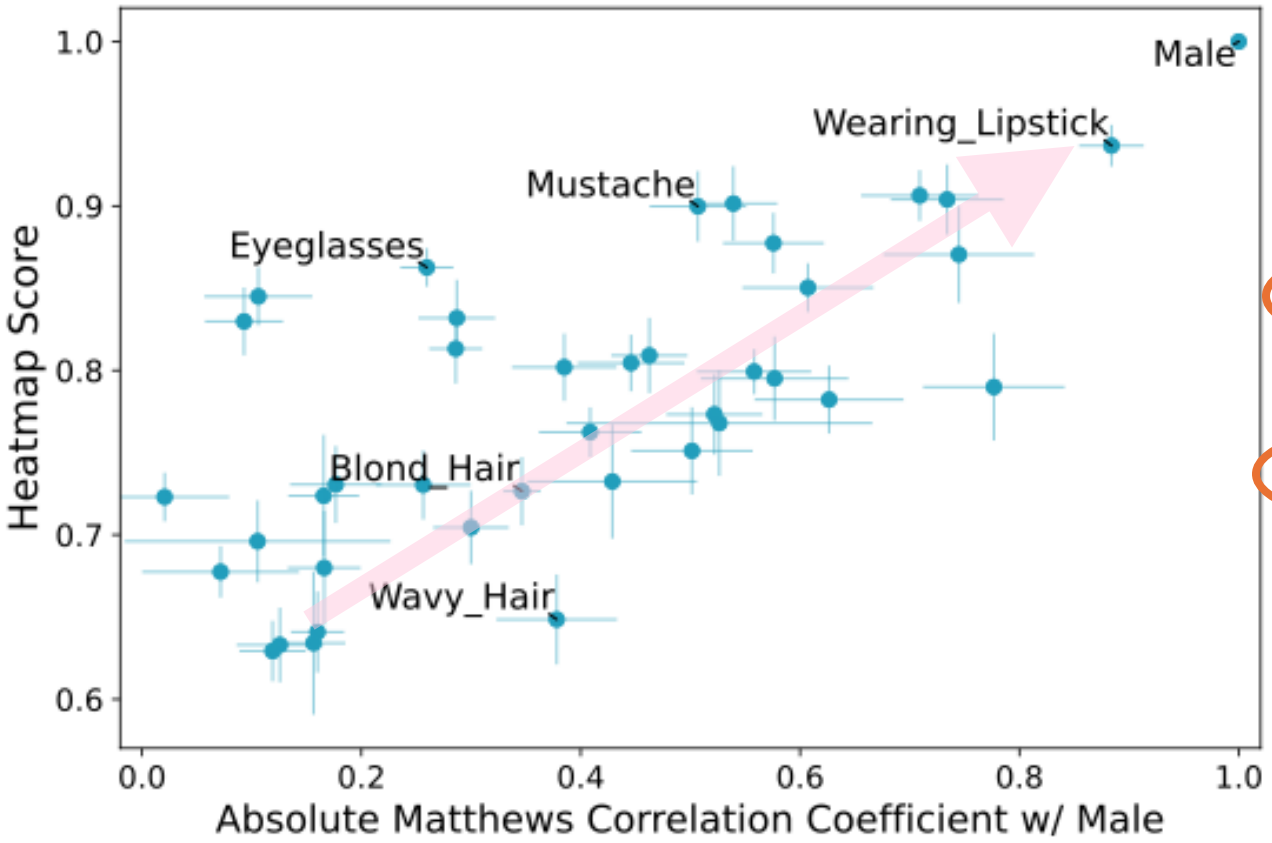
Comparison with Male Heatmap

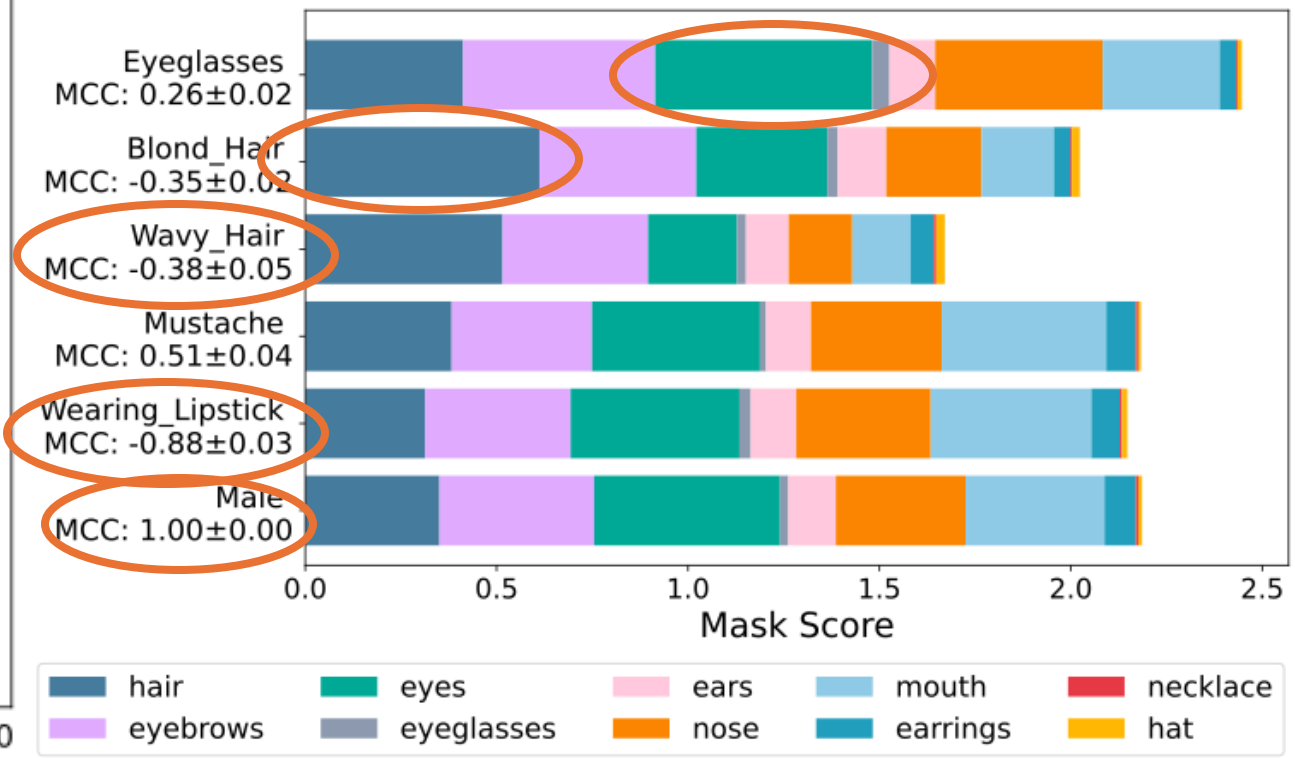




Trend between heatmap score and dataset labels, with outliers

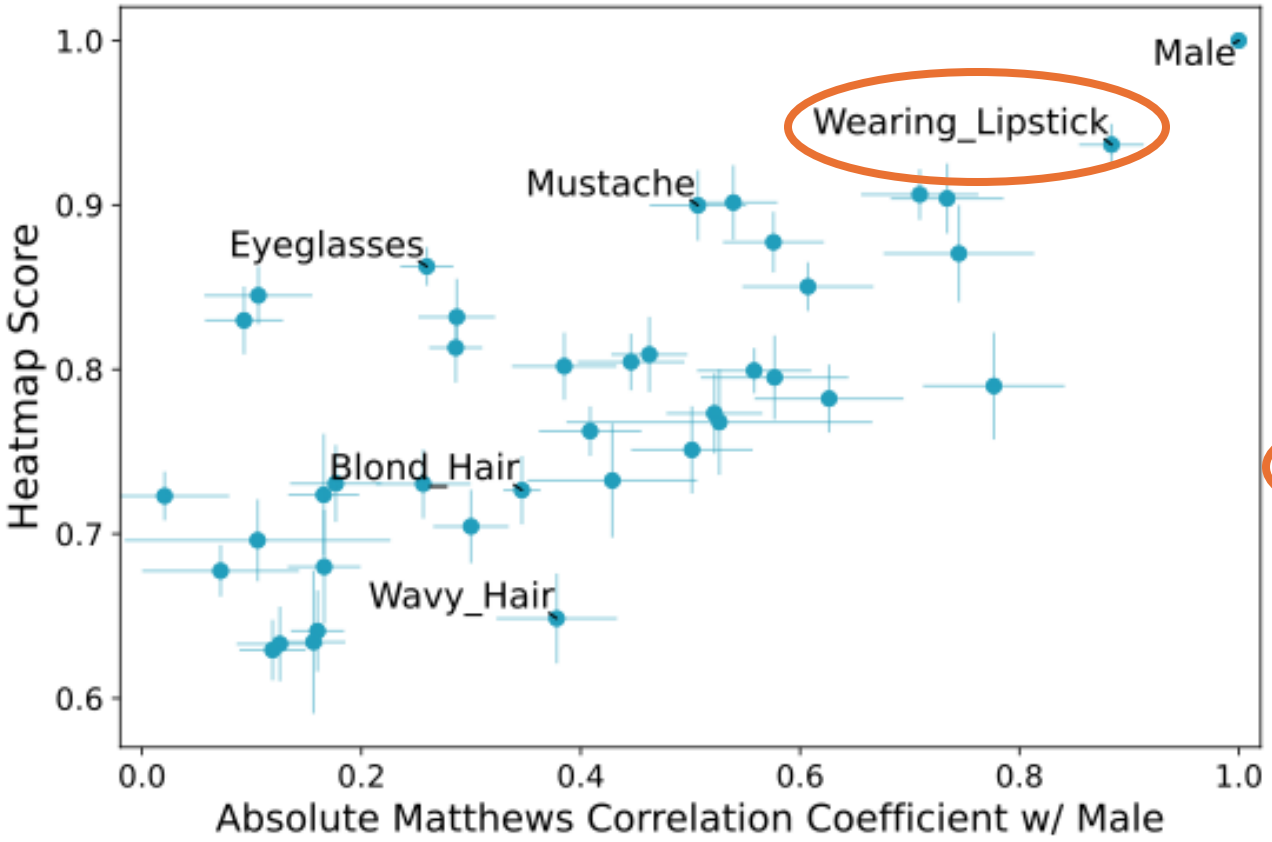
Comparison with Male Heatmap

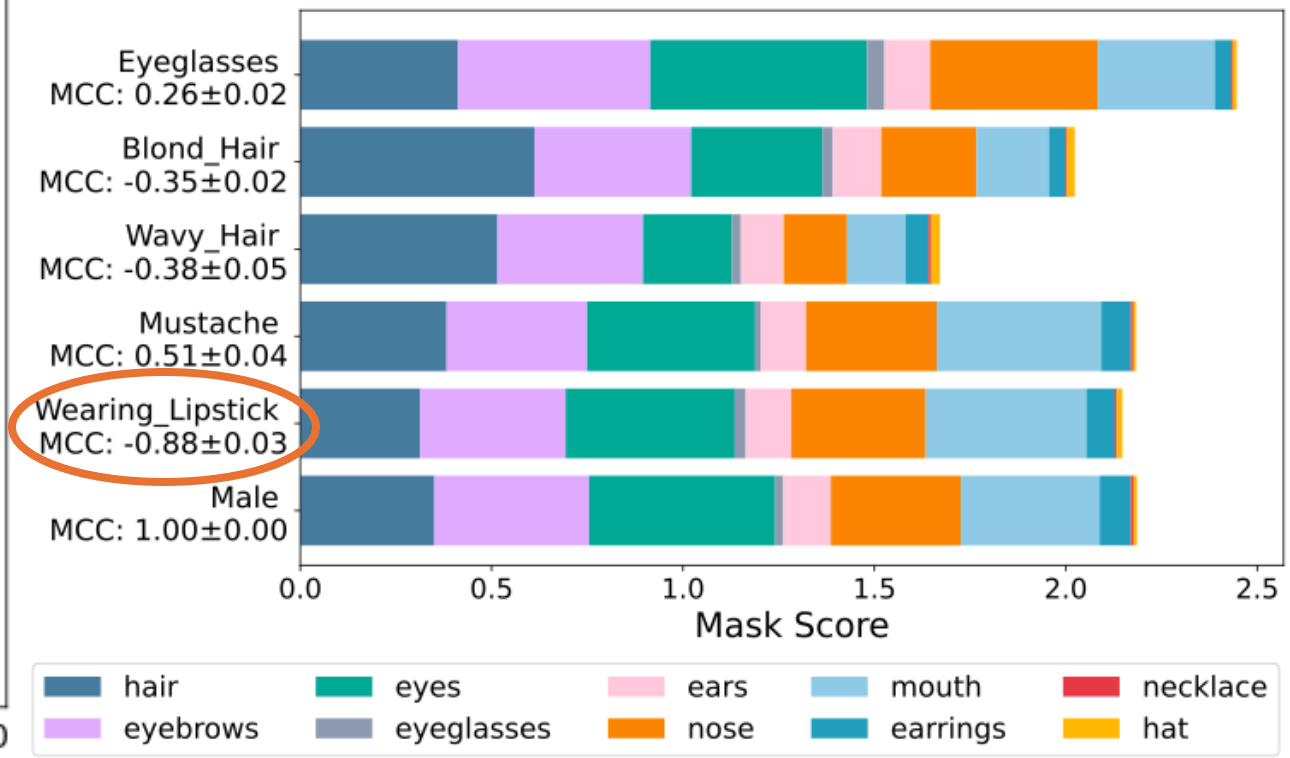




Trend between heatmap score and dataset labels, with outliers
Correlation is reflected in mask score

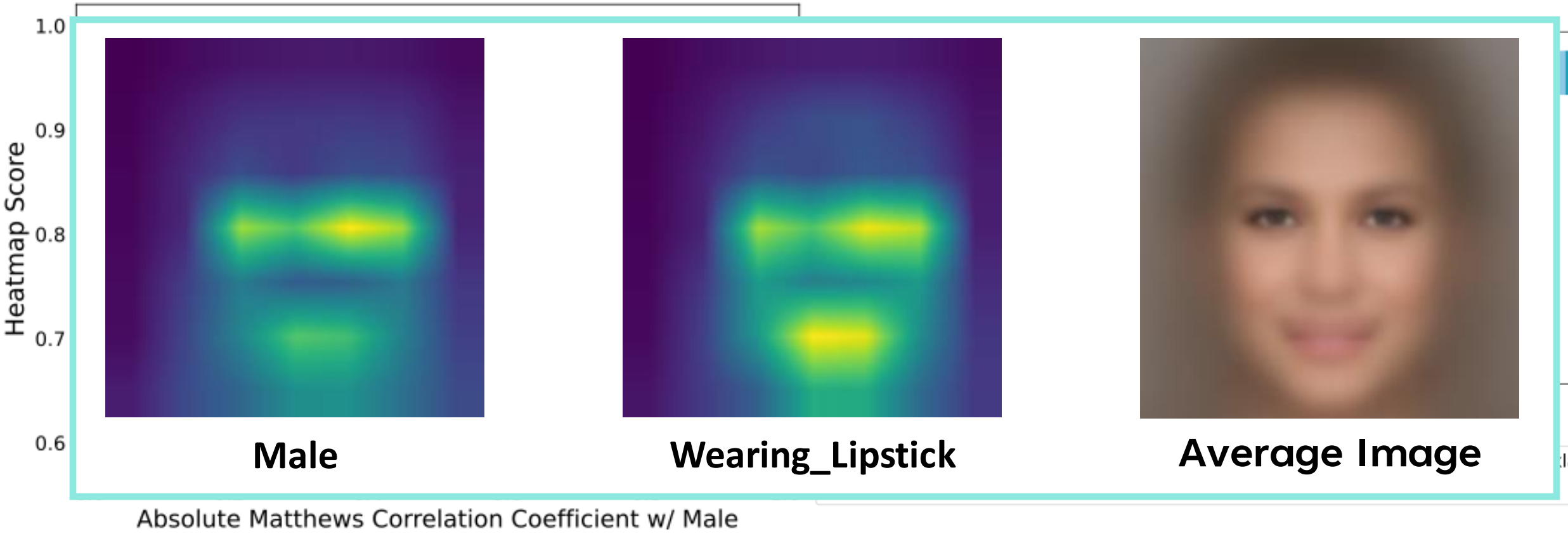
Wearing_Lipstick





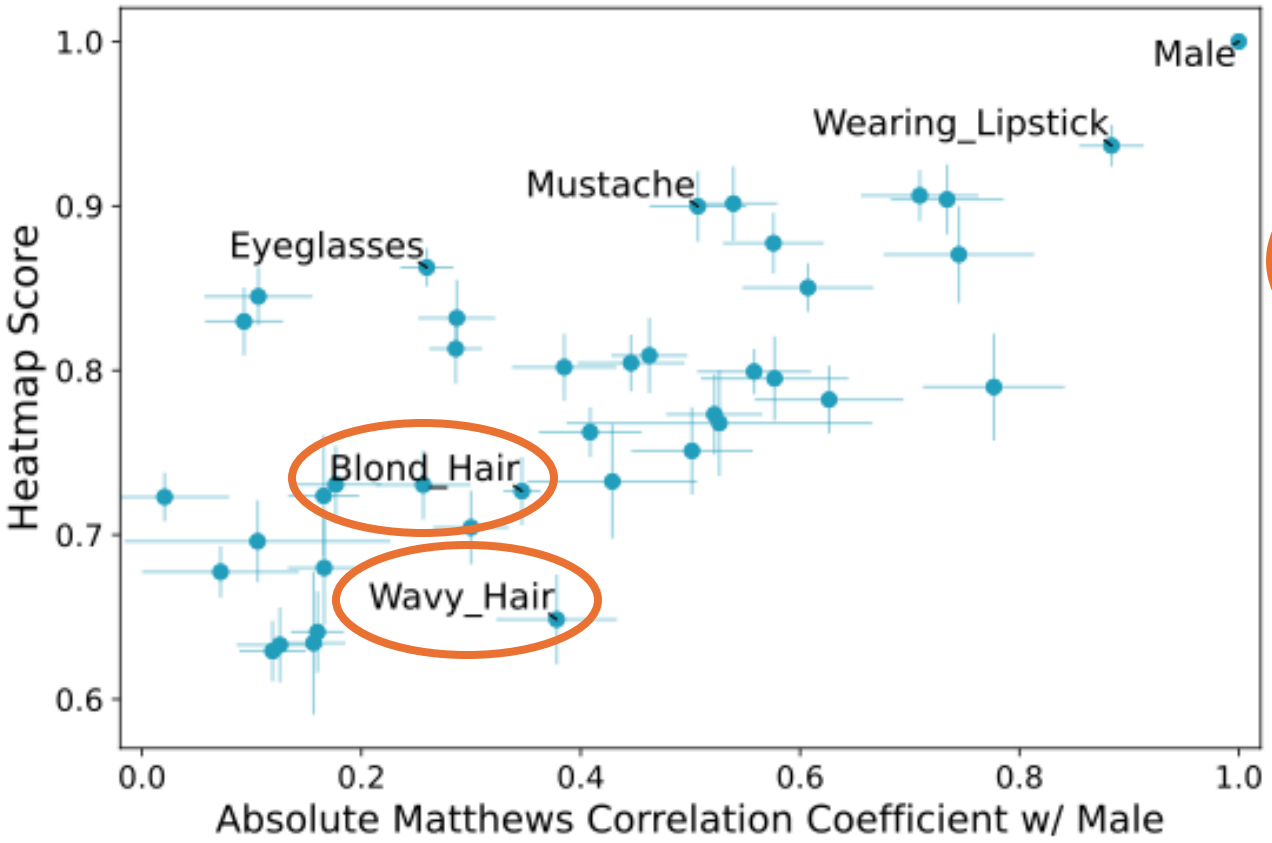
Highest correlation with Male
Reflected by heatmap and mask scores

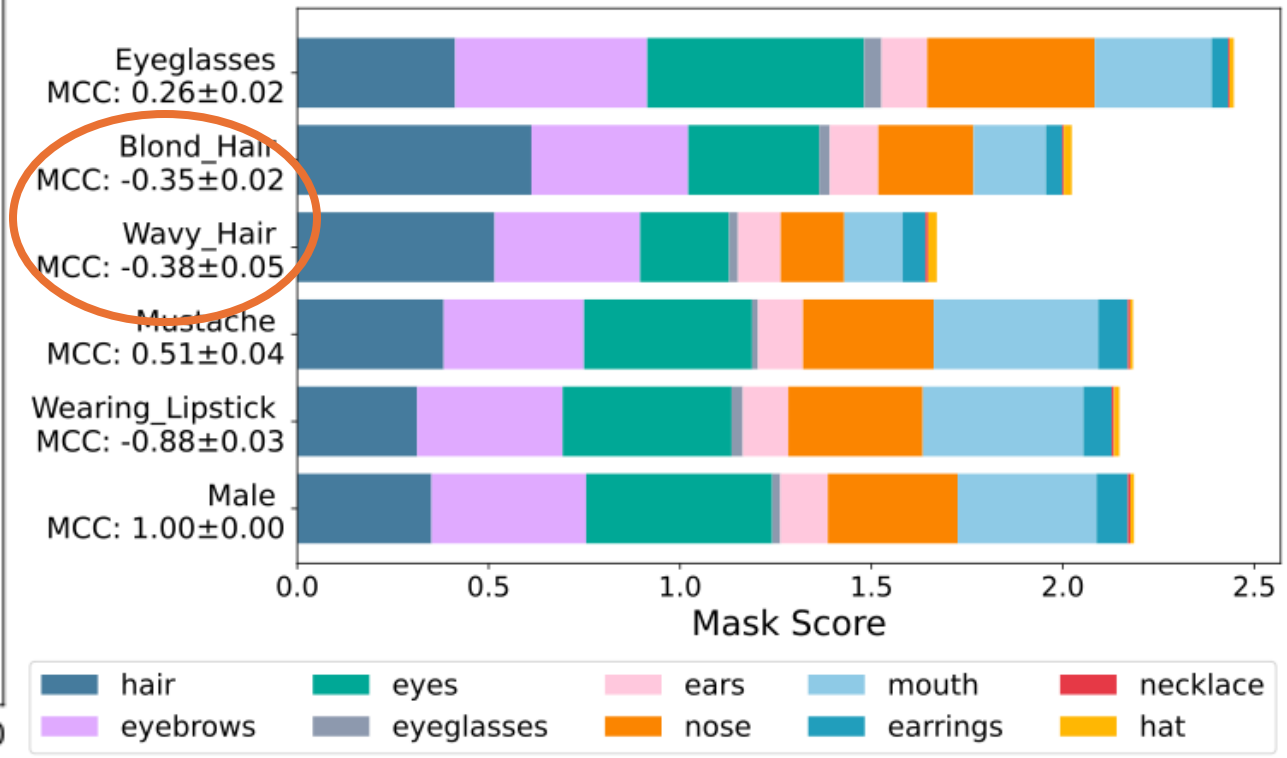
Wearing_Lipstick



Highest correlation with Male
Reflected by heatmap and mask scores

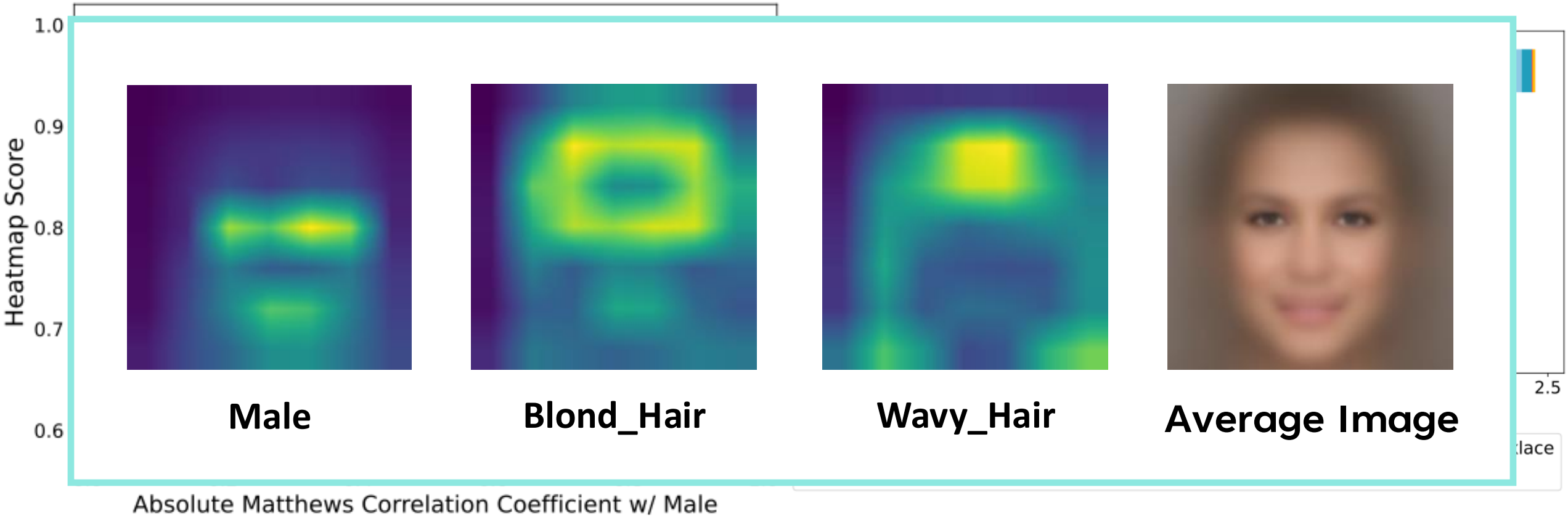
Blond_Hair and Wavy_Hair





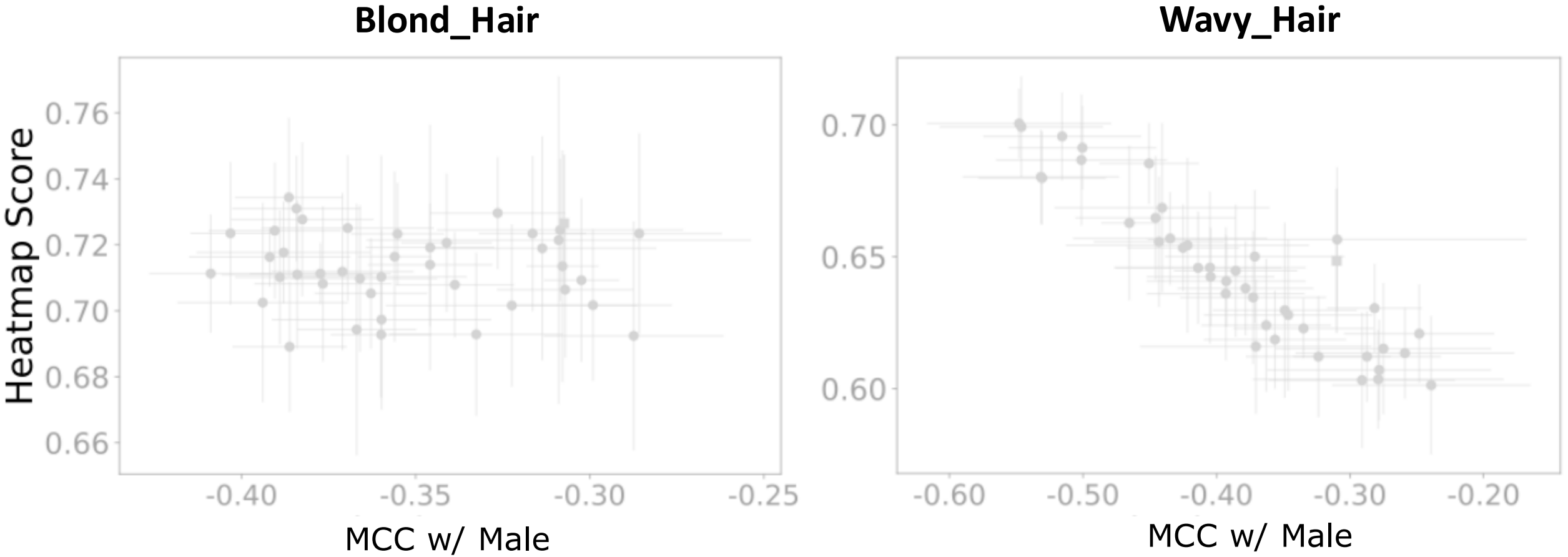
Two related attributes with similar correlations in model output, but different heatmap and mask scores

Blond_Hair and Wavy_Hair



Two related attributes with similar label correlations, but different heatmap and mask scores

Varying correlations in the training dataset



No change in correlation for Blond_Hair suggests an unlabeled confounder distinct from Male

Takeaways

- Attention-IoU can effectively measure many forms of bias in image classifiers using attention maps
- Identify specific ways in which attributes are biased in CelebA
- Can guide creation of better models and debiasing methods

Thank you!

Code and paper available at
https://github.com/aaronserianni/attention-iou

Email: serianni@princeton.edu