

Automated Detection of Visual Attribute Reliance with a Self-Reflective Agent



Christy Li, Josep Lopez Camuñas, Jake Touchet, Jacob Andreas, Agata Lapedriza, Antonio Torralba, Tamar Rott Shoham

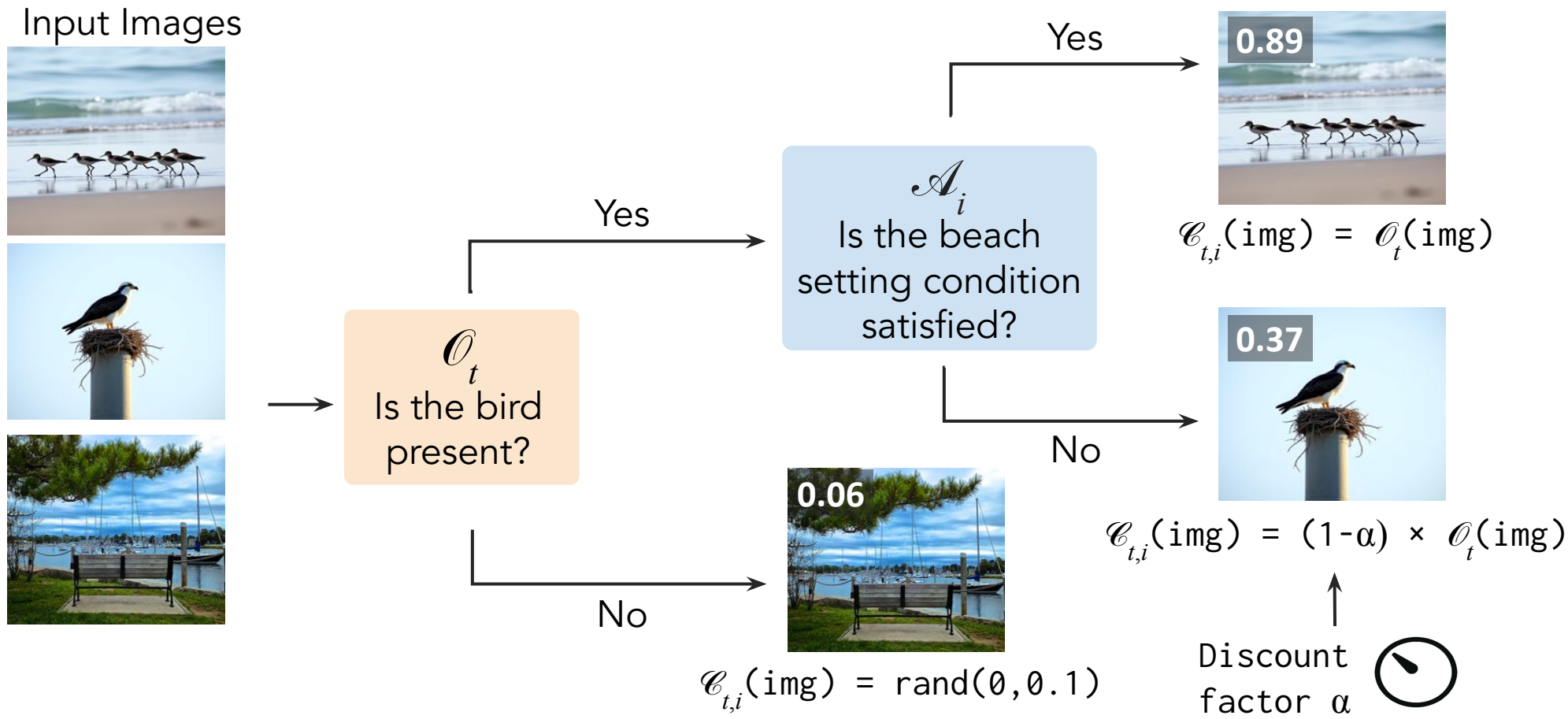


Motivation

- A well-known failure mode of vision models performing image recognition is over-reliance on **spurious visual attributes**, resulting in brittle predictions that fail to generalize.
- Existing methods for discovering the visual attributes behind model predictions often rely on manual inspection along a fixed set of predefined features, making them difficult to scale.
- We propose a **Self-reflective Automated Interpretability Agent (SAIA)** that produces natural-language descriptions of spurious attribute dependencies of visual models.

Attribute Reliance Detection Benchmark

Bird detector that relies on the presence of beach background



Taxonomy of Visual Attribute Reliances $\mathcal{A}_i$

Object attributes: color, material
Context attributes: image background, object state/configuration
Demographic attributes: gender, age of human subjects
Counterfactual demographic attributes: gender, age of human subjects

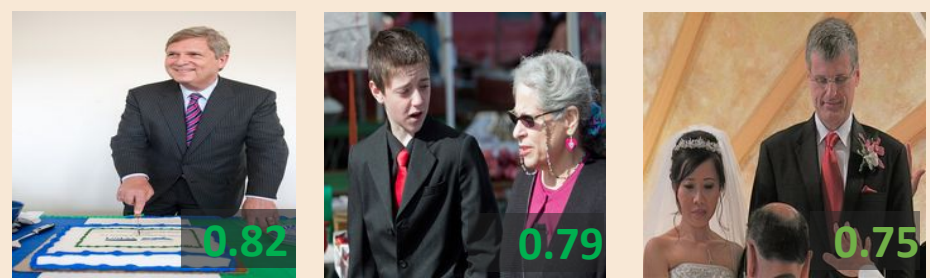
130 unique object recognition models, 18 types of visual attribute reliance

SAIA Framework

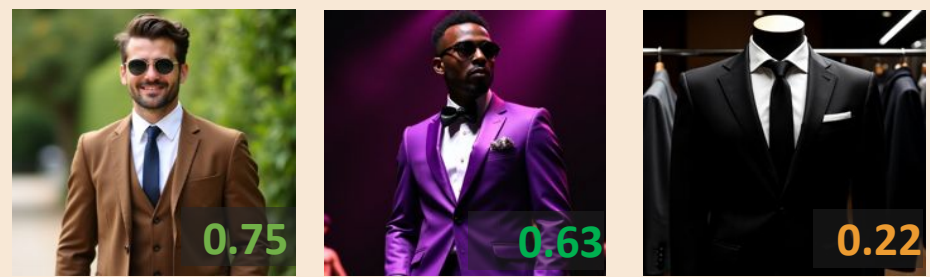
What are the visual features my model relies on to detect suit?

Hypothesis-Testing

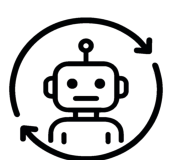
> Looking at images with highest classification score:



> Test: relies on black suits



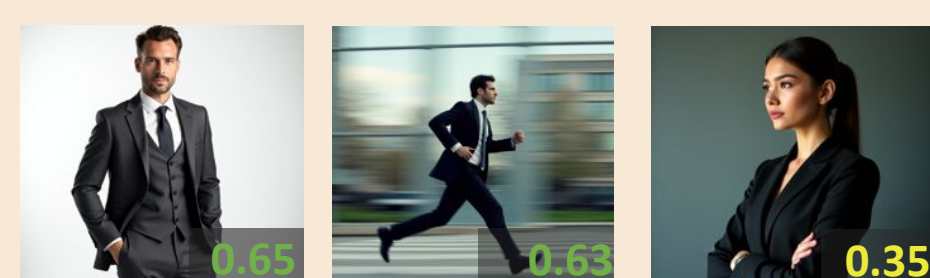
> Test: corporate settings



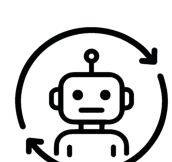
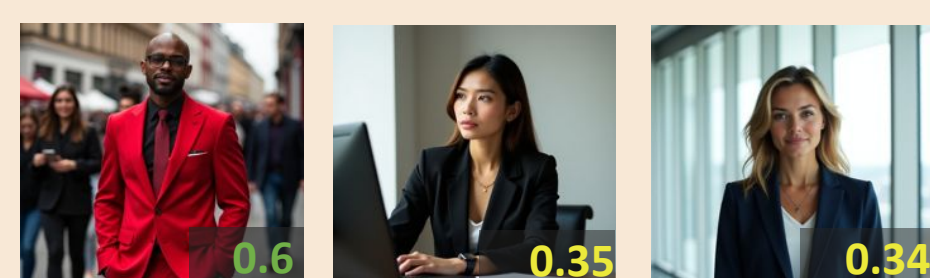
The model relies on corporate settings.

> Looking at images from self-evaluation and generating new hypotheses.

> Test: 45-degree views of suits



> Test: male vs. female

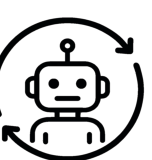
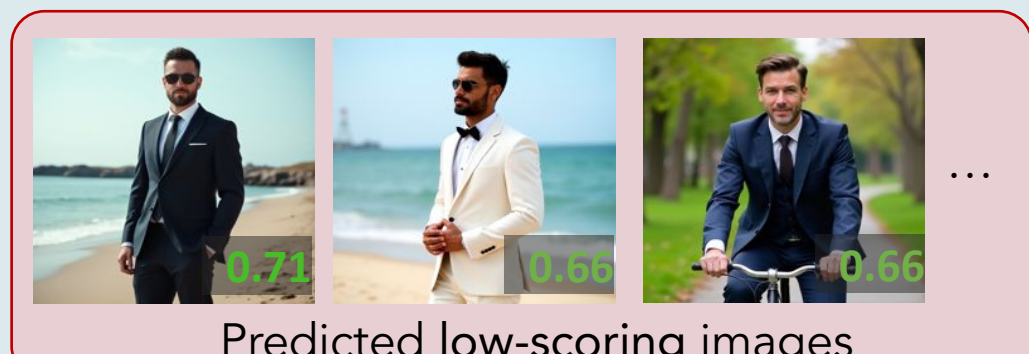
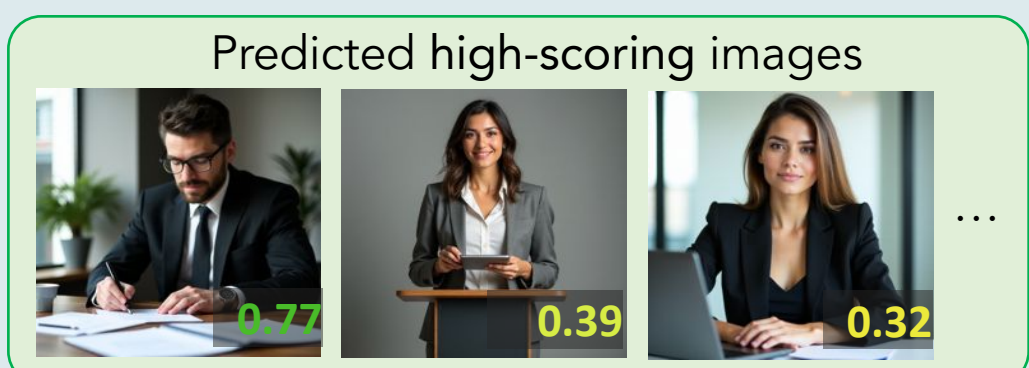


The model exhibits a strong gender bias, relying on a male wearing the suit.

Self-Reflection

> Let's evaluate the "corporate settings" conclusion. Let's look at images with and without this attribute and evaluate if they match model responses.

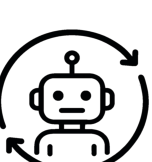
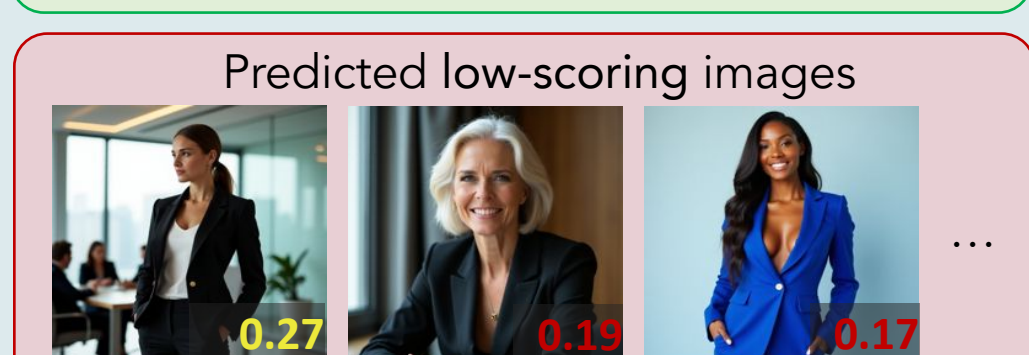
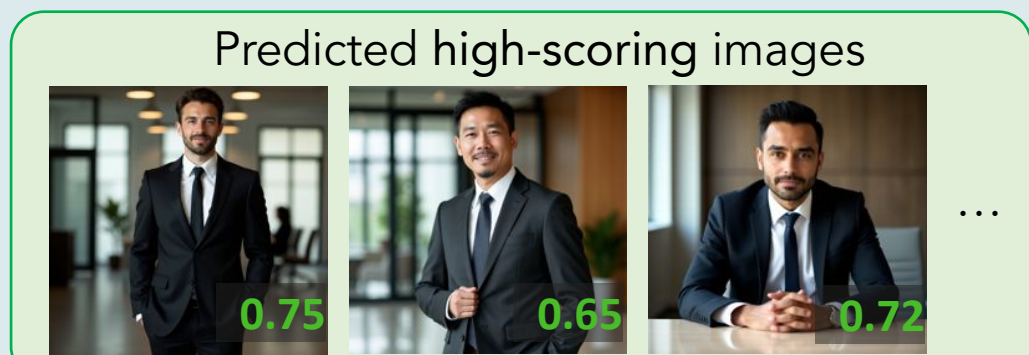
Self-Evaluation



Some images with corporate settings got low confidence scores, while some without corporate settings got high scores. Let's run another experiment.

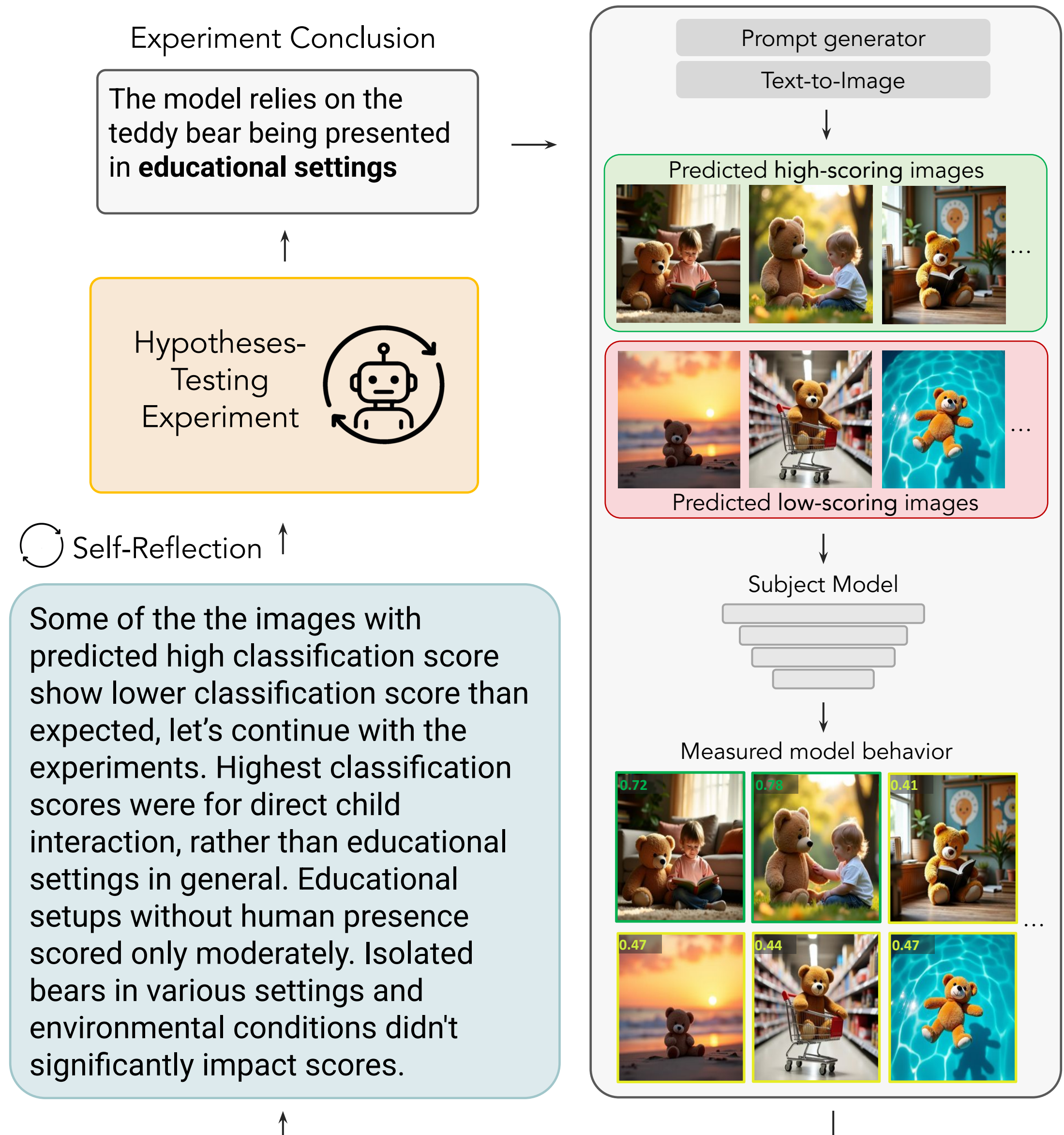
Self-Evaluation

> Let's evaluate the "gender bias" conclusion.



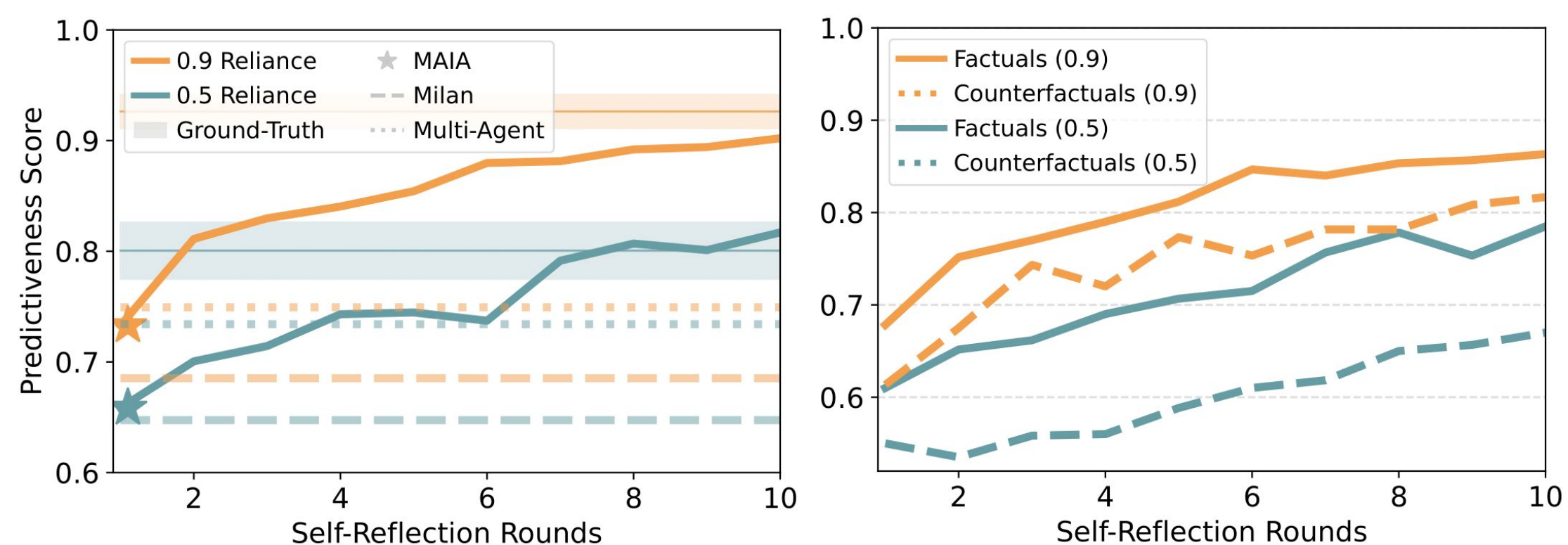
The conclusion matches the model behavior: The model relies on male presenting figures to detect the suit.

Self-Reflection



Evaluation

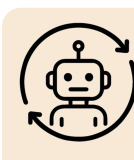
How does predictiveness score change with increasing rounds of self-reflection?



Applications to SOTA Models

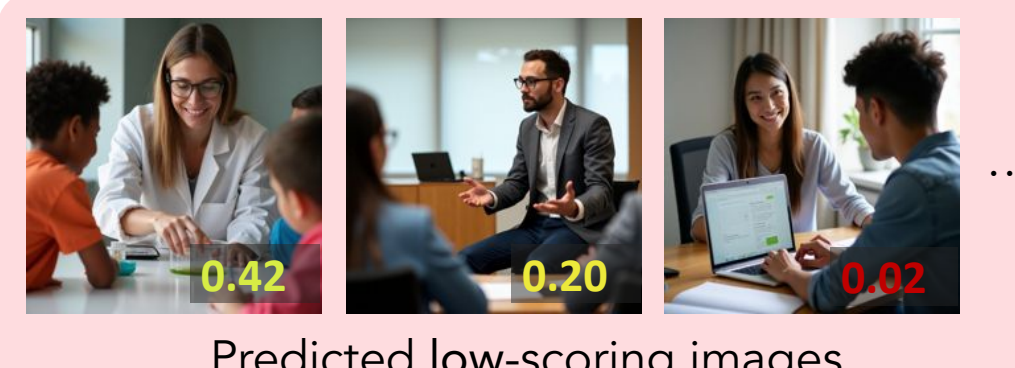
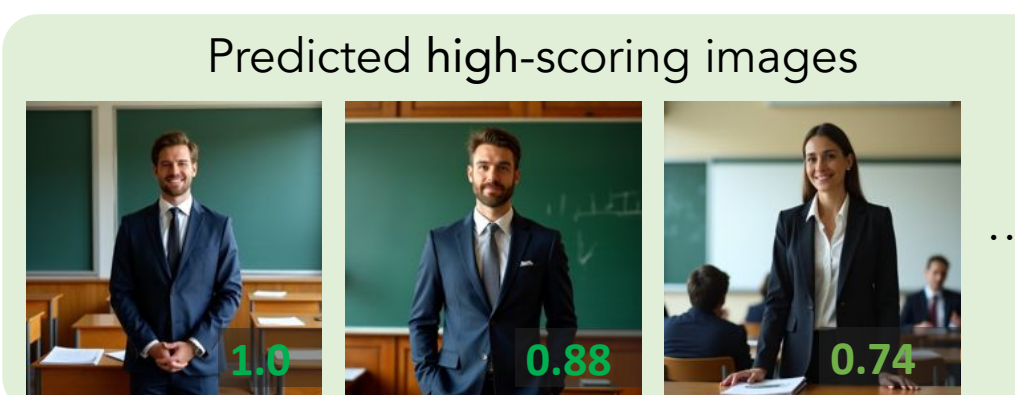


What are the visual attributes that CLIP relies on to recognize the concept teacher?

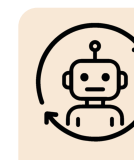


Let's conduct some experiments...

CLIP relies on authority signals in conventional classroom settings rather than actual pedagogical engagement.

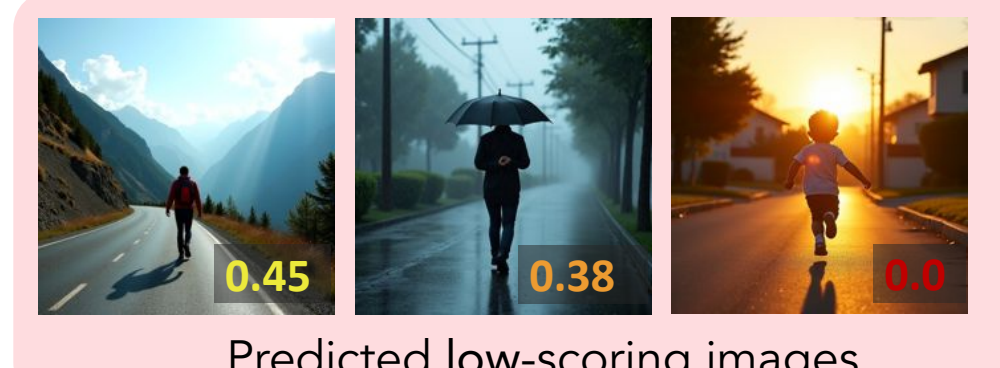
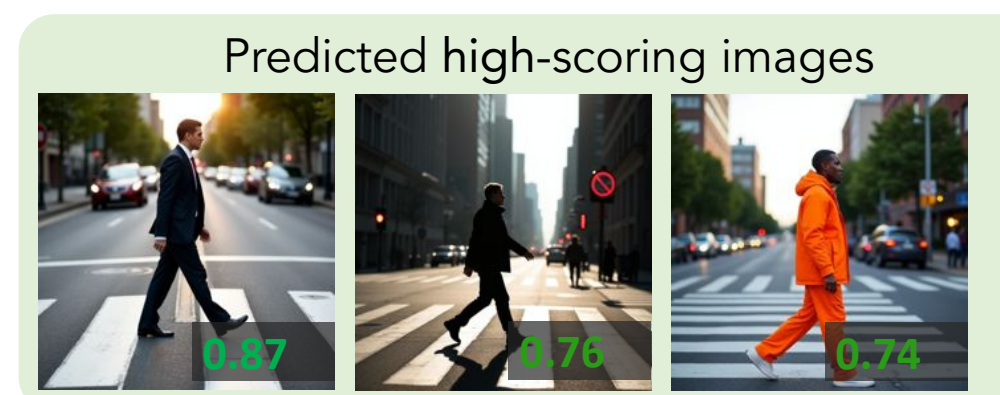


What are the visual attributes that YOLO relies on to detect pedestrians?



Let's conduct some experiments...

YOLO relies on side-view walking poses in urban crossing contexts.



We demonstrate that SAIA can identify feature reliances in state-of-the-art models including **CLIP's** vision encoder and the **YOLOv8** object detector.