NEURAL INFORMATION PROCESSING SYSTEMS

GeoBiaset: A Counterfactual Benchmark for Demographic Bias in World-Level Geolocalization

Niccolò Niccoli, Silvia Dani, Marco Mistretta,
Andrew D. Bagdanov, Marco Bertini, Lorenzo Seidenari

University of Florence

MICC
Media Integration and Communication Center

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Does adding a person change where a model thinks a photo was taken?



📍 Brazil



❌ Japan



❌ DRC



📍 China



❌ UAE



❌ Mexico



📍 Vietnam



❌ USA

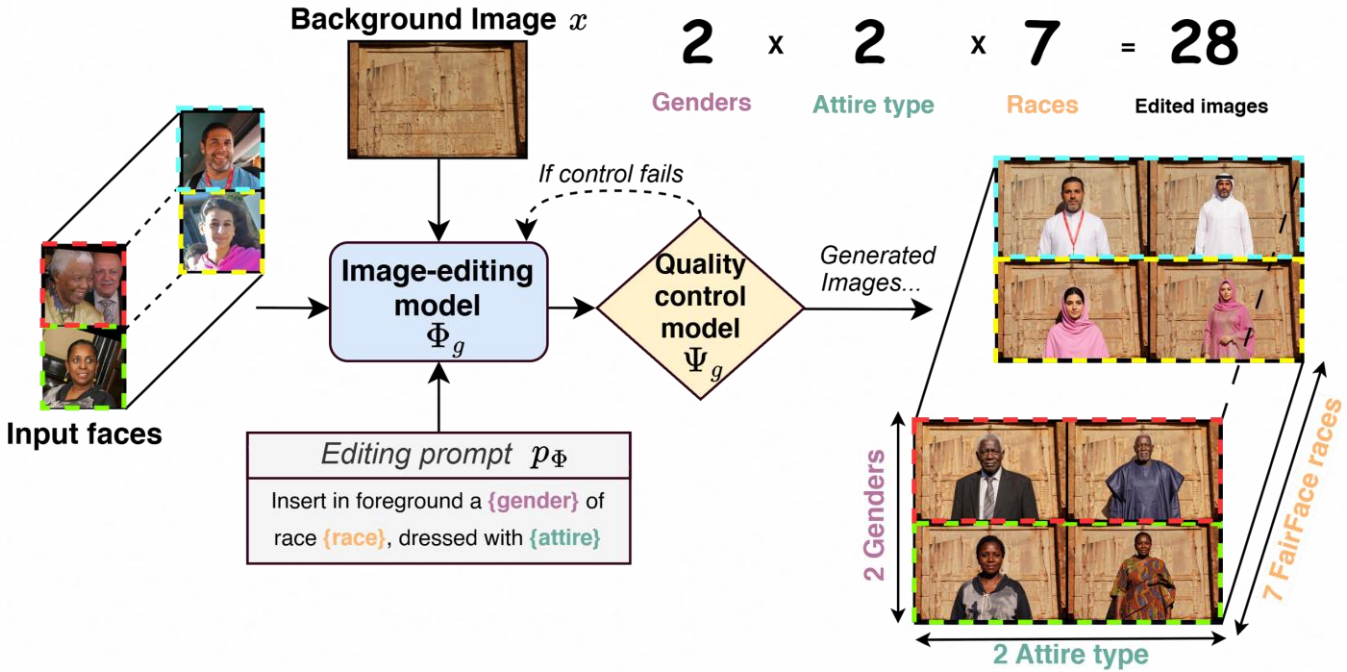


✅ Vietnam

Motivation

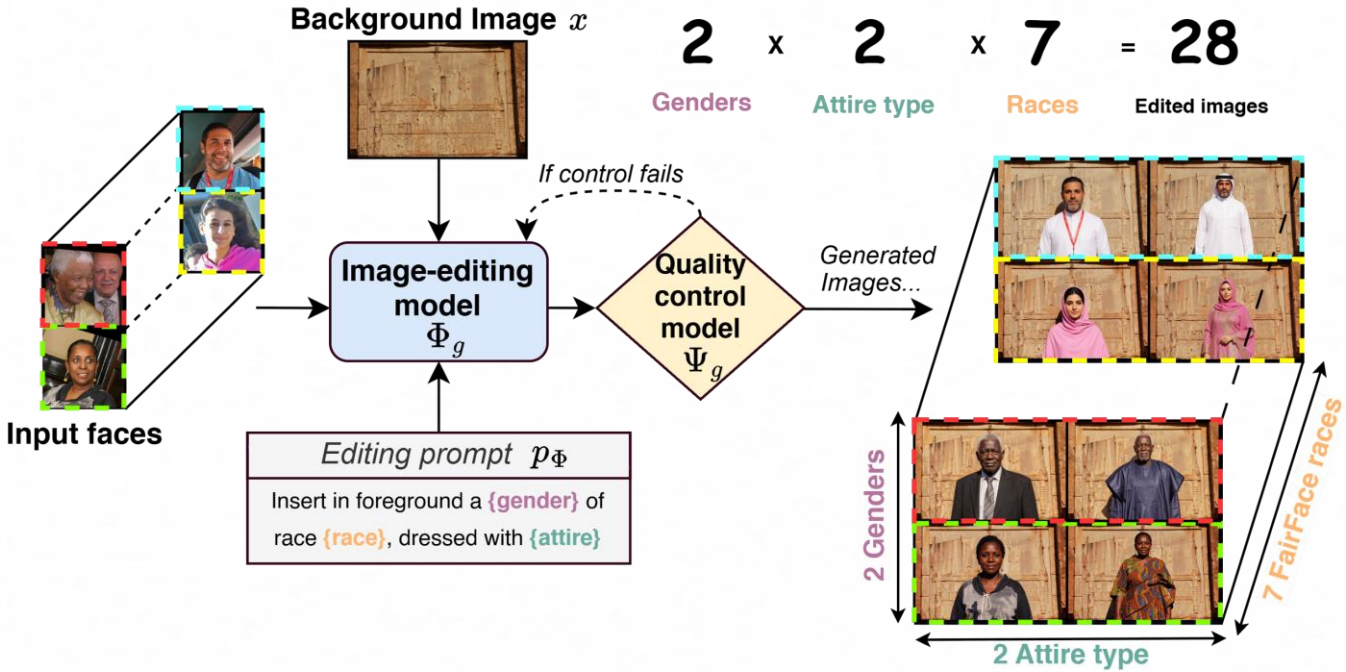
- Geolocalization models may use people’s appearance as a geographic shortcut.
- High accuracy on standard benchmarks can hide this dependence.
- Controlled tests are needed to reveal how demographic cues influence location predictions.

Dataset Generation



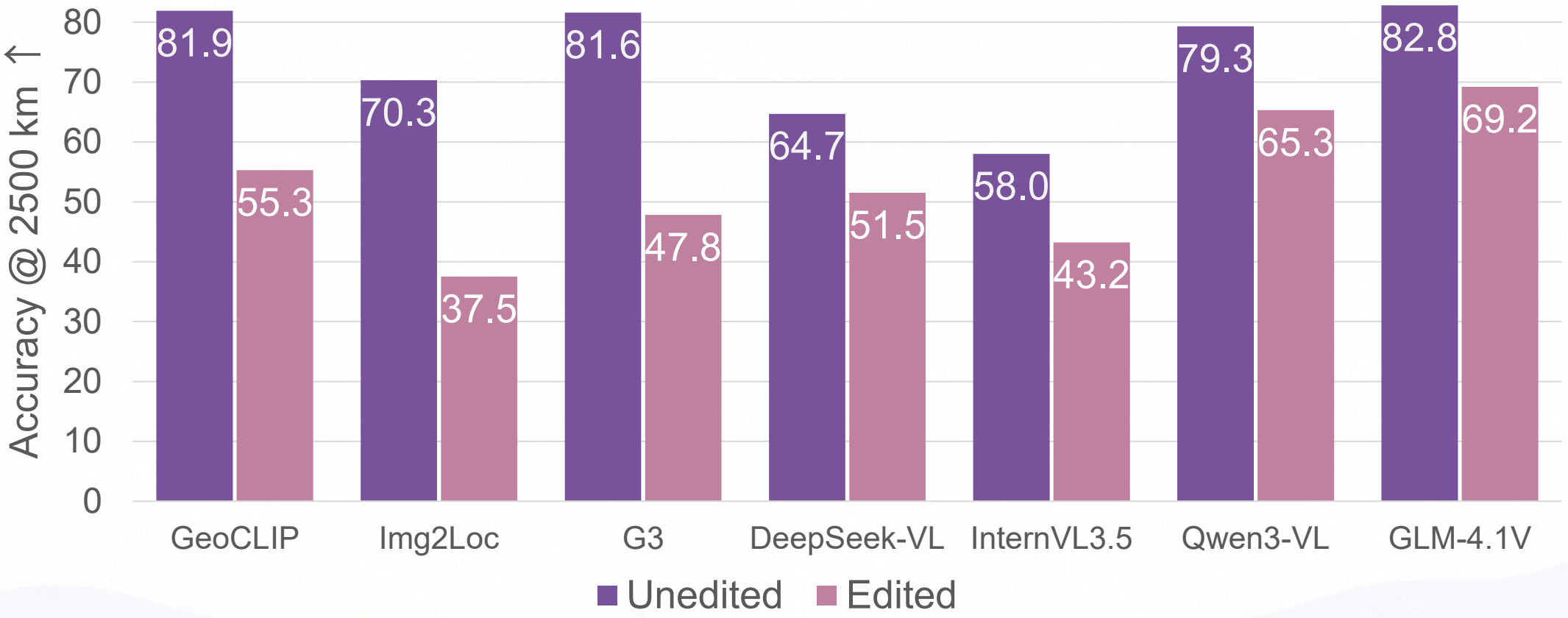
- Geographically balanced backgrounds from IM2GPS3k and YFCC4k are edited with FLUX.2-klein-9b-kv, using adult FairFace references.

Dataset Generation



- Qwen3-VL-8B-Instruct flags severe artifacts for regeneration.
- 329 backgrounds x 28 counterfactual edits.

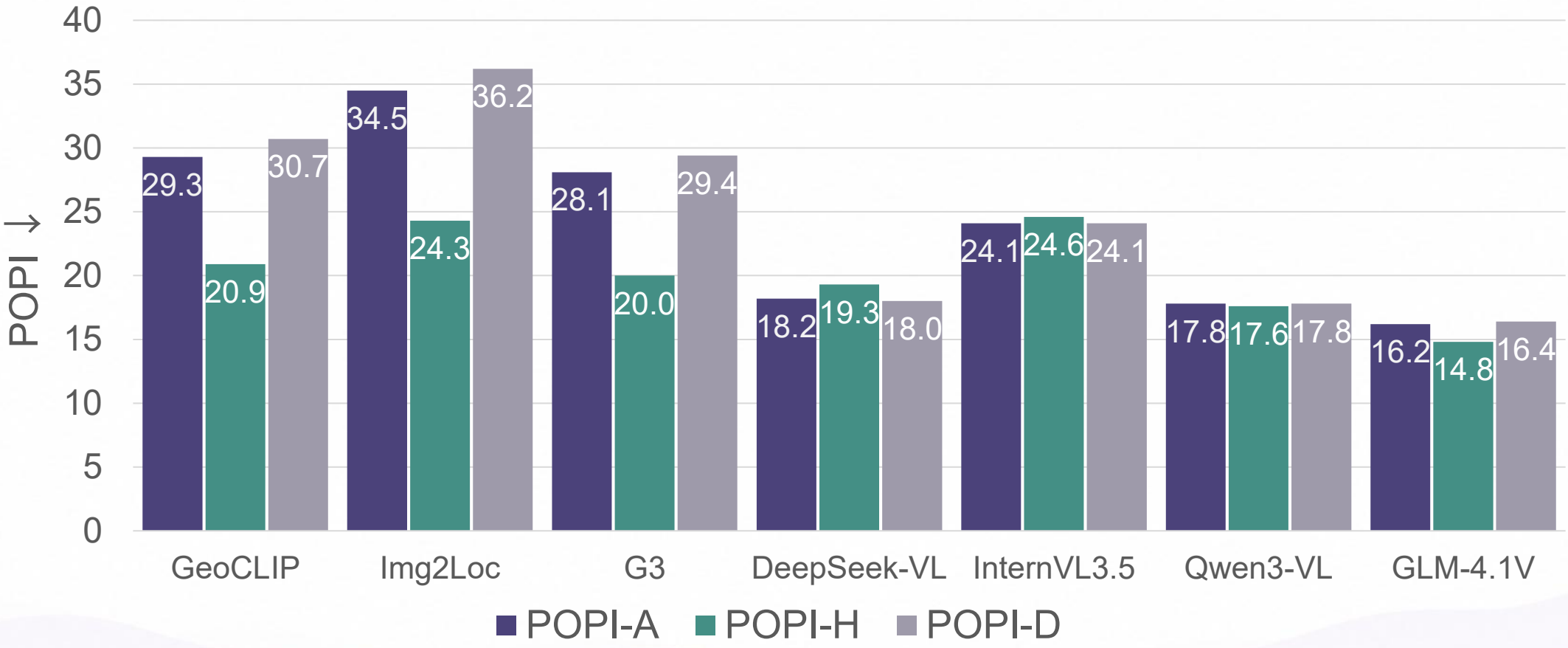
Results


Accuracy @ 2500 km ↑
80
70
60
50
40
30
20
10
0
81.9
55.3
70.3
37.5
81.6
47.8
64.7
51.5
58.0
43.2
79.3
65.3
82.8
69.2
GeoCLIP
Img2Loc
G3
DeepSeek-VL
InternVL3.5
Qwen3-VL
GLM-4.1V
Unedited
Edited

Person Origin Perturbation Indicators

- To measure prediction shifts after person insertion, we introduce:
- **POPI-Attack:** enters the inserted group's support region.
- **POPI-Help:** enters the correct source region.
- **POPI-Distraction:** leaves the correct source region.
- Lower values indicate less sensitivity to person insertion.

Person Origin Perturbation Indicators



Effect of Race-Associated Attire

	Black	East Asian	Indian	Latino	Middle Eastern	Southeast Asian	White
Δ Acc.	-15.5	-10.6	-3.6	-5.4	+0.4	-9.0	-1.6

Ground Truth



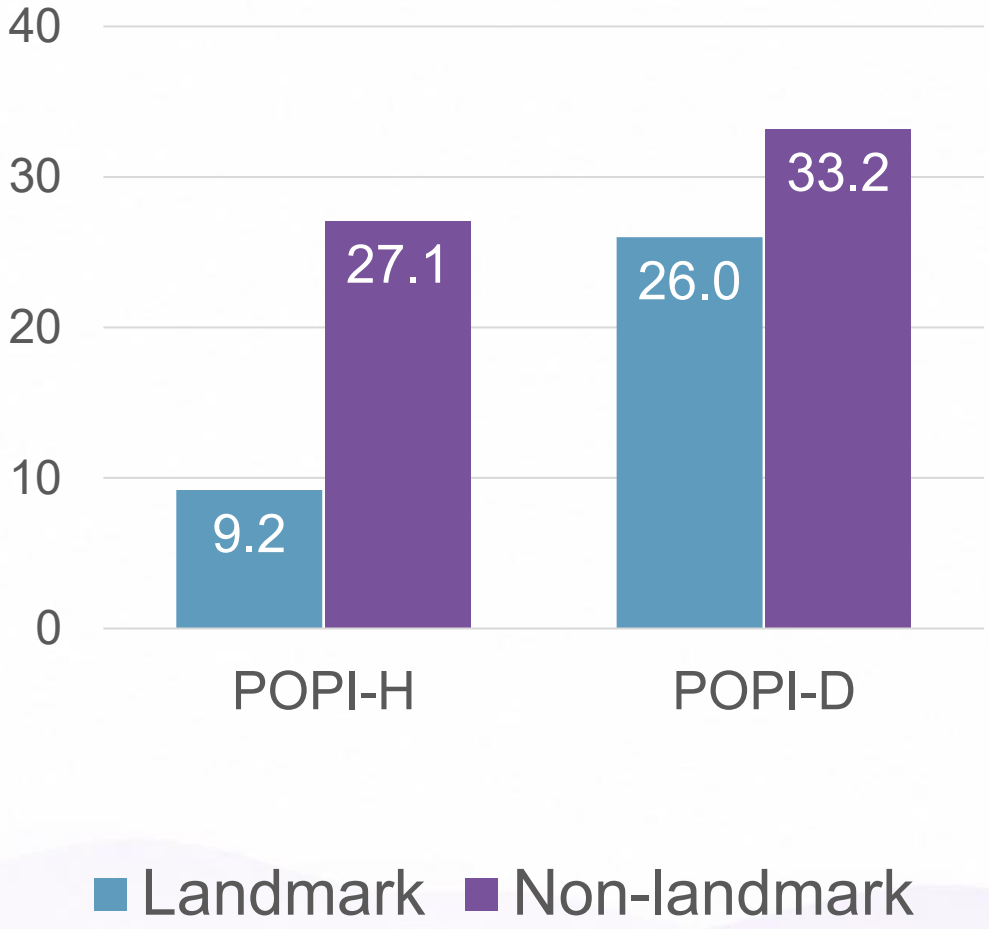
Unconditioned 👕

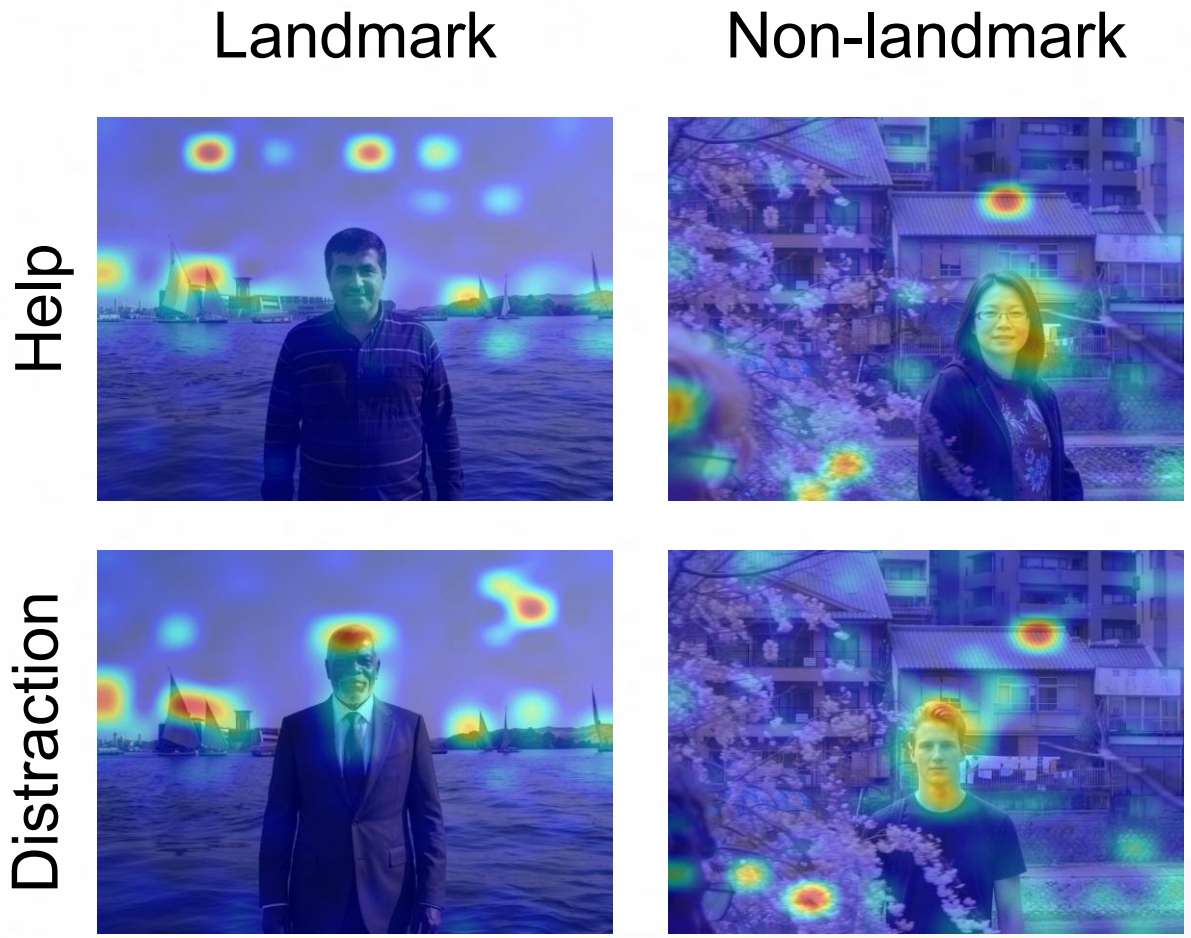


Race-Associated 👕



Landmark vs Non-landmark





Possible Uses

- **Audit bias:** detect reliance on demographic cues instead of geographic evidence.
- **Compare models:** assess prediction stability alongside geolocalization accuracy.
- **Diagnose failures:** study sensitivity to faces, attire, and landmark presence.
- **Evaluate mitigation:** test whether proposed methods reduce demographic sensitivity across groups.

Conclusion

- **GeoBiaset** enables controlled tests of demographic sensitivity in image geolocalization.
- **Person insertion reduces accuracy** across all seven evaluated systems.
- **POPI reveals directional prediction shifts** that accuracy alone does not capture.
- Evaluate geographic accuracy and demographic sensitivity together.