Multimodal Large Language Models Predict Urban Safety Perception but Encode Non-Neutral Demographic Priors

Ciro Beneduce[1] (✉), Bruno Lepri[1], and Massimiliano Luca[1]

Fondazione Bruno Kessler, Via Sommarive 18, Povo (TN), 38123, Italy
cbeneduce,lepri,mluca@fbk.eu

Understanding how people perceive urban environments is essential for inclusive and citizen-centred planning. Yet conventional surveys are costly, difficult to scale, and inevitably represent the perspectives of the populations from whom responses are collected. Street-view imagery and computer vision provide scalable alternatives for measuring perceived urban qualities, while Multimodal Large Language Models (MLLMs) additionally support zero-shot prediction and natural-language explanation. Their assessments, however, may combine visual evidence with geographic and demographic priors acquired during training. We therefore investigate not only whether MLLMs can predict perceived urban safety, but also how they calibrate those predictions and whose perspective their apparently Neutral responses resemble.

We evaluate four open and proprietary MLLMs spanning different architectures and capability levels: LLaVA 1.6 7B, Qwen2.5-VL-72B-Instruct, Gemini 2.5 Flash Lite, and GPT-5.1. The evaluation uses Place Pulse 2.0 [1], a crowdsourced urban-perception dataset covering 56 cities across six continents. To increase label reliability, we retain 1,530 images supported by at least 22 pairwise safety comparisons and binarise their TrueSkill scores at the global median, yielding 751 Safe and 779 Unsafe images. Each model evaluates every image twice without task-specific training and returns a Safe or Unsafe classification, three explanatory keywords, and a short rationale.

Our study design separates three interacting dimensions of image-grounded perception: visual competence, calibration bias, and persona-induced response shifts. A Neutral prompt specifies no observer identity, while persona prompts preserve the same task and wording while introducing one of six race or ethnicity categories, two genders, or three age groups.

All four models recover a meaningful visual signal without fine-tuning. Their city-macro Safe-F1 scores range from 65.4% to 69.4%, while predicted and observed city-level Unsafe rates correlate between $r = 0.60$ and $r = 0.71$. This apparent competence is nevertheless systematically miscalibrated. Safe recall ranges from 91.0% to 99.7%, but Safe precision never exceeds 58.7%. Consequently, the models classify only 5.5% to 24.9% of images as Unsafe, compared with 50.9% in the reference labels.

This Safe-bias is shared across model families. When all four models agree, their errors consist almost entirely of images unanimously classified as Safe despite being labelled Unsafe, whereas unanimous false-Unsafe errors are nearly absent. The resulting geographic performance gap is also partly driven by cal-

ibration rather than by a simple inability to interpret particular regions: the optimistic response strategy produces more errors in cities where Unsafe images are more prevalent, many of which are located in the Global South.

The models' explanations reveal a similarly convergent visual vocabulary. Safe judgements emphasise maintenance, greenery, residential character, daylight, and order. Unsafe judgements emphasise deterioration, isolation, poor lighting, restricted visibility, and limited pedestrian activity. Safety is represented as a relatively homogeneous ideal of the orderly and well-maintained street, whereas unsafety separates into more differentiated themes of physical decay, absence of street life, and spatial isolation. These associations echo established accounts of disorder and informal surveillance in urban environments [3, 2], although generated explanations should be interpreted as model-produced justifications rather than direct access to internal reasoning.

Persona prompting produces substantial and structured changes. Female personas classify more of the same images as Unsafe than Male personas in every model, with significant gaps ranging from 3.7 to 22.1 percentage points. Native American personas produce significant increases in Unsafe classifications across all four models, while Black/African American personas produce some of the largest shifts in three of four. Age effects are less uniform: Middle-aged personas generally provide the least risk-sensitive reference, but whether Young or Elderly personas produce more Unsafe classifications depends on the model. Accordingly, there is no universal model-independent age ordering.

The Neutral output is therefore unspecified rather than neutral. It consistently aligns most closely with the Male persona and generally remains close to the Middle-aged persona, while Black/African American and Native American personas frequently produce the largest departures. These results establish demographic prompt sensitivity, not demographic fidelity: Place Pulse does not provide group-specific human judgements for the same images, so we cannot determine whether the shifts reflect lived experience, training-data correlations, or simplified stereotypes.

Overall, MLLMs can extract scalable and interpretable signals of perceived urban safety, but they do so from a learned and non-neutral standpoint.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A.: Deep learning the city: Quantifying urban perception at a global scale. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 196–212. Springer (2016)
2. Jacobs, J.: The Death and Life of Great American Cities. Knopf Doubleday Publishing Group (1992)
3. Kelling, G.L., Wilson, J.Q.: Broken windows: The police and neighborhood safety. The Atlantic (March 1982), https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/