Wildly Efficient: Towards minimal sampling in urban visual perception studies in the wild

Matias Quintana
Singapore-ETH Centre
Singapore, Singapore
matias.quintana@sec.ethz.ch

Xiucheng Liang
Department of Architecture
National University of Singapore
Singapore, Singapore
xiucheng@u.nus.edu

Evelyn Loo
Institute for Human Development and Potential
Agency for Science, Technology and Research
Singapore, Singapore
evelyn_loo@a-star.edu.sg

Filip Biljecki
Department of Architecture, Department of Real Estate
National University of Singapore
Singapore, Singapore
filip@nus.edu.sg

Abstract

Urban visual perception remains a commonly used tool to quantify the human experience in cities. Studies continue to collect location and demographic-specific data as global models and synthetic data fails to capture residents' nuances. In this work, we explore a first analysis on the minimum data collection for demographic-specific perception prediction models. We found that more data does not always equal better models. While some perceptual dimensions consistently improve with more labeled data, some perceptual dimensions fluctuate. We call for researchers to carefully handle these dimensions, to provide guiding prompts in the perception rating or to have a specific strategy in image sampling. This preliminary results contributes to urban science guidelines and recommendations for more empirical and human-centric studies.

CCS Concepts

• **Human-centered computing**; • **Social and professional topics** → **Geographic characteristics**; • **Computing methodologies** → *Machine learning*; • **Applied computing** → *Law, social and behavioral sciences*;

Keywords

street view imagery, computer vision, human-centric, urban planning

ACM Reference Format:
Matias Quintana, Xiucheng Liang, Evelyn Loo, and Filip Biljecki. 2026. Wildly Efficient: Towards minimal sampling in urban visual perception studies in the wild. In *ACM Sustainability Week 2026 (ACM Sustainability Week Companion '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3765611.3815347

ACM Sustainability Week Companion '26, Banff, AB, Canada

ACM ISBN 979-8-4007-2199-1/26/06
https://doi.org/10.1145/3765611.3815347

1 Introduction

Urban visual perception remains a commonly used tool to quantify the human experience in cities. Studies commonly rely on street-level imagery and human preferences as labels to develop perception prediction models [1]. For localized interventions and analyses, researchers continue to collect local data, both imagery and residents' preferences, to improve global models' performances [3], as these models were found to shift from residents' responses [5]. Despite advances in LLM-based responses, these models still need local context understanding to match human evaluative nuances [4], underscoring the need for data collection in areas in specific locations or preferences from specific populations. In this poster, we propose a first analysis on the minimum number of rated images for demographic-specific urban visual perception prediction models to provide researchers guidelines when deploying urban visual perception studies in the wild. Using a global dataset with demographic information [5], we incrementally fine-tuned a perception model to location-specific respondents perception ratings. We show that model performance is perception-dependent. Some perceptual models require at least 100 samples (i.e., rated images) to stabilize their performance across different demographics, while other indicators show marginal improvements even with more labeled data.

2 Dataset selection and experiment design

Dataset selection: We used the SPECS [5] dataset as ground truth, which includes street-level image pairwise comparisons rated by 200 participants per country (Chile, Netherlands, Nigeria, Singapore, and USA) across ten perceptual indicators. We selected the six traditional indicators (*safe*, *lively*, *wealthy*, *beautiful*, *boring*, and *depressing*) and used country-specific perception ratings [5], retaining images with at least four pairwise comparisons per indicator [1]. Grouping by country of residence yields the largest per-group datasets (273–306 participants).

Model selection: Following existing global perceptual models [2], we used a vision transformer backbone pre-trained on ImageNet with a classification head predicting high (≥5.0) or low (<5.0) ratings, with final scores computed as $10 \times logit + 1 \times (1 - logit)$.

Experiment design: For each country-specific dataset, we applied an 80/20 train/test split and fine-tuned the classification head

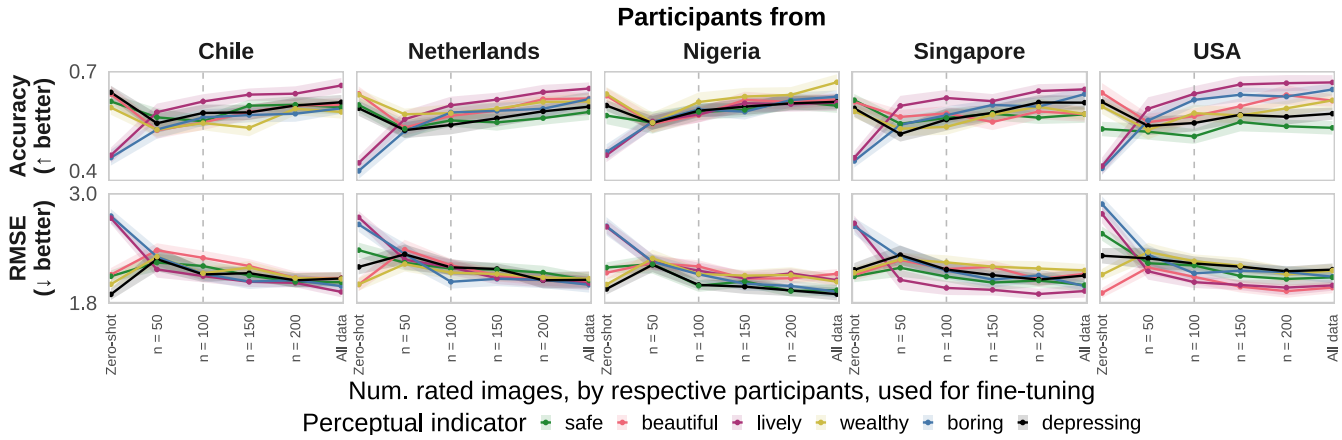

Figure 1: Fine-tuning performance metrics across zero-shot and incremental training samples for each perceptual indicator and demographic-specific dataset. Lines show the average values and shaded areas the 95% CI across 30 iterations.

(backbone frozen) on progressively larger training subsets. We report accuracy and RMSE from zero-shot to incremental predictions, repeated 30 times with different random splits (Figure 1). Implementation details are available at https://github.com/matqr/perception-fine-tuning.

3 Preliminary results

Across all country-specific datasets, *lively* and *boring* perception models consistently improve with more training data, stabilizing at roughly 100 samples. Conversely, the remaining four perceptual dimensions (*safe*, *beautiful*, *wealthy*, and *depressing*) show inconsistent behavior, with marginal performance gains observed primarily in the Netherlands and Nigeria datasets. This suggests that certain perceptual dimensions are interpreted similarly across regions, potentially reducing the data requirements for their modeling. However, others show more variability, highlighting the inherent complexity of modeling subjective preferences A critical factor in this variability is the street-level imagery itself; the selection and composition of these images are essential for capturing human experience. Notably, performance values remained relatively low across all indicators and country-specific datasets. This was anticipated, as perception prediction is notoriously challenging, with global models reporting average accuracies ranging from 61.6% to 76.7% [2]. This modest performance is further explained by the model architecture: the base model retains weights pre-trained on ImageNet (an object recognition task), while the classification head is trained from scratch for perception prediction. Ultimately, our goal was not to achieve state-of-the-art performance, but to demonstrate the potential of fine-tuning models for targeted interventions using publicly available datasets.

4 Conclusions and discussion

These preliminary findings motivate further exploration and increased transparency in urban visual perception modeling, particularly as the research landscape shifts toward more complex, LLM-based architectures [4]. While numerous studies continue to prioritize the optimization of model performance, we shift our attention to the nuances of image selection and data collection protocols. Ultimately, this work poses the fundamental question of whether the future of the field lies in fine-tuning increasingly larger models or in developing specialized architectures trained from scratch for perceptual tasks.

Acknowledgments

This research is supported by The National Research Foundation, Singapore, and Ministry of National Development, Singapore under its Cities of Tomorrow R&D Programme (CoT Award COT-H1-2025-2). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore and Ministry of National Development, Singapore.

References

[1] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep Learning the City : Quantifying Urban Perception at a Global Scale. (2016). https://doi.org/10.48550/arxiv.1608.01769 arXiv:1608.01769

[2] Yujun Hou, Matias Quintana, Maxim Khomiakov, Winston Yap, Jiani Ouyang, Koichi Ito, Zeyu Wang, Tianhong Zhao, and Filip Biljecki. 2024. Global Streetscapes — A Comprehensive Dataset of 10 Million Street-Level Images across 688 Cities for Urban Science and Analytics. *ISPRS Journal of Photogrammetry and Remote Sensing* 215 (2024), 216–238. https://doi.org/10.1016/j.isprsjprs.2024.06.023

[3] Yuhao Kang, Jonatan Abraham, Vania Ceccato, Fábio Duarte, Song Gao, Lukas Ljungqvist, Fan Zhang, Per Näsman, and Carlo Ratti. 2023. Assessing Differences in Safety Perceptions Using GeoAI and Survey across Neighbourhoods in Stockholm, Sweden. *Landscape and Urban Planning* 236 (2023), 104768. https://doi.org/10.1016/j.landurbplan.2023.104768

[4] Milad Malekzadeh, Elias Willberg, Jussi Torkko, and Tuuli Toivonen. 2025. Urban Attractiveness According to ChatGPT: Contrasting AI and Human Insights. *Computers, Environment and Urban Systems* 117 (2025), 102243. https://doi.org/10.1016/j.compenvurbsys.2024.102243

[5] Matias Quintana, Youlong Gu, Xiucheng Liang, Yujun Hou, Koichi Ito, Yihan Zhu, Mahmoud Abdelrahman, and Filip Biljecki. 2025. Global Urban Visual Perception Varies across Demographics and Personalities. *Nature Cities* 2, 11 (2025), 1092–1106. https://doi.org/10.1038/s44284-025-00330-x