

AI-Driven Urban Perception Mapping: Comparing Gemini and LLaVA

Ayse Giz Gulnerman(✉)

Ankara Hacı Bayram Veli University, Ankara 06560, Turkey
ayse.gulnerman@hbv.edu.tr

Abstract. This study evaluates the capability of Generative AI (GenAI) tools in determining urban perception through street-level imagery (SLI) in the central part of Ankara. Specifically, Gemini 1.5 Flash and Large Language Vision Assistant (LLaVA) 13B models are utilized to assess predefined perception aspects. While Gemini operates via an application programming interface (API), LLaVA is executed locally using the Ollama platform. The findings indicate that Gemini 1.5 Flash provides consistent results across English, German, and Turkish prompts, whereas LLaVA 13B struggles with non-English queries and occasionally returns NA values for specific perception aspects. Beyond model performance, this study explores the potential contributions of GenAI to urban perception mapping. By integrating auxiliary data such as air quality, noise levels, and temperature, GenAI could enhance urban analytics beyond visual interpretation. However, challenges remain, including limitations in capturing personal experiences, cultural backgrounds, and subjective perception variations. Despite these constraints, GenAI presents a promising approach for large-scale, automated urban perception assessment, supporting sustainable and inclusive urban planning. Future studies should focus on refining multimodal GenAI models, incorporating diverse data sources, and addressing biases to improve the accuracy and applicability of urban perception analysis.

Keywords: Urban Perception Mapping · GenAI · Street-Level Imagery · Mapillary · Gemini · LLaVA

1 Introduction

1.1 Urban Design and Human Perception

Urban area's influence on human well-being has been widely discussed since 1960 [1–3]. Urban designers, architects, psychologists, and sociologists produced theories and empirical studies on how urban space influences human perception and behaviour [4–8] and momentary-emotion and long-term subjective well-being [9, 10].

Urban designers shape cities by arranging and categorizing activity places turning them into functional environment. Lynch [11], emphasizes that urban space cannot be defined solely by its physical arrangement; it must also encompass a meaningful purpose alongside its physical order. In this perspective, when an individual perceives the history,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15899, pp. 339–349, 2026.
https://doi.org/10.1007/978-3-031-97663-6_30

customs, nature, and essence of the city in visited locations, a joyful connection with the space emerges.

Montgomery [5] defines the sense of urban space with three elements: activity, form, and image. According to this approach, the design affects its user preference and it is a fact that spaces actively used throughout the day increase the perception of reliability and preference rate. This design, called mixed use, keeps the space alive at all hours of the day. Thus, the space gives the human a sense of trust. Similarly, Jacobs [6] states that human immediately notices the diverse urban spaces where there is mobility, uses and supports that place. The four components must be fulfilled to ensure diversity in the space namely: 1- multifunctionality of the space, 2- low rise and short size buildings blocks, 3- harmonious blend of old and new buildings, 4- sufficient level of population density. Jacobs [6] argues that urban diversity cannot be achieved when even one of these four components is missing.

Lynch [4], Norberg-Schulz [7] define urban perception according to movement experience over the constituent parts of the urban area. Therefore, the urban perception might change based on the motion pattern what is gained from urban experience. Though the individual perception of the urban images formed in the mind in an urban area where these experiences (relations) exist, the presence of common imaginary data has the power to produce social perception. Lynch [4] states that this common imaginary can be analyzed through five elements; paths, nodes, districts, edges, and landmarks which helps to picture mind maps. Jojic [12] visualized Lynch's elements; paths, node, landmark, edge and district by matching them with the real-world examples.

Human urban perception is explained by different perspectives based on urban environment features. Bentley [13] claim that human prefer permeability, diversity, robustness, visual suitability, richness, customization, and readability of urban space. From a different angle, Gestalt theory [8] explains urban perception as focusing on the entire space rather than individual urban elements. According to this theory, the overall image holds a distinct meaning from its components. Humans mentally code entire urban space based on the principles of unity, proximity, continuity, similarity, and simplicity.

Holahan [14] suggests that living environments reflect human personality, as individuals shape urban experiences based on personal needs. Human motion also influences perception. Rapoport [15] classifies it as dynamic (e.g., walking) and static (e.g., sitting), noting that dynamic users prefer narrow, linear spaces, while static users favor open areas. Factors like movement purpose, direction, mode, and speed further affect perception [16]. Argin et al. [17] highlight how smartphones reshape visual attention, altering urban perception in the post-flâneur era.

Urban designers influence experiences through design strategies grounded in nudge theory. Defined by Thaler and Cass [18] and Almaas [19], nudging guides individuals toward better decisions without restricting choices. Examples include a rooftop entry at an opera house and a visually enhanced staircase in a library [19]. While nudging is integrated into architectural design, tracking its effects in complex urban environments shaped by visuals, noise, lighting, smell, and crowding remains a significant challenge [20].

1.2 Mapping Urban Perception and Human Wellbeing

Mapping human-urban interaction is important for re-designing healthy urban environments that can contribute to human well-being, but it remains challenging. Studies show that human emotions (as an indicator of instant human-urban interaction) are volatile and can easily be distorted over time as other emotions interfere with the original experience [21, 22].

Reichert et al. [23] presents the technological developments in the last decade, enabling experiments on the relationship between human mental health and urban upbringing as well as city living in a real-life laboratory. While the impact of urbanization on mental health problems is well-established, determining which of the multiple elements of urban living conveys risk remains elusive [23].

Researches start urban perception studies to unravel these elements in real-life urban context. Mainly studies focus on ambulatory assessment (AA) which means combining ecological momentary assessment with smartphones e-diaries, and physical behaviour data collected by accelerometers, GPS-trackers and hearth rate monitors [24, 25]. AA has important role in psychological assessment by recording urban-live experience. While merging momentary assessment with the urban space, remote sensing data for land-use and weather information, volunteered geographic information (e.g., open street map) for base map, social media data for travel behaviour and connotation of locations etc. are adopted in several studies [23, 24, 26].

Beside AA, there are other approaches focusing on just measuring human emotional states related with urban spaces. Some studies, focus on specifically one functional space (e.g., parks) influence on human sentiment. Kovacs-Györi [27] adopts a lexicon that classifies unigrams (single words) into 8 emotional states (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) for classifying social media (tweet) content to present human emotional states within or around parks. Another approach to understand environmental influence on human is urban perception. Salesses et al. [28] aims to map which areas of cities are perceived as safer, livelier, wealthier, more active, beautiful and friendly (6 emotional perceptions) over geotagged images from Google Street View imagery under The MIT Media Lab's Place Pulse project. A pioneering study enhanced the former dataset and present Place Pulse 2.0 dataset including near 111K Google Street View images from 56 cities (28 country) with pairwise labelling by over 81K volunteers to tackle limited urban perception studies at global scale [29]. Similarly, Wang et al. [30] assesses street quality over near 126K street view images based on human perception and additionally investigate the perception relationship with the semantics of objects (e.g., sky, road, and plant) in images.

1.3 Large Language Models Use in Mapping

Recent research has increasingly explored the role of Large Language Models (LLMs) in mapping urban perceptions and spatial understanding. For instance, Verma et al. [31] introduced generative agents powered by LLMs that navigate urban environments and simulate human-like perceptions, enabling new ways to assess urban livability and safety. Similarly, Aman and Matisziw [32] combined language and vision models to map urban sentiment, highlighting how transformer-based models can analyze geotagged social

media content to understand public emotional responses to urban spaces. In the same direction, Yan et al. [33] proposed UrbanCLIP, a framework that integrates textual and visual data through contrastive pre-training to improve urban region profiling across cities. Complementing these developments, Hochmair et al. [34] benchmarked multiple LLMs including ChatGPT-4, Gemini, Claude-3, and Copilot, on their performance in spatial tasks, demonstrating their potential for geographic reasoning. These studies show that LLMs, especially when integrated with geospatial and visual data, provide a promising foundation for mapping and interpreting urban perception.

1.4 Aim of the Research

The influence of urban environments on human well-being, perception, and mental health has been extensively researched for decades. Various theories, approaches, and techniques have been developed to identify how urban spaces and their elements affect individuals. Generative AI (GenAI), a powerful tool across numerous domains, has the potential to become a key method for understanding and mapping urban influences on human life. This research aims to explore how GenAI tools process urban scene datasets from the perspective of human perception.

2 Methodology

The exploration in this study begins with the selection of the case area (urban space), followed by querying, retrieving, and tidying the open dataset (urban scene images), as well as identifying the GenAI tools and process steps used. This section consists of three subsections: case area, data retrieval and tidying, and data processing.

2.1 Case Area

Three features, urban center, pedestrian-based area, and daily population density are used to determine the case area. Based on these criteria, the Meşrutiyet District, located in the central part of Ankara, the capital city of Türkiye, is selected as the case area (Fig. 1). The Meşrutiyet District features crowded, pedestrian-based streets surrounded by busy roads throughout the day. These streets are lined with old, low to mid-rise buildings that have shops, bookstores, cafes, and bars on the ground floors, while the upper floors primarily contain private offices and workplaces.

2.2 Data Retrieval and Tidying

Mapillary [35] is a pioneering platform that provides street-level imagery (SLI) collected through crowdsourcing. Mapillary offers open SLI data for research via its application programming interface (API). The Mapillary API [36] allows users to query SLIs by defining a bounding box. In the first instance, we queried 532 images (Fig. 2(a)) within the border of Mesrutiyet District. Since Mapillary data is crowdsourced, the content often contains several discrepancies, such as varying camera angles, resolution, thematic content and point of view. The retrieved SLIs include irrelevant data for this research and require filtering.

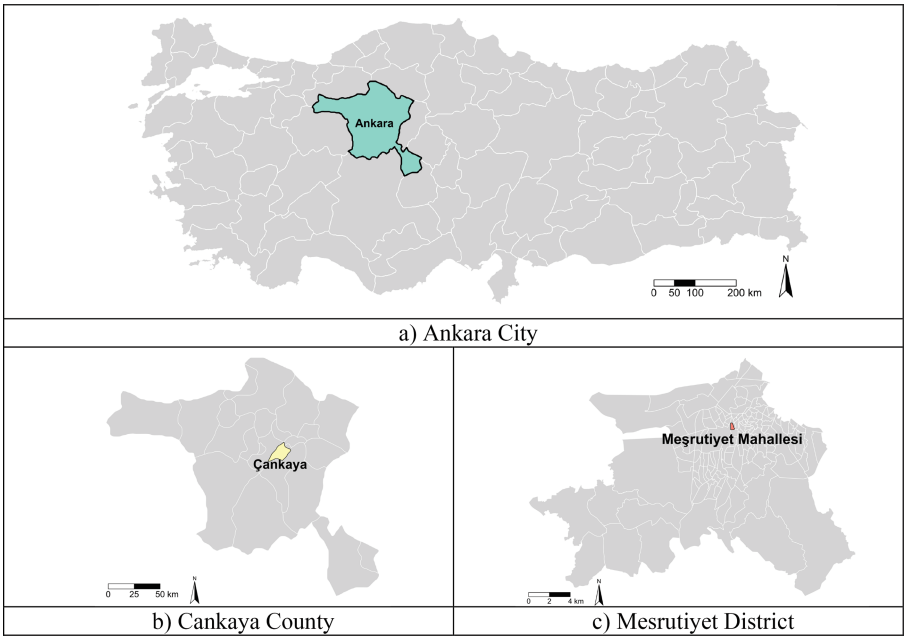

Fig. 1. Location of the case area Mesrutiyet District.

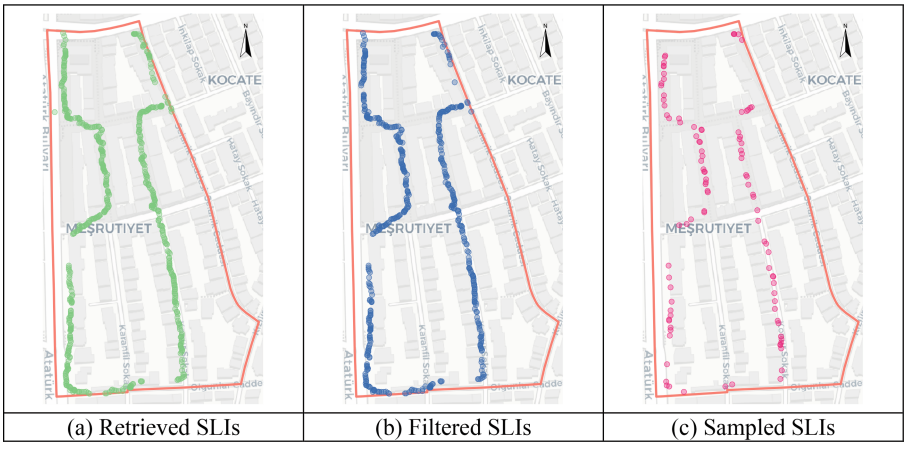

Fig. 2. Retrieved, filtered and sampled SLIs within the bbox of Mesrutiyet District.

After manually filtering the retrieved SLIs, 470 images were retained for further assessment using the techniques explained in Sect. 2.3. During the manual filtering, we considered (1) image clarity, (2) the point of view focusing on covering the street rather than a single object (such as a building, advertising board, or car), and (3) a plausible ground-to-sky ratio. Figure 2(b) displays the filtered 470 SLIs. Although more images

are available, we randomly sampled 100 photos for assess in this study, as shown in Fig. 2(c).

2.3 Data Processing

GenAI refers to a sub area of artificial intelligence to create new content based on the given instructions over existing dataset [37]. GenAI is applied to numerous domains to generate text, image, music, code, and data. In this study we used special form of GenAI which combines language models with vision models in order to interpret our SLIs.

Two different GenAI models, Gemini [34] and LLaVA (Large Language Vision Assistant) [37] are used in this study to generate text descriptions for images based on questions asked about the image contents. Gemini is a GenAI tool powered by Google AI operating multimodality actions i.e., understanding and processing different modes of data (text, code, images etc.). In this study we adopt Gemini 1.5. Flash model through Gemini API. This model is recommended excellent performance for diverse and repetitive tasks. LLaVA is another multimodal GenAI model that has the capability of connecting text and visual data. In this study we used Ollama platform to run LLaVA 13B model locally. We designed the sequence of prompt engineering questioning in the following order:

(1) "Which human perception aspects for urban space do you consider? List the aspects."
(2) "Analyze the image and provide a score from 1 to 10 (1 = negative, 5 = neutral, 10 = positive) for each of the following human urban perception aspects.

In order to compare the results, we asked the first question to Gemini and applied the returned perception aspects as input for the second question to both models. We also make the both models return the answers in a csv file format that help to visualize and compare the performance. In the Gemini model we have also tested prompt engineering in English (en), German (de), and Turkish (tr). Due to low local response returning performance of LLaVA 13B, we did not apply German and Turkish prompt in the model.

3 Results

In response to the first question, the Gemini model provides scores for human perception aspects of urban spaces, including: (1) Visual Appeal, (2) Activity, (3) Spaciousness, (4) Order, (5) Comfort, (6) Accessibility, (7) Natural Elements, (8) Cultural Significance, (9) Emotional Tone, and (10) Sensory Stimulation. The Gemini model assigns scores to each aspect prompted in three languages - English (en), German (de), and Turkish (tr). Figure 3 illustrates the scores for each aspect, as queried in English (a), German (b), and Turkish (c) with the Gemini model, corresponding to each sampled image. Based on the prompted query, which specifies a scoring scale from 1 to 10 (1 = negative, 5 = neutral, 10 = positive), pink shades represent negative perceptions, while green shades indicate positive perceptions. The darkest pink corresponds to the most negative perception, whereas the darkest green represents the most positive perception. Similarly, Fig. 3(d) displays the perception scores returned for each sampled image by the LLaVA model. In Fig. 3(d), LLaVA returns 0 values for some processed images, indicating 'Na' (not available) values. These are represented in red.

When comparing the result plots (maps), the Gemini model demonstrates consistency across languages. However, the results prompted in English (en) clearly indicate a more positive perception compared to those prompted in German (de) and Turkish (tr). On the other hand, LLaVA model returns less consistent and capable results after several try,

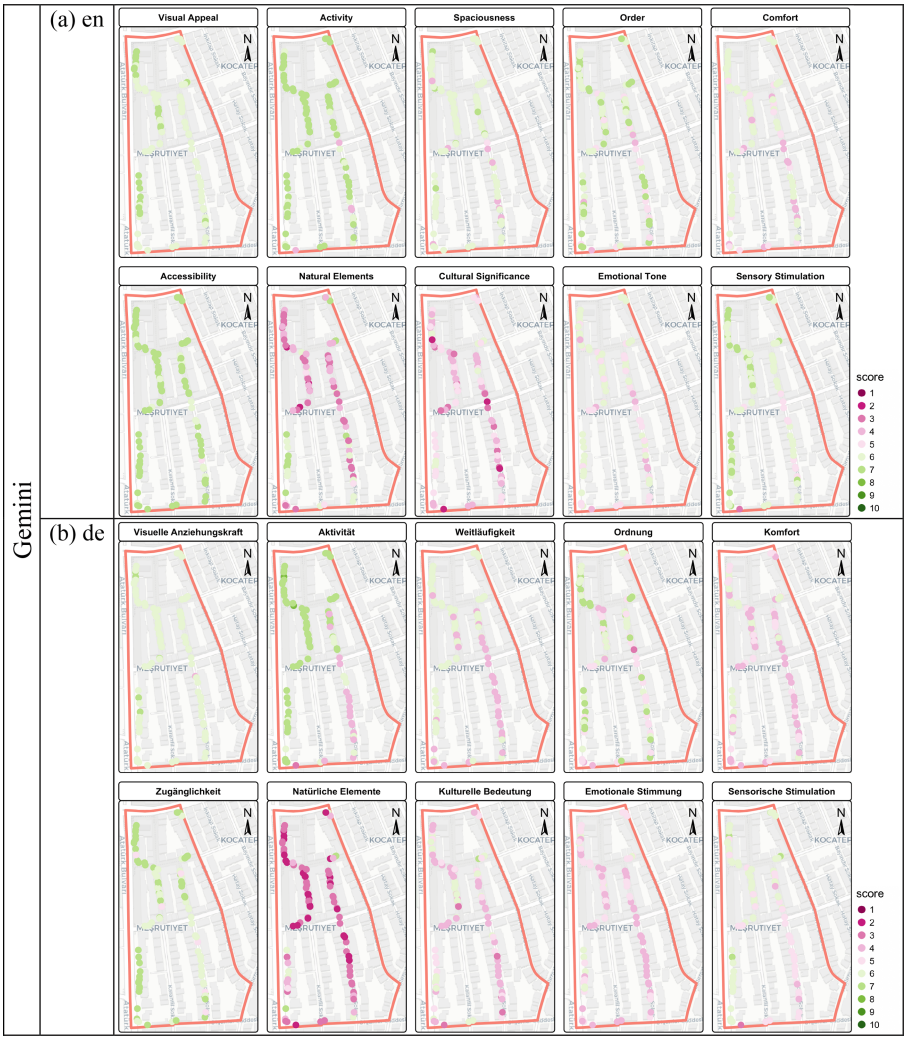

Fig. 3. Human urban perception mapping based on perception aspects using Gemini and LLaVA models.

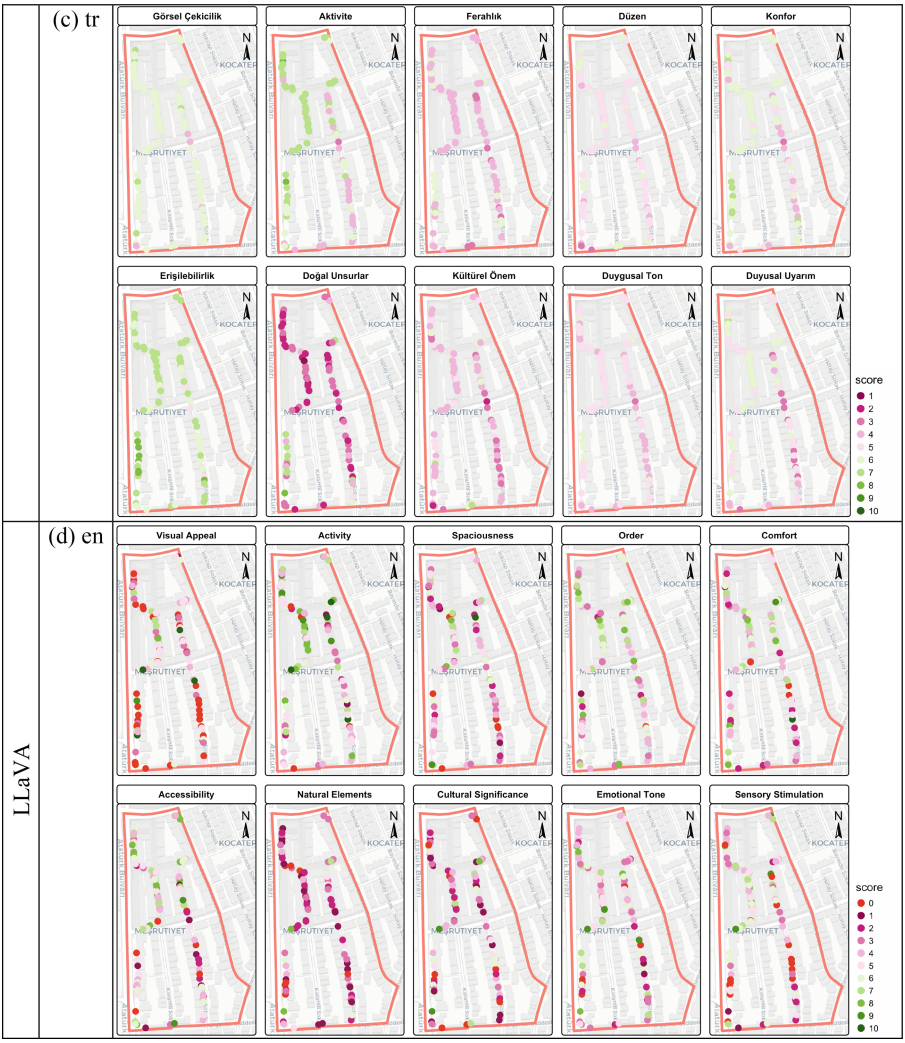

Fig. 3. (*continued*)

we end up the scoring process with some missing scores for some images. In addition, returned scores also have negative perception tendency compared to Gemini results.

Table 1 summarizes the perception scores generated by Gemini and LLaVA. Gemini (en) prompts yielded the highest scores overall, indicating a more positive perception compared to Gemini (de) and Gemini (tr) prompts. Gemini (de) prompts had slightly lower scores than Gemini (en), but still followed similar trend. The Gemini (tr) prompts had the lowest average scores across most perception aspects particularly for "Spaciousness" and "Natural Elements". LLaVA (en)* which is a refined version the summary, aligning somewhat more closely with Gemini however still presents more negative perception compared to Gemini (en) scores.

Table 1. Average scores of human perception aspect by models.

Model	Language	Perception Aspects									
		1	2	3	4	5	6	7	8	9	10
Gemini	En	6.26	6.69	5.9	6.11	5.51	6.83	4.38	4.42	5.46	6.15
	de	6.03	6.31	5.31	5.77	4.81	6.47	3.22	4.79	4.69	5.63
	tr	5.96	6.06	4.28	5.05	5.61	6.63	3.22	4.16	4.87	5.19
LLaVA	en	3.27	5.29	4	4.99	4.61	4.7	2.64	3.52	4.15	4.1
LLaVA	en*	5.11	5.63	4.35	5.14	4.85	5.05	2.97	3.87	4.61	4.77

* Average score excluding 0/NA values.

4 Conclusion

This study investigates the capability of GenAI tools in determining urban perception in the central part of Ankara. The Gemini 1.5 Flash and LLaVA 13B models are employed for this purpose based on predefined perception aspects. While Gemini is accessed via an API, LLaVA operates locally using the Ollama platform.

The findings indicate that Gemini 1.5 Flash demonstrates consistent performance across different language prompts. In contrast, LLaVA 13B returns NA values for some perception aspects in certain sampled street-level images (SLIs) and exhibits limited accuracy when prompted in German and Turkish.

When comparing the two models, Gemini 1.5 Flash provides more reliable and stable results across all three languages, whereas LLaVA 13B struggles to return perception scores consistently, even in English. However, Gemini 1.5 Flash offers only limited free access, while LLaVA 13B runs locally and is free to use, making it a viable alternative for applications requiring offline processing.

GenAI has the potential to revolutionize urban perception mapping by efficiently analyzing SLI in combination with auxiliary data sources such as air quality, noise levels, and temperature to assess environmental comfort, social media geotagged posts for real-time sentiment analysis, and mobility data for accessibility insights. By integrating multi-source data, GenAI can support data-driven urban planning aimed at enhancing walkability, livability, and well-being in cities.

However, GenAI's interpretative capabilities remain in their early stages, with several challenges in accurately assessing human-centered perception aspects. While GenAI can process and analyze visual data, ambient environmental factors such as sound, air quality, and temperature remain underexplored. Moreover, personal attributes (e.g., personality, gender, and age) significantly influence urban perception, as do individual experiences, familiarity with similar spaces, and cultural backgrounds—all of which GenAI alone cannot fully capture when generating urban scene scores.

Despite these limitations, urban perception assessment remains a complex yet crucial research area with significant implications for human well-being. As GenAI models continue to evolve, future studies should focus on enhancing multimodal AI frameworks that integrate spatiotemporal, sensory, and demographic data to provide a more

holistic understanding of urban environments. Additionally, improving language adaptability and cross-cultural interpretation capabilities in GenAI models will be critical for broadening their applicability in global urban studies.

Acknowledgments. This study has not yet received funding from any partners.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Jackson, L.E.: The relationship of urban design to human health and condition. Landsc. Urban Plan. **64**(4), 191–200 (2003)
2. Canter, D.V., Kenneth, H.C.: Environmental psychology. J. Environ. Psychol. **1**(1), 1–11 (1981)
3. Barker, R.G.: Explorations in ecological psychology. Am. Psychol. **20**(1), 1 (1965)
4. Lynch, K.: The image of the environment. The image of the city **11**, 1–13 (1960)
5. Montgomery, J.: Making a city: urbanity, vitality and urban design. J. Urban Des. **3**(1), 93–116 (1998)
6. Jacobs, J.: The death and life of great American. Cities **21**(1), 13–25 (1961)
7. Norberg-Schulz, C.: Intentions in Architecture, no. 74. MIT press, Cambridge (1968)
8. Günay, B.: Gestalt theory and city planning education (2007)
9. Weijs-Perrée, M., Dane, G., van den Berg, P., van Dorst, M.: A multi-level path analysis of the relationships between the momentary experience characteristics, satisfaction with urban public spaces, and momentary-and long-term subjective wellbeing. Int. J. Environ. Res. Public Health **16**(19), 3621 (2019)
10. Dane, G., Aloys, B., Tao, F.: Subjective immediate experiences during large-scale cultural events in cities: a geotagging experiment (2019)
11. Lynch, K.: The city as environment. Sci. Am. **213**(3), 209–221 (1965)
12. Jojic, S.: City branding and the tourist gaze: City branding for tourism development. Eur. J. Soc. Sci. Educ. Res. **5**(3), 150–160 (2018)
13. Bentley, I.: Responsive Environments: A Manual for Designers. Routledge, Abingdon (1985)
14. Holahan, C.J.: Environmental psychology. Annu. Rev. Psychol. **37**(1), 381–407 (1986)
15. Rapoport, A.: The Meaning of the Built Environment: A Nonverbal Communication Approach. University of Arizona Press (1990)
16. Lang, J.T.: Creating Architecture Theory: The Role of the Behavioural Sciences in Environment Design (1987)
17. Argin, G., Pak, B., Turkoglu, H.: Between post-flâneur and smartphone zombie: smartphone users' altering visual attention and walking behavior in public space. ISPRS Int. J. Geo Inf. **9**(12), 700 (2020)
18. Thaler, R.H., Sunstein, C.R.: Nudge: Improving decisions about health, wealth, and happiness. Penguin (2009)
19. Almaas, I.H.: Nudges for a better architecture? An Online Review of Architecture by Arkitektur N (2013). https://www.architecturenorway.no/questions/cities-sustainability/nudge/
20. Gurkaynak, I.: Environmental psychology: nature, history, method. Ankara Univ. J. Fac. Educ. Sci. **21**(1), 1–9 (1988)
21. Levine, L.J.: Reconstructing memory for emotions. J. Exp. Psychol. Gen. **126**(2), 165 (1997)

22. Scherer, K.R.: What are emotions? And how can they be measured? Soc. Sci. Inf. **44**(4), 695–729 (2005)
23. Reichert, M., et al.: Studying the impact of built environments on human mental health in everyday life: methodological developments, state-of-the-art and technological frontiers. Curr. Opin. Psychol. **32**, 158–164 (2020)
24. Reichert, M., et al.: Using ambulatory assessment for experience sampling and the mapping of environmental risk factors in everyday life. Die Psychiatrie **13**(02), 94–102 (2016)
25. Trull, T.J., Ebner-Priemer, U.: The role of ambulatory assessment in psychological science. Curr. Direct. Psychol. Sci. **23**(6), 466–470 (2014)
26. Šćepanović, S., et al.: The healthy states of America: creating a health taxonomy with social media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 15 (2021)
27. Kovacs-Györi, A., et al.: Beyond spatial proximity—classifying parks and their visitors in London based on spatiotemporal and sentiment analysis of Twitter data. ISPRS Int. J. Geo-Inf. **7**(9), 378 (2018)
28. Salesses, P., Schechtner, K., Hidalgo, C.A.: The collaborative image of the city: mapping the inequality of urban perception. PloS one **8**(7), e68400 (2013)
29. Dubey, A., et al.: Deep learning the city: quantifying urban perception at a global scale. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part I 14. Springer, Heidelberg (2016)
30. Wang, L., et al.: Measuring residents' perceptions of city streets to inform better street planning through deep learning and space syntax. ISPRS J. Photogram. Remote Sens. **190**, 215–230 (2022)
31. Verma, D., Mumm, O., Carlow, V.M.: Generative agents in the streets: exploring the use of Large Language Models (LLMs) in collecting urban perceptions. arXiv preprint arXiv:2312.13126 (2023)
32. Aman, J., Matisziw, T.C.: Urban sentiment mapping using language and vision models in spatial analysis. Front. Comput. Sci. **7**, 1504523 (2025)
33. Yan, Y., et al.: Urbanclip: learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In: Proceedings of the ACM Web Conference, pp. 4006–4017 (2024)
34. Hochmair, H.H., Juhász, L., Kemp, T.: Correctness comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for Spatial Tasks. Trans. GIS **28**(7), 2219–2231 (2024)
35. https://www.mapillary.com/dataset/vistas
36. https://www.mapillary.com/developer/api-documentation?locale=tr_TR
37. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Adv. Neural. Inf. Process. Syst. **36**, 34892–34916 (2023)