Computers, Environment and Urban Systems 117 (2025) 102243

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Urban attractiveness according to ChatGPT: Contrasting AI and human insights

Milad Malekzadeh *, Elias Willberg, Jussi Torkko, Tuuli Toivonen

Digital Geography Lab, Department of Geosciences and Geography, University of Helsinki, Finland

ARTICLE INFO

Keywords:
Street view imagery
Spatial planning
AI-human comparison
Urban design
Multimodal large language models

ABSTRACT

The attractiveness of urban environments significantly impacts residents' satisfaction with their living spaces and their overall mood, which in turn, affects their health and well-being. Given the resource-intensive nature of gathering evaluations on urban attractiveness through surveys or inquiries from residents, there is a constant quest for automated solutions to streamline this process and support spatial planning. In this study, we applied an off-the-shelf AI model to automate the analysis of urban attractiveness, using over 1800 Google Street View images of Helsinki, Finland. By incorporating the GPT-4 model, we assessed these images through three criteria-based prompts. Simultaneously, 24 participants, categorised into residents and non-residents, were asked to rate the images. To gain insights into the non-transparent decision-making processes of GPT-4, we employed semantic segmentation to explore how the model uses different image features. Our results demonstrated a strong alignment between GPT-4 and participant ratings, although geographic disparities were noted. Specifically, GPT-4 showed a preference for suburban areas with significant greenery, contrasting with participants who found these areas less attractive. Conversely, in the city centre and densely populated urban regions of Helsinki, GPT-4 assigned lower attractiveness scores than participant ratings. The semantic segmentation analysis revealed that GPT-4's ratings were primarily influenced by physical features like vegetation, buildings, and sidewalk. While there was general agreement between AI and human assessments across various locations, GPT-4 struggled to incorporate contextual nuances into its ratings, unlike participants, who considered both context and features of the urban environment. The study suggests that leveraging AI models like GPT-4 allows spatial planners to gather insights into the attractiveness of different areas efficiently. However, caution is necessary, while we used an off-the-shelf model, it is crucial to develop models specifically trained to understand the local context. Although AI models provide valuable insights, human perspectives are essential for a comprehensive understanding of urban attractiveness.

1. Introduction

Urban environments have a profound role on our satisfaction, travel behaviour, and our health and well-being (Giles-Corti et al., 2016; Nieuwenhuijsen, 2020). Growing evidence shows how well-designed, pleasant urban environments that attract people to walk and cycle, can lead to a range of positive impacts, from higher physical activity and improved mood, to increased active travel, and economic vibrancy (Nieuwenhuijsen, 2020; Saelens & Handy, 2008; St-Louis, Manaugh, van Lierop, & El-Geneidy, 2014; van Wee & Ettema, 2016). Yet, it is also increasingly understood that unpleasant environments can inflict various negative emotions, including fear and anxiety, and be prone to adverse consequences such as crime, car dependence, and avoidance (Carver, Timperio, & Crawford, 2008; de Jong & Fyhri, 2023; Giles-Corti et al., 2016). Obviously, it is highly relevant for local planners and decision-makers to understand and locate such features when developing cities.

While the key components of attractive urban environments were identified long ago (Gehl, 1987; Jacobs, 1992), and have been recognised by the research community (Ewing & Handy, 2009; Forsyth, 2015; Giles-Corti & Donovan, 2003; Saelens & Handy, 2008), the operationalisation of these principles into meaningful and robust spatial indicators is highly dependent on the availability of data. In this respect, recent years have witnessed a dramatic increase in the availability of

* Corresponding author.
E-mail addresses: milad.malekzadeh@helsinki.fi (M. Malekzadeh), elias.willberg@helsinki.fi (E. Willberg), jussi.torkko@helsinki.fi (J. Torkko), tuuli.toivonen@helsinki.fi (T. Toivonen).

https://doi.org/10.1016/j.compenvurbsys.2024.102243
Received 16 September 2024; Received in revised form 15 November 2024; Accepted 13 December 2024
Available online 24 December 2024

micro-scale data collected from the street level, representing the immediate urban environment. The emergence of street view imagery (SVI) in many cities has provided a rich source of data with which to assess the visual quality of streets (Biljecki & Ito, 2021). Together with rapidly developed computer vision techniques for object detection, SVI have allowed researchers to capture detailed street features automatically, thereby overcoming some of the limitations of less detailed neighbourhood-level metrics or field-based audit data collection (Ki, Chen, Lee, & Lieu, 2023). A burgeoning literature has applied SVI to a variety of urban use cases, including the assessment of walkability (Li, Yabuki, & Fukuda, 2022; Nagata et al., 2020), pedestrian and cycling safety (Hamim & Ukkusuri, 2024; Mooney et al., 2020), pedestrian and cycling volume and behaviour (Chen et al., 2020; Liu, Ettema, & Helbich, 2023), street greenery (Li et al., 2015; Ye et al., 2019), microclimate (Gong, Zeng, Ng, & Norford, 2019; Sun et al., 2021), and physical disorder of streets (Keralis et al., 2020; Mayne et al., 2018).

From the planning perspective, the process of turning the visual information contained by SVI into environmental indicators applicable to planning practice nevertheless remains a challenge. Existing literature on SVI and environmental quality has largely focused on searching for the most important correlation between visual street features and travel behaviour, improving the accuracy of existing street quality indicators, and mapping the spatial distribution of distinct environmental features (e.g., visual complexity or street enclosure) in the study cities (Biljecki & Ito, 2021; Liu & Sevtsuk, 2024). Despite the obvious potential of SVI, translating their potential into planning practice is a non-trivial issue. Common hardships include distilling multiple attributes into conceptually and methodologically robust but simple indicators, ensuring spatial coverage, high requirements for technical and methodological know-how, and the need for computational capacity. It is therefore necessary to continue the search for ways which will lower entry barriers and streamline the process of harnessing SVI when evaluating urban environmental quality in planning.

Ensuring a high likelihood of finding these ways lies in the recent breakthroughs in artificial intelligence (AI). The emergence of Multimodal Large Language Models (MLLMs) like GPT-4 (Open AI) and Kosmos-2.5 (Microsoft) with capabilities to integrate the textual interaction capability with image analysis, holds great promise. The capacity of MLLMs to produce human-like text based on large amounts of data, opens new opportunities for applications requiring comprehensive analysis of both visual and textual information, such as the visual assessment of an urban area. Above all, the use of current MLLMs based on simple textual inputs can lower the barriers for using street view data in planning processes and can produce operable environmental quality indicators. However, it is still little known how well the results produced by MLLMs on the attractiveness reflect people's experience, which is a central requirement for their application. Overall, the potential of MLLMs in analysing the attractiveness of urban environments, remains largely underexplored.

For this study, we explored the potential of MLLMs to produce assessments of the attractiveness of urban environments. We applied an AI model to automate the analysis of over 1800 Google Street View (GSV) images collected in Helsinki, Finland. By incorporating the GPT-4 model with urban environmental quality criteria from the literature, we assessed these images through the set of three input prompts, from simpler, to more complex. To compare the ratings of environmental quality that we obtained from the AI model, we asked 24 participants, categorised into residents and non-residents, to rate the attractiveness of the urban environments in the images. By comparing the ratings of the AI models and our participants statistically and spatially, we revealed the potential, as well as the limitations, of MLLMs, when applied to environmental quality assessment. Semantic segmentation analysis was then used to identify and categorize key features in the images during GPT-4's evaluation, offering insights into how the model prioritizes various elements of the urban environment in its assessments. Finally, we discuss the implications and limitations of our approach in planning.

2. Literature review

2.1. AI in sentiment and multimodal analysis

The advent of Large Language Models (LLMs) such as GPT-3 (Brown et al., 2020) and BERT (Devlin, Chang, Lee, & Bert, 2018) has not only revolutionised AI with sophisticated text generation and understanding, but also democratised interactions with AI technologies (Bilgram & Laarmann, 2023). These models enable users to execute commands, optimise and fine-tune AI responses, and engage in nuanced interactions without requiring deep AI expertise. This suite of LLMs illustrates the leap towards intuitive, accessible technology, transforming user interactions across various domains. This capability is particularly crucial for our study, allowing for the customisation of analysis criteria. Nonetheless, the application of LLMs is limited, because it lacks the ability to process visual media, essential for assessing urban attractiveness.

The emergence of MLLMs like GPT-4 (Achiam et al., 2023) and Kosmos-2.5 (Huang et al., 2024; Peng et al., 2023), which integrate the textual interaction capability of LLMs with image analysis, offers novel potential for applications requiring integrated visual and textual interpretation. Early studies on MLLMs have focused on understanding how these models interpret and relate visual and textual inputs, leading to advancements in their ability to generate responses based on diverse data types, such as images and text descriptions. Notable foundational models include BLIP-2 (Bootstrapping Language-Image Pre-training-2) (Li, Li, Savarese, & Hoi, 2023), designed to generate detailed image captions and descriptive text; CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021), which matches images to textual descriptions based on contextual similarities; and LLaVA (Large Language-and-Vision Assistant) (Liu, Li, Wu, & Lee, 2024), which enhances interactive understanding by linking language processing with image features.

As MLLMs have evolved, their scope has broadened to support more complex multimodal tasks, using distinct components like modality encoders, which process specific input types (e.g., images or text), LLM backbones, and modality generators that produce outputs suited to each data type (Zhang et al., 2024). This architecture allows for efficient handling of complex data inputs, providing flexibility in input representation and enabling seamless integration of multiple data types—such as images, text, and even audio—into a unified framework for diverse analytical tasks.

This advancement provides new options for applications requiring comprehensive analysis of both visual and textual information, such as an urban area's visual assessment, that can potentially benefit spatial planning. Some early studies have explored this field, including the study by Jongwiriyanurak et al. (Jongwiriyanurak et al., 2023), which used LLaVA by prompting six questions to gather information on various factors considered critical in assessing motorcycle crash risks. Similarly, Liu et al. (Liu, Haworth, & Wang, 2023) employed CLIP to assess perceived walkability by analysing both tangible and subjective factors such as safety and attractiveness. Despite the progress made, the deployment of multimodal LLMs for a detailed analysis of urban attractiveness is still largely underexplored.

2.2. Determinants of urban attractiveness

During recent decades, the determinants of urban quality and attractiveness have become well established by an interdisciplinary research community. Within this literature, one of the key works is the book by Ewing et al., *Measuring Urban Design: Metrics for livable places* (Ewing et al., 2013), which details a framework of metrics for measuring the quality of urban environments, as well as the definitions and measurement protocols to operationalise these metrics. This framework is grounded in the multidisciplinary understanding of how physical spaces and urban design qualities interact to influence individual reactions and

behaviours, particularly walking behaviour, which is often a proxy for urban attractiveness. Ewing's framework is especially useful to our study as it provides a systematic foundation for identifying and assessing the key elements that shape urban attractiveness, allowing us to draw on well-established metrics and principles to evaluate the attractiveness of different urban environments.

The framework divides the relevant metrics into three main groups. The first group, *enduring physical features*, comprises elements such as sidewalk features for pedestrian activity, street design for traffic and activity, tree canopy and greenery, physical indicators of human activity, and permanent lighting. The second group encompasses *urban design qualities*, including imageability, legibility, human scale, transparency, linkage, complexity, and coherence. Thirdly, the last group extends the evaluation criteria to include *individual reactions*, reflecting personal and emotional responses to the urban environment.

In the literature, physical features like sidewalk width, street width, and tree canopy are directly observable and are believed to influence the more subjective urban design qualities (Clifton, Livi Smith, & Rodriguez, 2007; Pikora et al., 2002; Wimbardana, Tarigan, & Sagala, 2018). These features are often used in active transportation audit instruments, to measure the quality of the walking or bicycling environment (Day, Boarnet, Alfonzo, & Forsyth, 2006; Emery, Crump, & Bors, 2003; Khisty, 1994; Landis, 1994; Pikora et al., 2002; Wimbardana et al., 2018).

On the other hand, urban design qualities, while influenced by these physical features, contribute to a cumulative effect on the experience of walking down a street that is greater than the sum of the parts. For instance, imageability, a concept popularised by Lynch (Lynch, 1960), refers to the quality in a physical object which gives it a high probability of evoking a strong image in any given observer. It is what makes a space memorable and distinct. Similarly, qualities such as legibility, which is the ease with which a place can be recognised and organised into a coherent pattern, play a crucial role in how an individual perceives and engages with an urban space (Evans, Smith, & Pezdek, 1982; Lynch, 1960). The concept of transparency, derived from architecture and urban planning, refers to the literal and figurative visibility of a place. It affects how individuals perceive the openness and accessibility of a space, which is crucial for the sense of appealing (Arnold, 1980; Ewing & Handy, 2009; Jacobs, 1993). Complexity and coherence, on the other hand, reflect the visual richness and orderly arrangement of urban elements, which have been empirically linked to people's preference for and engagement with urban spaces (Rapoport, 2013). Moreover, enclosure, as described by Alexander (Alexander, 1977) and Jacobs (Jacobs, 1993), refers to the creation of well-defined outdoor spaces with clear shapes and boundaries, akin to rooms, which evoke feelings of safety, definition, and memorability. Lastly, human scale and linkage refer to how the proportions of space and elements correspond with human dimensions, ensuring comfort (Moudon & Lee, 2003), connectivity of different spaces (Craig, Brownson, Cragg, & Dunn, 2002), and facilitating movement and interaction (Gehl, 1987).

Moreover, the inclusion of subjective reactions in the evaluation criteria acknowledges the multifaceted nature of urban design qualities. While physical features can be measured objectively, their influence on individual perceptions and behaviours is subjective, and can vary widely. As such, an evaluation should consider individual reactions, such as a sense of safety, comfort, and interest, which are personal yet pivotal components of an environment's pleasantness. As Talen stated, the field of urban analysis has yet to reach consensus on the most appropriate measures to use (Talen, 2002). The literature reveals a diversity of approaches, with various studies opting for singular measures, others for combinations, but without a universally accepted standard.

3. Methodology

In this study, we used street view imagery as the input for MLLMs from which to evaluate urban attractiveness, which was then compared against human participant ratings. First, to find the evaluation criteria to optimise the AI model for our study, we defined a set of criteria and determinants of urban attractiveness. Our criteria were based on Ewing et al. (Ewing et al., 2013) (see section 4) to align with established theories and empirical evidence from the urban design literature. These criteria served as prompts in conjunction with visual data when engaging with the AI model. Initially, imagery data were collected and subjected to preliminary analysis with a subsample of the data to determine the most appropriate AI model. Following the preliminary analysis and model selection, participants were asked to rate the images, allowing for a comparative analysis between AI-generated and human assessments. Given that the selected model was a commercial, proprietary model, we employed various techniques to explore its black-box nature and gain insights into the mechanisms it uses to evaluate the attractiveness of urban environments. Detailed procedural steps are provided in the following.

3.1. Study area

Our study took place in Helsinki, the capital of Finland (Fig. 1-a). Helsinki is a medium-sized city with a population of about 650,000 and a total land area of 214 km^2 (Mäki & Sinkko, 2022) (Fig. 1-b). The city comprises various types of urban fabric, including a densely built urban core at the tip of the peninsula with the highest population density in the county of around 5550 people per km^2 and the centre of economic, cultural, and social activity (Fig. 1-c). Further away from the city centre are residential areas in the western, northern, and eastern parts of the city. The city, apart from its centre, is characterised by its greenery and multiple green spaces, which comprise approximately one-third of the total land area of Helsinki.

3.2. Acquisition of urban imagery

We acquired panoramic GSV images from Helsinki through Google Maps API with a key authorisation. The images had been collected between 2009 and 2017 over the study area (Fig. 1-b). In the acquisition process, images were sampled at 20 m intervals along the street network. Each GSV image included a time stamp referring to the month and year when the image was taken, as well as the coordinate location of the image. The size of the images was 640 × 640 with a field of view of 60° and pitch 0°. The panoramic 360° image that we used in the models and in the human evaluation, was composed of six directional images (0°, 60°, 120°, 180°, 240°, 300°) in each compass location, with 0° directed towards the north.

To select the images for AI and human evaluation, we first randomly sampled 1000 GSV image locations from the study area to ensure spatial coverage. To ensure that we did not miss important locations from the local residents' point of view, we consulted a survey by the City of Helsinki on the walkability of local neighbourhoods (Norppa & Kaunis, 2020). We located images that were up to 50 m to the important locations that the residents had mapped and added these to our sample. Finally, we removed duplicate images, which resulted into a total set 1967 images that we used as the input for the AI models and human evaluations.

3.3. AI evaluation of urban imagery

3.3.1. Preliminary analysis

In our preliminary analysis, we evaluated three multimodal large language models (MLLMs): CLIP, BLIP, and GPT-4, to identify the most suitable model for assessing urban attractiveness. Each model was examined for its ability to produce ratings that aligned with human evaluations. To establish a benchmark, we asked a small group of 10 non-residents to rate 10 street view images based on their general attractiveness, providing a comparative baseline against which we measured the performance of the AI models.

CLIP (Radford et al., 2021) was initially considered due to its

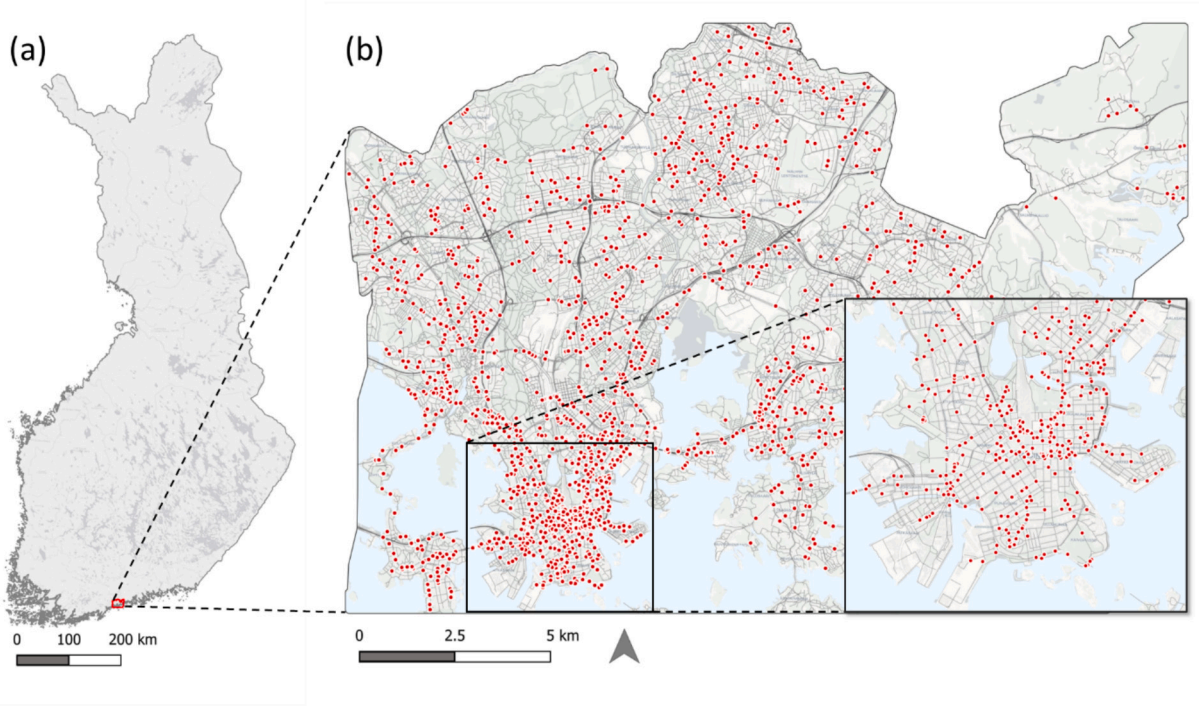

Fig. 1. (a) Location of Helsinki within Finland; (b) Helsinki's urban and suburban areas with red dots indicating the locations of Google Street View (GSV) images used in the study. Inset: A detailed view of Helsinki's urban core. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

strength in associating images with relevant textual descriptions. However, while CLIP excels at image classification tasks, such as categorizing an image based on content (e.g., "70% park, 20% residential area"), it lacks the capability to directly assess attractiveness. To work around this, we attempted to use CLIP's classifications to generate descriptive sentences and then applied VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto & Gilbert, 2014) for sentiment analysis. This approach proved inadequate, as sentiment analysis models like VADER evaluated the sentiment of the sentences rather than the attractiveness of the scenes described in the sentences. Moreover, when we used GPT-4 to interpret these sentences, the results were inconsistent and failed to capture the nuanced aspects of urban environments.

BLIP (Li et al., 2023) showed more promise in generating captions and responding to higher-level questions about the images. By leveraging a variant known as BLIP_VQA, we generated descriptive statements that GPT-4 could interpret more effectively. Despite minor grammatical errors, ChatGPT could understand and process these statements effectively. For an example of the statements, refer to Appendix A.

Ultimately, GPT-4 alone emerged as the most effective tool for our study. It provided consistent, sensible ratings that closely aligned with those from the preliminary human evaluations, with an average difference of just 0.56 (standardized values) compared to the 0.76 difference when using BLIP and GPT-4 together. Therefore, we chose GPT-4 as the primary model for our main study, which will be discussed in detail in the following section.

3.3.2. *Primary analysis*

Drawing on the determinants of attractiveness outlined in section 4, we developed three distinct prompt types for GPT-4, ranging from simple to complex, to assess the influence of these criteria on the attractiveness scores. These prompts were iteratively refined during the preliminary analysis by adjusting the wording to better reflect determinants of urban attractiveness, to enhance the alignment between AI-generated assessments and participant ratings, thereby reducing discrepancies. To avoid any unintended bias or exaggeration in the evaluation process, we replaced the term "attractiveness" with "visual appeal." This refinement allowed GPT-4 to focus more effectively on the specified criteria rather than overemphasizing individual features. Using OpenAI's API allowed us to automate the submission of prompts and to retrieve results for the extensive number of images, a necessity given the volume of data.

For the first prompt, we simply asked GPT-4 to rate the overall visual appeal on a scale from 1 (completely unappealing) to 7 (completely appealing), without specifying any criteria (Table 1). The second prompt incorporated a set of physical features. In the third prompt, we integrated urban design quality criteria, as well as subjective reactions, into the previous set of criteria. Each criterion in the second and third prompts included a brief definition. For all prompts, we instructed GPT-4 to disregard temporary elements such as weather or passing vehicles, to ensure consistency in responses. We also developed two distinct prompts for querying ChatGPT, tailored to reflect either a local resident's or a non-resident's perspective, which could be typical or atypical in terms of local aesthetic and environmental viewpoints. Hereafter, we have referred to these prompts as Model-1, Model-2, and Model-3, respectively, with the suffixes 'LR' for local residents and 'NR' for non-residents. Each prompt was accompanied by a panoramic image of the area. The exact wording of these prompts is available in Appendix B.

For each criterion, GPT-4 was instructed to provide a single integer rating. In the case of the first prompt, only one overall rating was requested. In this study, we used a simple average without weighting the criteria, treating all as being equally important. Assigning appropriate weights to different criteria is not a straightforward task, as it requires a nuanced understanding of how various factors contribute to attractiveness. This could be achieved through analysis of people's opinions and preferences or the input of domain experts. However, as the literature reveals (Boivin & Tanguay, 2019; Broitman et al., 2020; Li, Li, Jia, Zhou, & Hijazi, 2022; Liu, Silva, Wu, & Wang, 2017; Moura, Cambra, &

Table 1
Criteria used in each ChatGPT query prompt with their main relevant references.

Prompt	Criteria	References
Model-1 (Prompt 1)	Overall attractiveness	
Model-2 (Prompt 2)	**Enduring Physical Features** - Sidewalk Features for Pedestrian Activity - Street Design for Traffic and Activity - Tree Canopy and Greenery - Physical Indicators of Human Activity - Permanent Lighting	**(Clifton et al., 2007; Ewing & Handy, 2009; Pikora et al., 2002; Wimbardana et al., 2018)**
Model-3 (Prompt 3)	**Enduring Physical Features** - Sidewalk Features for Pedestrian Activity - Street Design for Traffic and Activity - Tree Canopy and Greenery - Physical Indicators of Human Activity - Permanent Lighting **Urban Design Qualities** - Imageability - Legibility - Enclosure - Human Scale - Transparency - Linkage - Complexity - Coherence **Subjective Reactions**	(Clifton et al., 2007; Ewing & Handy, 2009; Pikora et al., 2002; Wimbardana et al., 2018) (Alexander, 1977; Arnold, 1980; Craig et al., 2002; Evans et al., 1982; Ewing & Handy, 2009; Gehl, 1987; Jacobs, 1993; Lynch, 1960; Moudon & Lee, 2003; Rapoport, 2013) (Birenboim, 2018; Ewing & Handy, 2009; Zhang, Zheng, & Wang, 2022)

Gonçalves, 2017; Tang & Long, 2019), there is no consensus on a standardized weighting scheme that applies to our specific criteria, making the implementation of a weighted average a complex and context-dependent endeavour. Future experiments could explore the application of weighted averages, potentially to achieve results that align more closely with participants' ratings.

3.4. Human evaluation of urban imagery

To assess the ratings generated by GPT-4, we primarily recruited university students as participants, and personnel affiliated with the institutions of the authors. Recognising that familiarity with an area and associated memories can influence perceptions, we included both residents and non-residents in our participant pool, to evaluate these effects and to compare their ratings with those from GPT-4. The study involved 13 participants who were residents of Helsinki at the time of the experiment, and 11 non-residents. Participants received instructions via a concise guidance document. We intentionally did not highlight specific criteria, unlike the prompts used with GPT-4, allowing participants to rate the attractiveness based on their intuitive perception, as if they were physically present in the area. To maintain consistency with the AI prompts and minimize bias, we replaced the term "attractiveness" with "visual appeal" in the participant instructions, avoiding any "wow" effect the term "attractiveness" might imply. This approach encouraged a more subjective assessment rather than a detailed objective analysis of features. However, they were instructed to disregard any temporary features. For the exact wording of the instructions, please refer to Appendix C.

Given the substantial number of images involved and the time-intensive nature of the task, we requested that participants rate at least 500 images each, although they were permitted to rate more if they chose to. To ensure a balanced distribution of ratings and prevent any single image from being rated excessively or not at all, we strategically divided the batch of images. On average, each participant provided 1014 ratings, accumulating a total of 24,349 ratings in total. Each image was rated at least 9 times. On average, each image received 11 ratings.

To evaluate the reliability of the raters for the images, we conducted a Cronbach's alpha test as an inter-rater reliability measure (Agbo, 2010; de Vet, Mokkink, Mosmuller, & Terwee, 2017). Given that not all participants rated every image, we employed a pairwise deletion method in our analysis. This approach ensures that only cases with missing data on specific variables are excluded from the calculation, allowing each statistic to be based on the most complete sample available for that particular analysis (Enders, 2022).

3.4.1. Adjusting ratings for comparative analysis

Recognising the subjective nature of attractiveness, we noted significant variance in the average ratings provided by participants; the lowest individual average was 3.12, while the highest was 5.55. Such disparities indicated that comparing raw rating values could be problematic and not directly comparable. To address this, we standardized ratings of the individual to highlight the relative attractiveness of areas. To maintain consistency, we also standardized the GPT-4 ratings, thus adjusting the data across different evaluators and prompts.

To adjust for potential bias due of luminosity, we investigated whether luminosity influenced the ratings, based on the assumption that brighter and sunnier images might receive higher ratings. If this correlation had been significant, it would have been necessary to adjust for luminosity in our analysis. However, our findings showed no significant correlation between luminosity and the attractiveness ratings (Appendix D). Consequently, we did not adjust the images based on luminosity.

3.5. Statistical analysis

We observed that the ratings from the first prompt resulted in a non-normal distribution (refer to Appendix E). Consequently, when comparing the distribution of ratings from the first prompt with those of residents and non-residents, we employed the Wilcoxon test. In contrast, since the distributions from prompts two and three were normally distributed, we applied *t*-tests for comparisons. Additionally, we calculated the Pearson correlation coefficient to assess the degree of correlation between the ratings from GPT-4 and those provided by residents or non-residents.

To evaluate the ratings spatially, we first assessed the overall spatial autocorrelation using Moran's I. We then calculated the differences between the ratings from GPT-4 and participants to further analyse these discrepancies using Moran's I. Subsequently, we identified clusters of these differences using local Moran's I. To pinpoint the hot spots (areas where the differences between GPT-4 and participant ratings are positive and significant) and cold spots (areas where these differences are negative and significant), we employed the Getis-Ord G* statistic.

3.6. Decoding GPT-4's outputs

Given that GPT-4 is a proprietary model, its underlying architecture and the mechanisms behind its multimodal capabilities, particularly in image analysis, remain largely unknown. Although GPT-4 has demonstrated a relatively good level of proficiency in identifying objects and understanding spatial relationships (Chen et al., 2024) compared to other MLLMs, the process by which it derives values for attractiveness or specific criteria used in our study is not well understood. To explore this black box, we adopted an experimental approach. We began with semantic segmentation analysis to quantify the proportions of various object classes within the images. Using these identified classes, we then trained a random forest (RF) model to explore the relationship between these features and the responses generated by GPT-4 across all six models. This approach offers preliminary insights into which elements

of a scene might influence the GPT-4's scores. However, it is important to note that this analysis is just a small step towards transparency; many factors, such as the nuanced interactions between detected objects, were not explored in this study and remain outside the scope of our research.

To segment the images into distinct classes, we employed the InternImage semantic segmentation model (Wang et al., 2023), using the InternImage-XL backbone integrated with the UperNet architecture (Xiao, Liu, Zhou, Jiang, & Sun, 2018). We used a pre-trained variant of this model, which is trained on the Cityscapes dataset (Cordts et al., 2016). The segmentation was performed on the Finnish CSC's Puhti supercomputer, utilizing its Nvidia V100 GPU. This process produced pixel-level predictions for each image, classified according to the semantic categories defined by the Cityscapes dataset, which were then used as features in our analysis.

We selected a Random Forest (RF) model due to the high level of non-linearity and complexity observed in our preliminary tests, which ruled out linear regression models despite their interpretability. To address potential multicollinearity, we tested the features and excluded those with high correlation, and combined all vehicles related features to a new feature, road transport, resulting in a final set of eight features: vegetation, road, building, road transport, sky, person, sidewalk, and terrain. Optimal parameters for the RF models were determined using a grid-search cross-validation approach from the Scikit-learn package in Python (Pedregosa et al., 2011). To assess the significance of each feature within the model, we employed the permutation feature importance (PFI) method, a model-agnostic approach that evaluates the impact of each feature by measuring the decrease in model accuracy when the feature's values are randomly shuffled (Altmann, Toloşi, Sander, & Lengauer, 2010).

To visualize the relationship between the features and the GPT-4 attractiveness scores, we utilized accumulated local effects (ALE) plots (Apley & Zhu, 2020). ALE plots are particularly useful in the presence of correlated features, as they calculate the average effect of a feature on the dependent variable within specific value ranges. These plots provide a clearer and more accurate understanding of the marginal effects of each variable, helping to elucidate the non-linear patterns in the relationship between different features and the attractiveness scores generated by GPT-4.

4. Results

4.1. Descriptive findings

The results of the inter-rater reliability test yielded a Cronbach's alpha value of 0.94, indicating an exceptionally high level of agreement among the participants' ratings. This suggests that the ratings were highly consistent across different raters, underscoring the reliability of the subjective assessments provided. Such a strong inter-rater reliability further validates the robustness of the dataset and the credibility of the comparisons made between human and AI-generated evaluations, even with a relatively small number of participants.

In the comparison between the local resident and non-resident groups of participants, we observed a similar pattern in the ratings (Table 2). Local residents showed a slightly lower standard deviation, but a broader range of values compared to non-residents. For the GPT-4 Model-1 (LR and NR), the standard deviation was higher, and the range of values broader than those observed in participant ratings, with both models displaying a tendency towards lower median values. This suggests that despite a general trend towards lower ratings, the values above the average were considerably higher for GPT-4 compared to the participants. When examining GPT-4 Models 2 and 3 (LR and NR), the standard deviations aligned closely with those of participants, but the range of values was broader, especially on the lower, negative values. Nevertheless, the values for the second and third quartiles remained similar across all models and participant ratings.

In analysing the distribution of ratings between GPT-4 models and participants using Wilcoxon and T-tests, we found no significant differences, as indicated by the *p*-values (Table 3). This similarity in distributions suggests that the ratings from the GPT-4 models align closely with those provided by participants, with no statistically significant variations between the groups.

Analysis of Pearson's R correlation values between the ratings from GPT-4 models and participants reveals a moderate positive correlation, with the highest value being 0.54. The highest correlation within the local residents' group can be observed with GPT-4 Model-1 (Fig. 2-a), while for the non-residents group, GPT-4 Model-3 shows the highest correlation (Fig. 2-b). GPT-4 Model-2 and Model-3 were highly correlated with each other, whereas this high correlation did not hold between these two models and GPT-4 Model-1. However, these differences are slight and not substantial, all remaining below 0.06. This indicates minimal variation among the GPT-4 models in terms of their correlation levels with participant ratings.

4.2. Spatial statistics

To explore spatial autocorrelation within our dataset, we first analysed the global Moran's I for each set of ratings. The results indicated a significant positive spatial autocorrelation across all observations, ranging from 0.16 to 0.39 ($p = 0.001$ for all cases). Ratings from non-resident participants showed the highest level of spatial autocorrelation, at 0.39, suggesting that participants have more consistent and generalised perceptions of an urban area when it is unfamiliar to them. In contrast, residents, who have detailed and varied understanding based on their personal experiences and familiarity with the area, are likely to use more diverse criteria in their ratings, leading to lower spatial autocorrelation, with a Moran's I of 0.26.

Comparing the spatial autocorrelation of ratings derived from different GPT-4 models, we observe that the more complex the model (i.e., the more criteria in the prompt), the higher the spatial autocorrelation. We observed that Model-1 for local residents had a Moran's I of 0.19, whereas Model-2 and Model-3 yielded values of 0.26 and 0.29, respectively. A similar pattern was observed for non-residents, where Model-1 produced a Moran's I of 0.16, while Model-2 and Model-3 increased to 0.23 and 0.26. This result can be interpreted in two contradictory ways. First, simple models may apply fewer criteria and lack the sophistication to analyse complex urban features consistently,

Table 2
Summary statistics for GPT-4 models and participant ratings.

Ratings	Mean	Std	Min	25 %	50 %	75 %	Max
Local Residents	0	0.72	−2.37	−0.49	0.01	0.52	1.92
Non-Residents	0	0.75	−2.17	−0.51	0.07	0.56	1.96
Model-1 LR	0	1.00	−2.67	−0.83	−0.83	1.00	2.84
Model-2 LR	0	0.76	−3.97	−0.28	0.17	0.51	1.99
Model-3 LR	0	0.76	−3.76	−0.37	0.15	0.52	1.78
Model-1 NR	0	1.00	−4.50	−0.75	−0.75	1.11	1.11
Model-2 NR	0	0.76	−3.84	−0.30	0.16	0.48	1.82
Model-3 NR	0	0.75	−3.70	−0.36	0.15	0.52	1.80

Table 3
Statistical comparison of GPT-4 models and participant ratings' distributions (significance level = 0.05).

Ratings	Statistic	*p*-value
	Local Residents	
Model-1 LR (Wilcoxon)	805,480.0	0.50
Model-2 LR (T-Test)	0.95	0.33
Model-3 LR (T-Test)	0.96	0.33
	Non-Residents	
Model-1 NR (Wilcoxon)	808,881.0	0.60
Model-2 NR (T-Test)	0.89	0.36
Model-3 NR (T-Test)	0.92	0.35

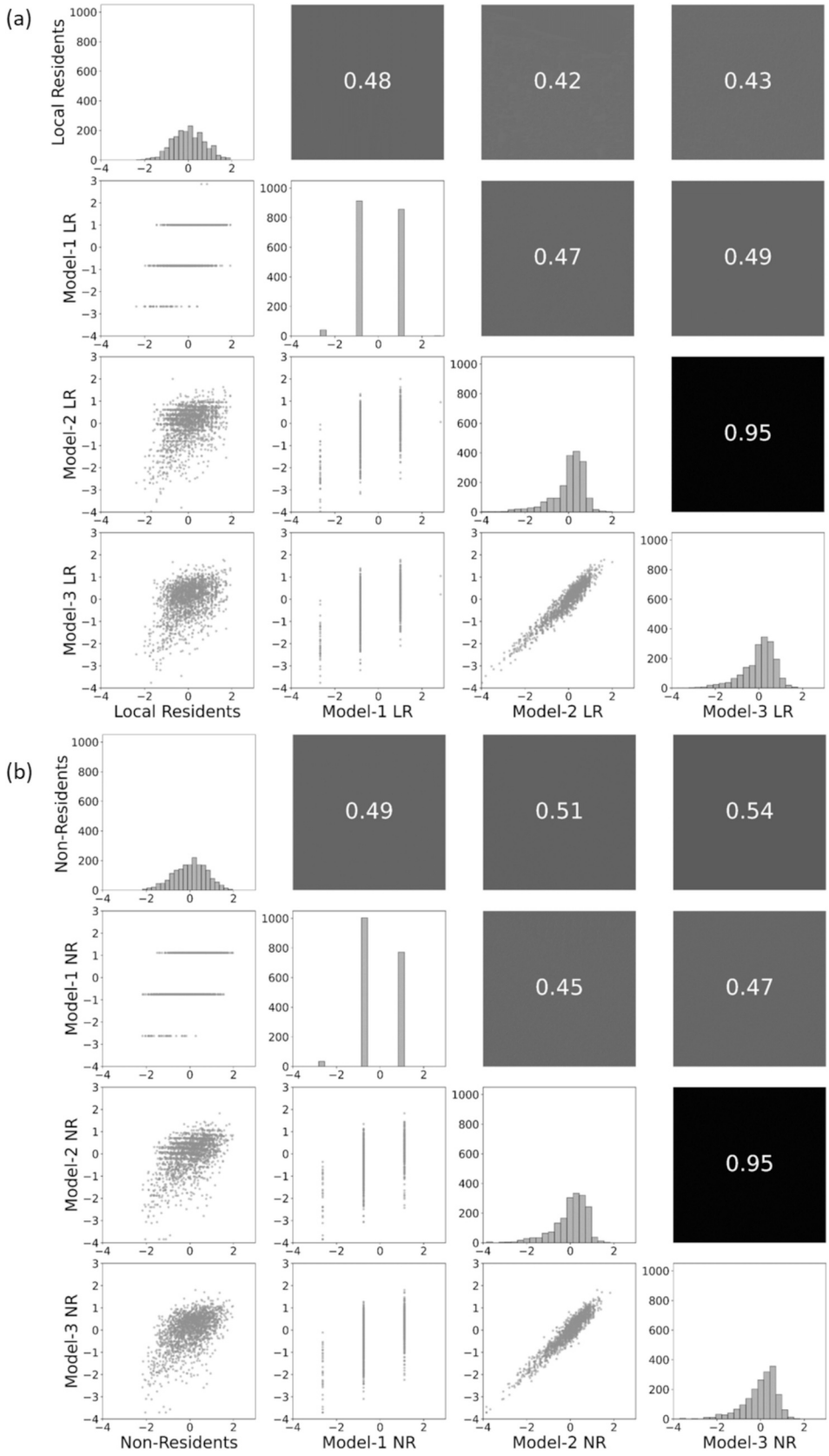

Fig. 2. Top half: Correlation heatmaps of Pearson's R values. Bottom half: Scatter plots. Diagonal: Histograms. The figure shows ratings derived from (a) GPT-4 LR models and local resident participants, and (b) GPT-4 NR models and non-resident participants.

leading to more varied ratings, while more complex models apply more criteria uniformly, resulting in more homogenised evaluations and increased spatial autocorrelation. Alternatively, the additional criteria in more complex models might neutralise each other, leading to moderate and more similar ratings. This is supported by the descriptive analysis in Table 2, in which the range of the middle 50 % of the data in Model-1 for both local residents and non-residents was higher than in Model-2 and Model-3. However, this analysis alone is insufficient to determine which interpretation is correct. To evaluate these views further, we also examined the spatial autocorrelation of differences between pairs of ratings (model-derived versus participant-derived) using Moran's I to evaluate the compatibility of these ratings with those of participants.

The analysis of spatial autocorrelation (Moran's I) of differences between GPT-4 model ratings and participant ratings revealed that local residents' ratings exhibit slightly lower spatial autocorrelation compared to non-residents' ratings, indicating greater spatial variation in differences between local residents' ratings and those from GPT-4 models. For local residents, spatial autocorrelation of differences increased from a Moran's I of 0.06 for Model-1 to 0.19 and 0.20 for Model-2 and Model-3, respectively. For non-residents, the spatial autocorrelation of differences was slightly higher overall, beginning with a Moran's I of 0.08 in Model-1 and rising to 0.21 and 0.23 in Model-2 and Model-3, respectively. Comparing the GPT-4 models, we could observe a significant increase in spatial autocorrelation from Model-1 to Model-2 and Model-3, with the latter two models showing almost three times

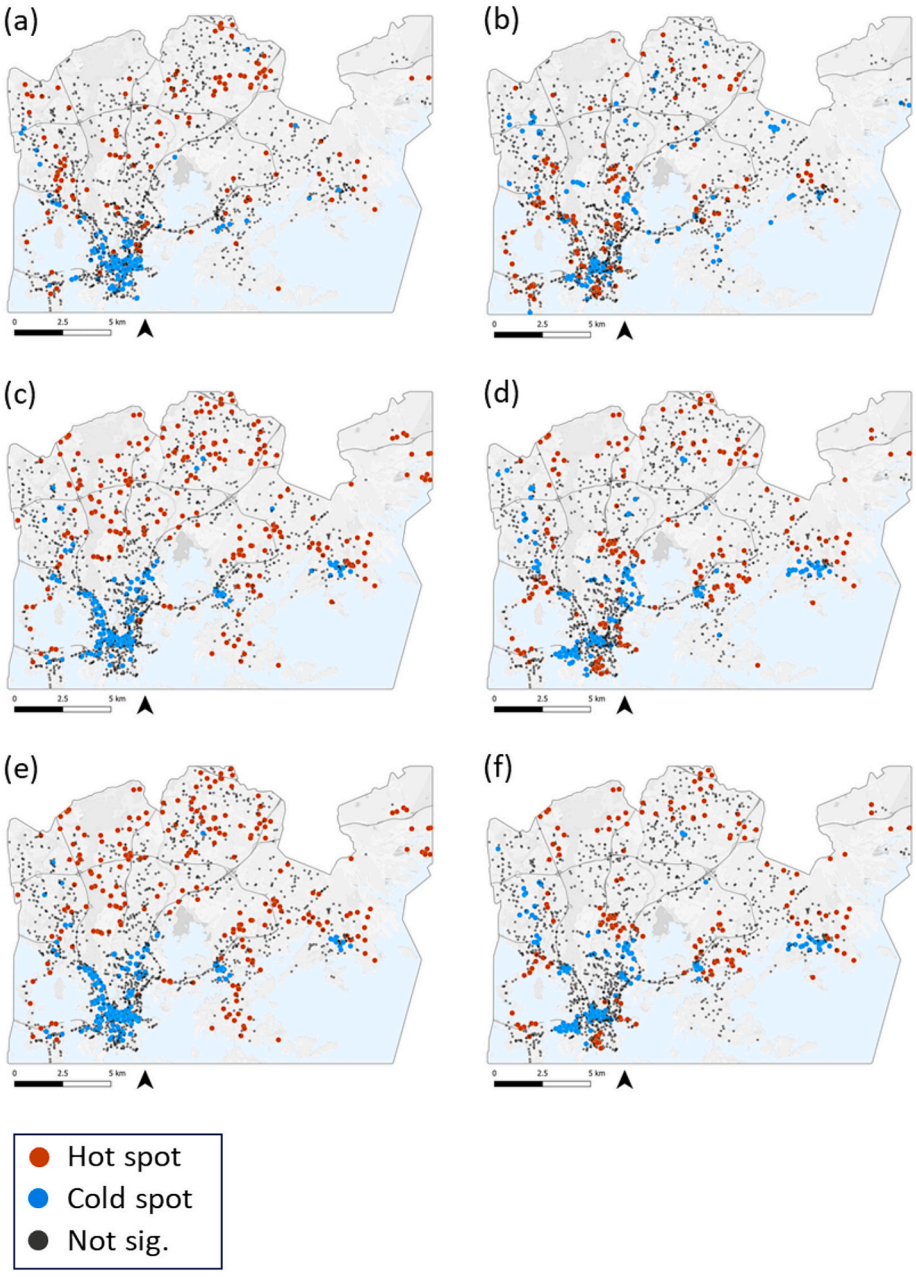

Fig. 3. Getis-Ord Gi hot spot analysis of the differences between GPT-4 model ratings and participant ratings (local residents and non-residents) for various models. Panels (a), (c), and (e) show the results for Model-1, Model-2, and Model-3 respectively, for local residents, while panels (b), (d), and (f) correspond to the same models for non-residents. Red dots represent significant hot spots, where GPT-4 model ratings are higher than participant ratings and are clustered. Blue dots represent significant cold spots where participant ratings are higher than GPT-4 model ratings and are clustered. Black dots indicate areas with no significant clustering. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

more spatial autocorrelation. This increase could be attributed to the specific criteria in Model-2 and Model-3, causing these models to treat areas that share partly similar characteristics uniformly.

To evaluate this further, we used Getis-Ord G* (Fig. 3) and Local Moran's I (Appendix F) analyses. The results showed a clear pattern of more hot spots in suburban and rural areas in which GPT-4 model ratings were higher than participant ratings, while the reverse pattern was observed in densely populated urban areas. Additionally, we found that simpler models and non-residents had fewer hot/cold spots, compared to more complex models and local residents, indicating lower local spatial autocorrelation in differences between simpler models and non-residents' ratings.

4.3. Association between semantic classes and GPT-4 outputs

The predominant features detected in the images were vegetation, road, sky, and buildings (Table 4). The proportion of pixels classified as person in the images was notably low, likely because the area a person occupies in an image is significantly smaller compared to other features, making it less prominent during semantic segmentation. Additionally, features like terrain and road transport appeared less frequently, with lower mean values. An example of the results from the semantic segmentation analysis is shown in Fig. 4.

The permutation feature importance (PFI) results reveal that in Model-1 LR, where GPT-4 was instructed to assume the perspective of a local resident, the model's decision-making was straightforward, with vegetation, road, and to a lesser extent, buildings having the highest degree of influence on the attractiveness score (Fig. 5). This suggests that the model, in this configuration, relied on a limited set of features to form its judgments. In Model-1 NR, where GPT-4 was prompted to consider the perspective of a non-resident, vegetation and road remained key factors, but the presence of people (feature: person) and sidewalks also gained importance, indicating a slight shift in focus based on the perceived perspective.

In Model-2 LR, all features except road, road transport, and terrain were relatively important, with vegetation being the most influential by a significant margin. It is particularly notable that, despite roads being the second most prominent feature in terms of proportion in the images, they exerted no significant influence on the ratings in the residents' model. This pattern suggests that even as the model's criteria are expanded, vegetation maintains a dominant role in determining the attractiveness. The results for Model-2 NR were similar, though with a reduced emphasis on vegetation and a slightly increased importance of roads.

In both Model-3 LR and NR, the results closely mirrored those of Model-2 s. This similarity suggests two potential interpretations. First, it may indicate that GPT-4 places greater emphasis on tangible criteria, enduring physical features, as the introduction of additional criteria did not significantly alter the feature importance. Alternatively, given that our analysis specifically investigates the relationship between physical features and attractiveness scores, it is expected that the introduction of additional criteria, urban design qualities and subjective reactions, would not substantially affect the association of these physical features with the scores.

Table 4
Mean and standard deviation (sd) of feature proportions detected in the images.

Feature	Mean (sd)
Road	0.20 (0.07)
Sidewalk	0.05 (0.03)
Building	0.12 (0.14)
Vegetation	0.31 (0.19)
Terrain	0.06 (0.05)
Sky	0.16 (0.1)
Person	0.0006 (0.001)
Road transport	0.02 (0.03)

The Accumulated Local Effects (ALE) plots reveal consistent patterns across different models for most features (Fig. 6). In the road class, Model-1 s exhibit a negative trend across the entire value range, whereas other models show minimal influence beyond the initial values. The sidewalk feature displays a generally positive and monotonic trend across all models, except for Model-1 LR. The initial presence of sidewalks increases attractiveness, but beyond 5 % of image coverage, the influence stabilizes and becomes negligible. For buildings, all models show heightened sensitivity to low building values (values near 0 % of image coverage), but once the proportion of buildings increases, the trend displays a slight positive slope.

Vegetation consistently demonstrates a strong, linear positive relationship with attractiveness across all models, reinforcing its dominant influence. In contrast, terrain shows minimal impact on attractiveness as evidenced by limited range of change in ALE, indicating its relative unimportance in GPT-4's assessment. The sky feature reveals a negative influence on attractiveness scores as its proportion increases in the image, except in Model-1 s, where its impact remains near zero. The presence of people in the images generally enhances attractiveness across all models, though the relationship is not strictly monotonic; notably, Model-1 NR shows an initial positive trend that declines before rising again. Lastly, road transport shows a weak negative trend with attractiveness, stabilizing near zero influence once it covers more than 2 % of the image.

5. Discussion

Attractive urban environments are associated with a range of positive impacts, including citizen satisfaction, well-being, and sustainable travel behaviours (Giles-Corti et al., 2016; Nieuwenhuijsen, 2020). Advances in artificial intelligence models, combined with visual data covering urban areas, have made it possible to analyse the urban attractiveness in unprecedented ways. However, the gap between advances in mapping and planning practice remains wide, due to the complexity and resource-intensity of the novel methods. In this study, we explored the applicability of an off-the-shelf AI model with simple text commands in producing reliable spatial estimates of urban attractiveness from Street View Imagery. We have also provided insights into the underlying logic behind how these AI-derived ratings are generated.

Our findings revealed a general alignment between the AI-generated ratings and human evaluations on urban attractiveness, but also notable contextual differences. GPT models generally assigned higher ratings to suburban areas, and lower ratings to densely populated urban areas, compared to the human participants. This discrepancy may arise from the models' inability to grasp the contextual and cultural nuances that humans use when evaluating urban environments. Humans might find densely-populated areas attractive due to their vibrant social and economic activities, which are integral to their daily lives and enhance the urban quality of life (Bardhan, Kurisu, & Hanaki, 2015), despite these areas being less green or visually complex. In Helsinki, GPT models underestimated the attractiveness, especially in the urban core with its dense population, active cultural and economic activity, and historical architecture, but less greenery. Conversely, suburban and rural areas, while often perceived as more attractive by the models due to their greenery and open spaces, might lack the dynamic elements that contribute to human satisfaction in urban settings, as the results indicated in Helsinki.

The analysis indicated that local residents exhibited lower spatial autocorrelation in their ratings compared to non-residents. This finding suggests that residents' evaluations are more diverse and influenced by their personal experiences and attachments to specific areas. In contrast, non-residents, who lack such personal connections, tend to rate urban areas based on more uniform and generalised criteria. This difference highlights how familiarity and a sense of place can shift evaluative criteria and perceptions (Soini, Vaarala, & Pouta, 2012). When living in

Fig. 4. Example of semantic segmentation analysis showing the classification of image features. Some features shown here were either eliminated or combined in our analysis after multicollinearity testing.

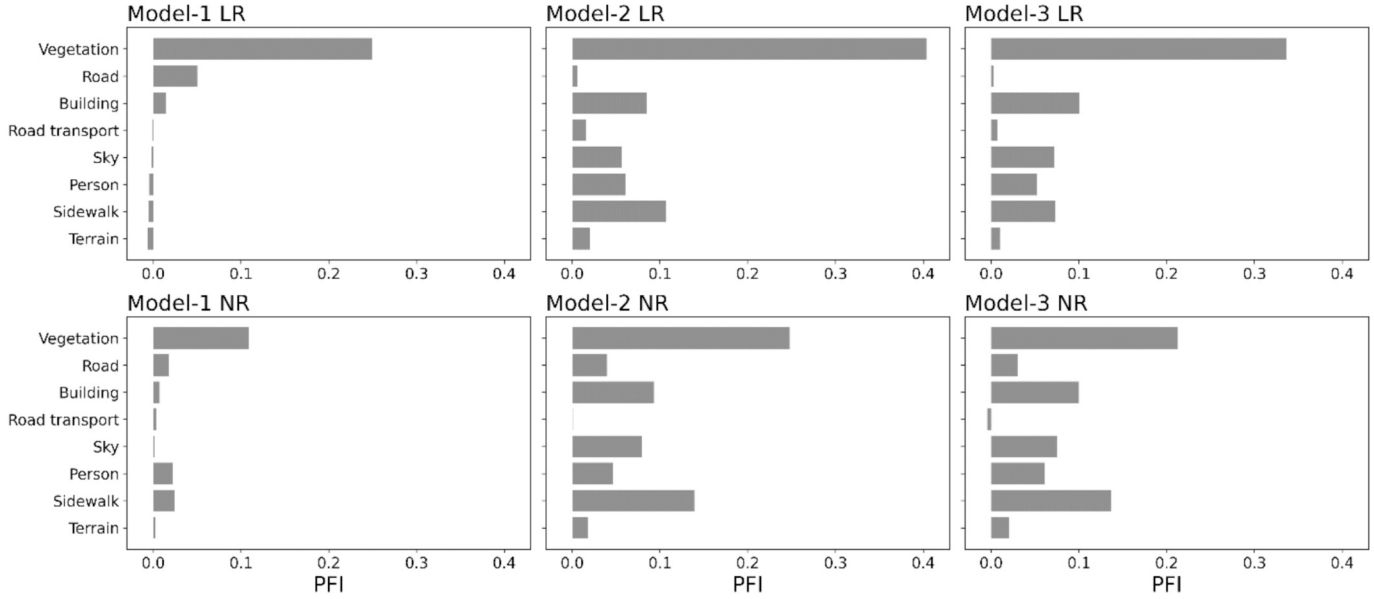

Fig. 5. Permutation feature importance (PFI) scores for each feature across various models (Model-1, Model-2, Model-3) for both Local Resident (LR) and Non-Resident (NR).

an area, the daily interactions and memories associated with specific locations play a significant role in shaping one's evaluation of those places.

The analysis of spatial autocorrelation, both globally and locally, along with the differences between GPT model ratings and participant ratings, indicates that simpler models (Model-1) exhibited lower spatial autocorrelation compared to more complex models (Model-2 and Model-3). This suggests that simpler models, which lack specific evaluative criteria in their prompts, tend to provide more random and less consistent ratings. It should be noted that the complexity of human perception, which encompasses emotional, cultural, and experiential factors, presents a significant challenge for AI models (Assunção, Patrão, Castelo-Branco, & Menezes, 2022). While current MLLMs like GPT-4 offer a promising start, they require guided prompting with certain criteria and/or fine-tuning to approach the depth of human evaluative processes.

Our efforts to demystify the non-transparent processes behind GPT-4's attractiveness scoring yielded insightful results. The strong association between certain image features, particularly vegetation, and the model's outputs suggests that semantic segmentation plays a crucial role in how GPT-4 assesses attractiveness. The observed trends, especially for highly influential features, align well with expected outcomes. However, it is important to recognize that semantic segmentation is likely only one part of a more complex process, which may also involve the analysis of spatial relationships between objects. Despite these findings, the full extent of GPT-4's decision-making process remains unclear. This ambiguity is particularly significant in light of existing research that highlights the challenges LLMs like GPT-4 face in spatial reasoning and understanding (Chen et al., 2024; Majumdar et al., 2024). Future studies should continue to explore these underlying mechanisms, aiming to enhance the transparency and interpretability of AI models.

An important reason for choosing off-the-shelf models was to democratise the use of AI for researchers and users who are not AI experts. These models eliminate the need for users to train the model, which often requires large amounts of training data and computational resources. Instead, these models offer a straightforward and accessible way to utilise AI without extensive technical expertise. While we initially considered using other models for the analysis, ChatGPT outperformed the alternatives. ChatGPT is widely accessible and familiar to many users (Bilgram & Laarmann, 2023), making our workflow more user-friendly and reproducible. By providing our prompts, we have enabled others to modify and experiment with the criteria easily, engaging with ChatGPT in a conversational manner. This accessibility is crucial for ensuring that advanced AI tools can be used broadly in urban planning and design without requiring extensive technical expertise.

Despite its advantages, the use of ChatGPT comes with limitations. It

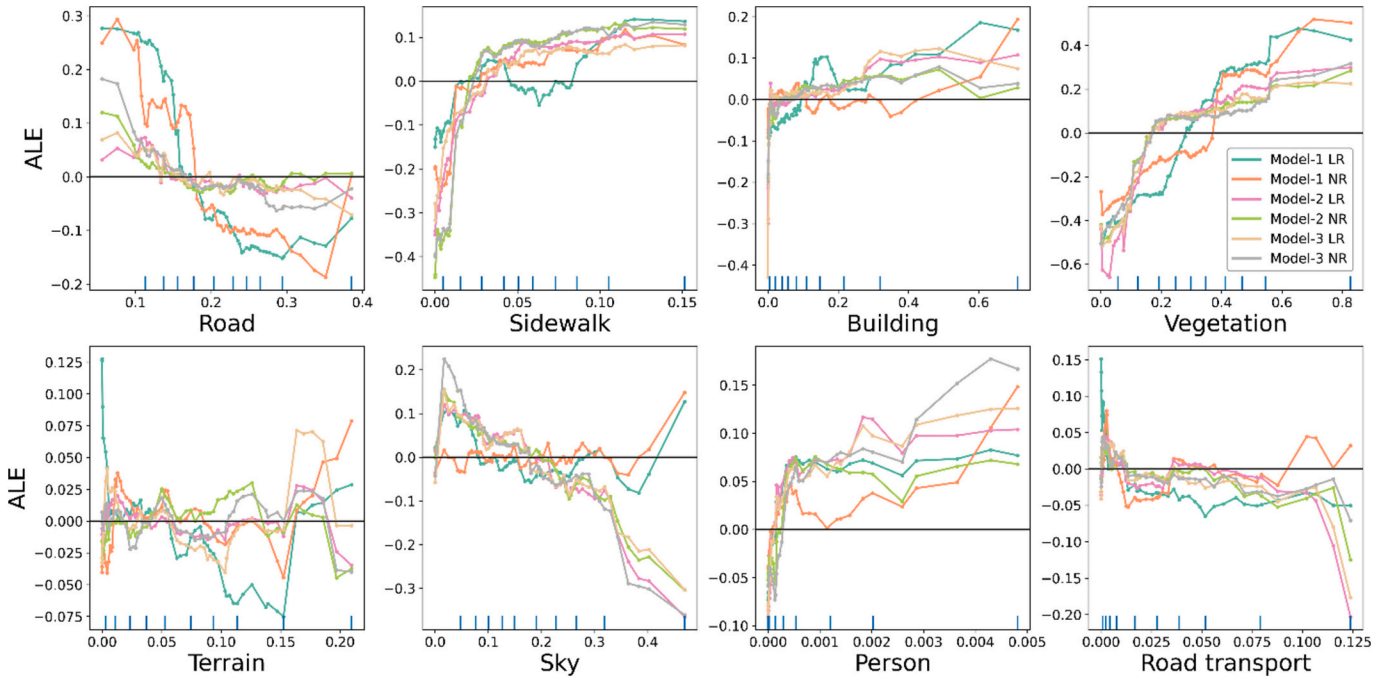

Fig. 6. Accumulated Local Effects (ALE) plots showing the influence of various features on GPT-4's attractiveness scores across various models (Model-1, Model-2, Model-3) for both Local Resident (LR) and Non-Resident (NR).

is proprietary model and has usage constraints, even in its premium version (Achiam et al., 2023). This can be a limiting factor for users needing to process numerous requests in a short period, as demonstrated in our workflow. Consequently, cities with extensive street view imagery, like Helsinki, would require either a sampling algorithm to optimise the number of images for assessment or the collection of their own street view images, thereby minimising the cost and time of API usage. It is important to note that the proprietary nature of models like GPT-4 presents significant challenges, as they function as "black boxes." While their ease of use and accessibility contribute to the democratization of AI, this non-transparency can be restrictive for researchers. Specifically, in our study, the inability to fully comprehend the decision-making processes and underlying mechanisms that GPT-4 employs to generate attractiveness scores is a clear limitation even if we took steps towards disclosing some of the inner logics behind the rating-producing mechanisms. The specific steps and criteria the model uses remain undisclosed and unclear. Nevertheless, despite these challenges, the high level of agreement between GPT-4's outputs and participant ratings demonstrates its practical utility and potential value. Future research should focus on developing open-access models tailored to urban landscape analysis, allowing users to fine-tune and adapt these models without associated costs.

An important limitation of this research is that the results cannot be fully compared to the actual experiences of people physically present in an area. While this workflow and method can highlight areas with high or low attractiveness, the ratings are based on images rather than the real-life experience of being in a space. Capturing a multisensory, real-life experience in a relatively low resolution 360-degree image is impossible. Real-life experiences involve a combination of visual, auditory, tactile, and even olfactory stimuli that images alone cannot convey (Bruce, Condie, Henshaw, & Payne, 2015; Gjerde, 2010; Thibaud, 2011). Consequently, the ratings generated from these images only correspond to a partial representation of the actual environment. It should be noted that all images were captured during the daytime. Although we assessed the impact of luminosity on ratings, variations in attractiveness perception between daytime and nighttime contexts remain unexplored in this study. Additionally, while images were taken across different seasons, we did not examine the effect of seasonality on perceived attractiveness. This is particularly relevant in a city like Helsinki, where seasonal changes significantly alter greenery (Klein, Willberg, Korpilo, & Toivonen, 2024) and other natural features, potentially impacting attractiveness assessments.

We also acknowledge that the primary aim of this study was to explore the potential of off-the-shelf AI models in evaluating urban attractiveness. Our focus was on capturing participants' immediate responses to images, without asking them to evaluate specific qualities or using various scales related to different features. Future studies could build on this by incorporating established tools and concepts from the literature to evaluate aesthetic qualities more comprehensively (Subiza-Pérez, Hauru, Korpela, Haapala, & Lehvävirta, 2019; Tveit, Ode, & Fry, 2006). Additionally, improvements could be made to the study design by refining the selection and ordering of images, increasing the number of participants, and ensuring a more diverse sample with different demographic backgrounds. Expanding the concept of residency to include different scales, such as residential or workplace neighbourhoods, could further enhance the robustness of the findings.

6. Conclusion

While AI models like GPT-4 offer significant potential for streamlining the evaluation of urban attractiveness, our study highlights the need to incorporate human perspectives to capture the full range of contextual and experiential nuances. AI models can serve as an increasingly valuable tool for preliminary assessments, identifying areas that may require further human investigation. This approach can reduce operational costs and time expended by urban planners and designers, providing a more efficient pathway to understanding urban environments. However, caution should be exercised when relying solely on AI models for policymaking decisions, especially in areas in which human experiences and perceptions play a crucial role. The results of AI assessments should complement, rather than replace, human evaluations. Future research should focus on enhancing AI models to mimic human perception better, by integrating more sophisticated and nuanced criteria.

Our work demonstrates the potential for AI to aid in the assessment of urban landscapes, but it also underscores the limitations of current technology. The ongoing development of AI models that can better understand and replicate human experiences will be critical for their effective application in urban planning. Existing survey materials could be leveraged to build a training pipeline, enhancing AI's ability to provide meaningful insights tailored to the local context. Ultimately, a hybrid approach that leverages both AI and human insights will provide the most comprehensive understanding of urban attractiveness, ensuring that planning and design decisions effectively promote healthy and satisfying living environments.

Ethics approval statement

This study did not require ethics approval.

Funding statement

This study is part of the GREENTRAVEL project (2023-2027) funded by the European Union (ERC, project 101044906). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work was also supported by the Finnish Ministry of Education and Culture's Pilot for Doctoral Programmes (Pilot project Mathematics of Sensing, Imaging and Modelling). Additionaly, Digital Geography Lab is supported by the University of Helsinki and the Flagship Program of Advanced Mathematics for Sensing, Imaging, and Modelling (FAME).

CRediT authorship contribution statement

Milad Malekzadeh: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Elias Willberg:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Jussi Torkko:** Writing – review & editing, Visualization, Software, Methodology, Formal analysis, Data curation. **Tuuli Toivonen:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

No potential conflict of interest was reported by the authors.

Acknowledgements

We are grateful for the valuable discussions and feedback provided by members of the Digital Geography Lab. We also extend our thanks to the reviewers for their constructive comments, which significantly improved our work. We sincerely thank our 24 participants who volunteered for this study, making this research possible. Special thanks go to Dr. Silviya Korpilo for her helpful comments and support. We are deeply grateful to Dr. Olle Järv for his expert guidance and thoughtful advice. Lastly, we acknowledge CSC – IT Center for Science and Geoportti, Finland, for providing the supercomputing resources used in our analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2024.102243.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). *Gpt-4 technical report. arXiv preprint arXiv:230308774*.

Agbo, A. A. (2010). Cronbach's alpha: Review of limitations and associated recommendations. *Journal of Psychology in Africa, 20*(2), 233–239.

Alexander, C. (1977). *A pattern language: Towns, buildings, construction*. Oxford University Press.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics, 26*(10), 1340–1347.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology, 82*(4), 1059–1086.

Arnold, H. F. (1980). *Trees in urban design*. Trees in urban design.

Assunção, G., Patrão, B., Castelo-Branco, M., & Menezes, P. (2022). An overview of emotion in artificial intelligence. *IEEE Transactions on Artificial Intelligence, 3*(6), 867–886.

Bardhan, R., Kurisu, K., & Hanaki, K. (2015). Does compact urban forms relate to good quality of life in high density cities of India? Case of Kolkata. *Cities [Internet], 48*, 55–65. Available from: https://www.sciencedirect.com/science/article/pii/S026427511500089X.

Bilgram, V., & Laarmann, F. (2023). Accelerating innovation with generative AI: AI-augmented digital prototyping and innovation methods. *IEEE Engineering Management Review, 51*(2), 18–25.

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning, 215*, Article 104217.

Birenboim, A. (2018). The influence of urban environments on our subjective momentary experiences. *Environment and Planning B: Urban Analytics and City Science, 45*(5), 915–932.

Boivin, M., & Tanguay, G. A. (2019). Analysis of the determinants of urban tourism attractiveness: The case of Québec City and Bordeaux. *Journal of Destination Marketing & Management [Internet]., 11*, 67–79. Available from: https://www.sciencedirect.com/science/article/pii/S2212571X16303560.

Broitman, D., Kourtit, K., Kourtit, K., Neuts, B., Nijkamp, P., & Wahlström, M. H. (2020 Dec 15). A structural equation model for place-based city love: An application to Swedish cities. *International Regional Science Review [Internet]., 44*(3–4), 432–465. Available from: https://doi.org/10.1177/0160017620979638.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Bruce, N., Condie, J., Henshaw, V., & Payne, S. R. (2015). Analysing olfactory and auditory sensescapes in English cities: Sensory expectation and urban environmental perception. *Ambiances Environnement Sensible, Architecture et Espace Urbain.*

Carver, A., Timperio, A., & Crawford, D. (2008). Playing it safe: The influence of neighbourhood safety on children's physical activity—A review. *Health Place [Internet]., 14*(2), 217–227. Available from: https://www.sciencedirect.com/science/article/pii/S1353829207000536.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., et al. (2024). Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14455–14465).

Chen, L., Lu, Y., Sheng, Q., Ye, Y., Wang, R., & Liu, Y. (2020). Estimating pedestrian volume using Street View images: A large-scale validation test. *Computers, Environment and Urban Systems [Internet]., 81*, Article 101481. Available from: https://www.sciencedirect.com/science/article/pii/S0198971519304351.

Clifton, K. J., Livi Smith, A. D., & Rodriguez, D. (2007). The development and testing of an audit for the pedestrian environment. *Landscape and Urban Planning [Internet]., 80*(1), 95–110. Available from: https://www.sciencedirect.com/science/article/pii/S0169204606001101.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3213–3223).

Craig, C. L., Brownson, R. C., Cragg, S. E., & Dunn, A. L. (2002). Exploring the effect of the environment on physical activity: A study examining walking to work. *American Journal of Preventive Medicine [Internet]., 23*(2, Supplement 1), 36–43. Available from: https://www.sciencedirect.com/science/article/pii/S0749379702004725.

Day, K., Boarnet, M., Alfonzo, M., & Forsyth, A. (2006). The Irvine–Minnesota inventory to measure built environments: development. *American Journal of Preventive Medicine, 30*(2), 144–152.

Devlin, J., Chang, M. W., Lee, K., & Bert, T. K. (2018). *Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805*.

Emery, J., Crump, C., & Bors, P. (2003). Reliability and validity of two instruments designed to assess the walking and bicycling suitability of sidewalks and roads. *American Journal of Health Promotion, 18*(1), 38–46.

Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.

Evans, G. W., Smith, C., & Pezdek, K. (1982). Cognitive maps and urban form. *Journal of the American Planning Association, 48*(2), 232–244.

Ewing, R., Clemente, O., Neckerman, K. M., Purciel-Hill, M., Quinn, J. W., & Rundle, A. (2013). *Measuring urban design: Metrics for livable places* (Vol. 200). Springer.

Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *The Journal of Urban Design (Abingdon)., 14*(1), 65–84.

Forsyth, A. (2015). What is a walkable place? The walkability debate in urban design. *URBAN DESIGN International [Internet]., 20*(4), 274–292. https://doi.org/10.1057/udi.2015.22

Gehl, J. (1987). *Life between buildings: Using public space*. Island Press.

Giles-Corti, B., & Donovan, R. J. (2003). Relative influences of individual, social environmental, and physical environmental correlates of walking. *American Journal of Public Health, 93*(9), 1583–1589.
Giles-Corti, B., Vernez-Moudon, A., Reis, R., Turrell, G., Dannenberg, A. L., Badland, H., et al. (2016). City planning and population health: A global challenge. *The Lancet, 388*(10062), 2912–2924.
Gjerde, M. (2010). Visual aesthetic perception and judgement of urban streetscapes. In *Paper for building a better world: CIB world congress* (pp. 12–22). Citeseer.
Gong, F. Y., Zeng, Z. C., Ng, E., & Norford, L. K. (2019). Spatiotemporal patterns of street-level solar radiation estimated using Google Street View in a high-density urban environment. *Building and Environment [Internet]., 148*, 547–566. Available from: https://www.sciencedirect.com/science/article/pii/S0360132318306437.
Hamim, O. F., & Ukkusuri, S. V. (2024). Towards safer streets: A framework for unveiling pedestrians' perceived road safety using street view imagery. *Accident Analysis & Prevention [Internet]., 195*, Article 107400. Available from: https://www.sciencedirect.com/science/article/pii/S0001457523004475.
Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., et al. (2024). Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems, 36*.
Hutto, C., & Gilbert, E. V. (2014). A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (pp. 216–225).
Jacobs, A. B. (1993). *Great streets*. University of California Transportation Center.
Jacobs, J. (1992). *The death and life of great American cities. 1961* (p. 321). New York: Vintage, 9783839413272–099.
de Jong, T., & Fyhri, A. (2023). Spatial characteristics of unpleasant cycling experiences. *The Journal of Transport Geography [Internet]., 112*, Article 103646. Available from: https://www.sciencedirect.com/science/article/pii/S0966692323001187.
Jongwiriyanurak, N., Zeng, Z., Wang, M., Haworth, J., Tanaksaranond, G., & Boehm, J. (2023). Framework for motorcycle risk assessment using onboard panoramic camera (short paper). In *12th international conference on geographic information science (GIScience 2023). Schloss Dagstuhl-Leibniz-Zentrum für Informatik*.
Keralis, J. M., Javanmardi, M., Khanna, S., Dwivedi, P., Huang, D., Tasdizen, T., et al. (2020). Health and the built environment in United States cities: Measuring associations using Google Street View-derived indicators of the built environment. *BMC Public Health [Internet]., 20*(1), 215. Available from: https://doi.org/10.1186/s12889-020-8300-1.
Khisty, C. J. (1994). Evaluation of pedestrian facilities: Beyond the level-of-service concept. *Transportation Research Record, 45*.
Ki, D., Chen, Z., Lee, S., & Lieu, S. (2023). A novel walkability index using google street view and deep learning. *Sustainable Cities and Society [Internet]., 99*, Article 104896. Available from: https://www.sciencedirect.com/science/article/pii/S2210670723005073.
Klein, R., Willberg, E., Korpilo, S., & Toivonen, T. (2024). Temporal variation in travel greenery across 86 cities in Europe. *Urban Forestry & Urban Greening, 102*, Article 128566.
Landis, B. W. (1994). Bicycle interaction hazard score: A theoretical model. *Transportation Research Record, 1438*, 3–8.
Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730–19742). PMLR.
Li, X., Li, Y., Jia, T., Zhou, L., & Hijazi, I. H. (2022). The six dimensions of built environment on urban vitality: Fusion evidence from multi-source data. *Cities [Internet]., 121*, Article 103482. Available from: https://www.sciencedirect.com/science/article/pii/S0264275121003814.
Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry and Urban Greening [Internet]., 14*(3), 675–685. Available from: https://www.sciencedirect.com/science/article/pii/S1618866715000874.
Li, Y., Yabuki, N., & Fukuda, T. (2022). Measuring visual walkability perception using panoramic street view images, virtual reality, and deep learning. *Sustainable Cities and Society [Internet]., 86*, Article 104140. Available from: https://www.sciencedirect.com/science/article/pii/S221067072200453X.
Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in Neural Information Processing Systems, 36*.
Liu, J., Ettema, D., & Helbich, M. (2023). Street view environments are associated with the walking duration of pedestrians: The case of Amsterdam, the Netherlands. *Landscape and Urban Planning [Internet]., 235*, Article 104752. Available from: https://www.sciencedirect.com/science/article/pii/S0169204623000713.
Liu, L., & Sevtsuk, A. (2024). Clarity or confusion: A review of computer vision street attributes in urban studies and planning. *Cities [Internet]., 150*, Article 105022. Available from: https://www.sciencedirect.com/science/article/pii/S0264275124002361.
Liu, L., Silva, E. A., Wu, C., & Wang, H. (2017). A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Computers, Environment and Urban Systems [Internet]., 65*, 113–125. Available from: https://www.sciencedirect.com/science/article/pii/S0198971516301831.
Liu, X., Haworth, J., & Wang, M. (2023). A new approach to assessing perceived walkability: Combining street view imagery with multimodal contrastive learning model. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on spatial big data and AI for industrial applications* (pp. 16–21).
Lynch, K. (1960). The image of the environment. *The image of the city, 11*, 1–13.
Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., et al. (2024). Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16488–16498).
Mäki, N., & Sinkko, H. (2022). Helsingin väestö vuodenvaihteessa 2021/2022 ja väestönmuutokset vuonna 2021 [Internet]. Helsinki [cited 2024 May 28]. Available from: https://www.hel.fi/hel2/tietokeskus/julkaisut/pdf/23_01_10_Tilastoja_7_Maki_Sinkko.pdf.
Mayne, S. L., Jose, A., Mo, A., Vo, L., Rachapalli, S., Ali, H., et al. (2018). Neighborhood disorder and obesity-related outcomes among women in Chicago. *International Journal of Environmental Research and Public Health, 15*(7), 1395.
Mooney, S. J., Wheeler-Martin, K., Fiedler, L. M., LaBelle, C. M., Lampe, T., Ratanatharathorn, A., et al. (2020). Development and validation of a Google Street View pedestrian safety audit tool. *Epidemiology, 31*(2), 301–309.
Moudon, A. V., & Lee, C. (2003). Walking and bicycling: An evaluation of environmental audit instruments. *American Journal of Health Promotion, 18*(1), 21–37.
Moura, F., Cambra, P., & Gonçalves, A. B. (2017). Measuring walkability for distinct pedestrian groups with a participatory assessment method: A case study in Lisbon. *Landscape and Urban Planning [Internet]., 157*, 282–296. Available from: https://www.sciencedirect.com/science/article/pii/S0169204616301268.
Nagata, S., Nakaya, T., Hanibuchi, T., Amagasa, S., Kikuchi, H., & Inoue, S. (2020). Objective scoring of streetscape walkability related to leisure walking: Statistical modeling approach with semantic segmentation of Google Street View images. *Health Place [Internet]., 66*, Article 102428. Available from: https://www.sciencedirect.com/science/article/pii/S1353829220302720.
Nieuwenhuijsen, M. J. (2020). Urban and transport planning pathways to carbon neutral, liveable and healthy cities; A review of the current evidence. *Environment International [Internet]., 140*, Article 105661. Available from: https://www.sciencedirect.com/science/article/pii/S0160412020302038.
Norppa, M., & Kaunis, H. H. (2020). vihreä ja rauhallinen jalan kaupunginosissa: Asukaskyselyn tulokset [Internet]. Helsinki [cited 2024 May 28]. Available from: https://ahjojulkaisu.hel.fi/9ABAF725-8151-C780-95A2-7B7346000000.pdf.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., et al. (2023). *Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv: 230614824*.
Pikora, T. J., Bull, F. C. L., Jamrozik, K., Knuiman, M., Giles-Corti, B., & Donovan, R. J. (2002). Developing a reliable audit instrument to measure the physical environment for physical activity. *American Journal of Preventive Medicine [Internet]., 23*(3), 187–194. Available from: https://www.sciencedirect.com/science/article/pii/S0749379702004981.
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.
Rapoport, A. (2013). *History and precedent in environmental design*. Springer Science & Business Media.
Saelens, B. E., & Handy, S. L. (2008). Built environment correlates of walking: A review. *Medicine and Science in Sports and Exercise, 40*(7 Suppl), S550.
Soini, K., Vaarala, H., & Pouta, E. (2012). Residents' sense of place and landscape perceptions at the rural–urban interface. *Landscape and Urban Planning [Internet]., 104*(1), 124–134. Available from: https://www.sciencedirect.com/science/article/pii/S0169204611002908.
St-Louis, E., Manaugh, K., van Lierop, D., & El-Geneidy, A. (2014). The happy commuter: A comparison of commuter satisfaction across modes. *Transportation Research Part F: Traffic Psychology and Behaviour [Internet]., 26*, 160–170. Available from: https://www.sciencedirect.com/science/article/pii/S1369847814001107.
Subiza-Pérez, M., Hauru, K., Korpela, K., Haapala, A., & Lehvävirta, S. (2019). Perceived Environmental Aesthetic Qualities Scale (PEAQS) – A self-report tool for the evaluation of green-blue spaces. *Urban Forestry and Urban Greening [Internet]., 43*, Article 126383. Available from: https://www.sciencedirect.com/science/article/pii/S1618866719301359.
Sun, Q.(C.), Macleod, T., Both, A., Hurley, J., Butt, A., & Amati, M. (2021). A human-centred assessment framework to prioritise heat mitigation efforts for active travel at city scale. *Science of The Total Environment [Internet]., 763*, Article 143033. Available from: https://www.sciencedirect.com/science/article/pii/S0048969720365633.
Talen, E. (2002). Pedestrian access as a measure of urban quality. *Planning Practice and Research, 17*(3), 257–278.
Tang, J., & Long, Y. (2019). Measuring visual quality of street space and its temporal variation: Methodology and its application in the Hutong area in Beijing. *Landscape and Urban Planning [Internet]., 191*, Article 103436. Available from: https://www.sciencedirect.com/science/article/pii/S0169204618310119.
Thibaud, J. P. (2011). The sensory fabric of urban ambiances. *The Senses and Society., 6*(2), 203–215.
Tveit, M., Ode, Å., & Fry, G. (2006 Jul 1). Key concepts in a framework for analysing visual landscape character. *Landscape Research [Internet]., 31*(3), 229–255. Available from: https://doi.org/10.1080/01426390600783269.
de Vet, H. C. W., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman–Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology, 85*, 45–49.
Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., et al. (2023). Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14408–14419).
van Wee, B., & Ettema, D. (2016). Travel behaviour and health: A conceptual model and research agenda. *The Journal of Transport & Health [Internet]., 3*(3), 240–248. Available from: https://www.sciencedirect.com/science/article/pii/S2214140516302006.

Wimbardana, R., Tarigan, A. K., & Sagala, S. (2018). Does a pedestrian environment promote walkability? Auditing a pedestrian environment using the pedestrian environmental data scan instrument. *Journal of Regional and City Planning, 29*, 57–66.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 418–434).

Ye, Y., Richards, D., Lu, Y., Song, X., Zhuang, Y., Zeng, W., et al. (2019). Measuring daily accessed street greenery: A human-scale approach for informing better urban planning practices. *Landscape and Urban Planning [Internet]., 191*, Article 103434. Available from: https://www.sciencedirect.com/science/article/pii/S0169204618309940.

Zhang, D., Yu, Y., Li, C., Dong, J., Su, D., Chu, C., et al. (2024). *Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:240113601*.

Zhang, N., Zheng, X., & Wang, X. (2022). Assessment of aesthetic quality of urban landscapes by integrating objective and subjective factors: A case study for riparian landscapes. *Frontiers in Ecology and Evolution, 9*, Article 735905.