

Enhancing Urban Walkability Assessment with Multimodal Large Language Models

Ivan Blečić[1(✉)], Valeria Saiu[1], and Giuseppe A. Trunfio[2]

[1] Department of Civil and Environmental Engineering and Architecture, University of Cagliari, Cagliari, Italy
{ivanblecic,v.saiu}@unica.it

[2] Department of Biomedical Sciences, University of Sassari, Sassari, Italy
trunfio@uniss.it

Abstract. Recent advances in computer vision and Convolutional Neural Networks (CNNs) have facilitated the use of street view imagery (SVI) for the automatic assessment of physical and perceived attributes of walkable areas. However, these methods still overlook the broader urban context and fail to capture and communicate to the user the qualitative factors influencing the assessed walkability score. This paper addresses these challenges by leveraging a Multimodal Large Language Model (MLLM) to provide a holistic assessment of walkability, consisting of both quantitative scores and linguistic qualitative insights. This approach offers a more comprehensive understanding of the factors contributing to the walkability score attributed to the image and enhances the interpretability and practical applicability of the assessments for urban planners and policymakers. Preliminary experiments demonstrate that a MLLM-based methodology can effectively capture a diverse range of factors of walkability, suggesting a promising direction for future developments of evaluation tools aimed at supporting the design of pedestrian-friendly environments.

Keywords: walkability assessment · machine learning · deep learning · multimodal large language model · street view

1 Introduction

Walkability, the measure of how conducive an environment is to pedestrian activity, has attracted significant attention in urban planning and public health research [1–4]. Numerous studies have investigated the impact on walking behaviour of the physical attributes of neighbourhoods, including factors such as residential density, land use mix, street connectivity, public transport accessibility, accessible amenities, highlighting a positive correlation between favourable neighbourhood design features and increased pedestrian activity [5,6]. Other key factors influencing pedestrian behaviour are those related to the physical walking environment, including the specific pedestrian-friendly infrastructures, which

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14819, pp. 394–411, 2024.
https://doi.org/10.1007/978-3-031-65282-0_26

determine pedestrian comfort and safety and which, taken together, account for the so-called *perceived walkability* [7–9].

Traditionally, experts have relied on surveys or time-consuming on-site assessments to identify key factors influencing perceived walkability. However, recent advancements in software and hardware technologies enabled novel approaches, largely based on computer vision methodologies applied to street view imagery (SVI), which can be nowadays considered as an important data source for urban analytics [10]. The proprietary Google Street View (GSV) service is the most well-known and widespread service providing SVI. However, in some areas it is also available Baidu Maps (https://map.baidu.com/) and a notable trend is that of crowdsourced street view imagery, facilitated by services such as Mapillary [11], KartaView (https://kartaview.org), in some cases providing adequate geographical coverage and temporal granularity, at a permissive licence.

According to the recent literature, semantic segmentation based on Convolutional Neural Networks (CNN) is one of the techniques that has been applied to SVI and has emerged as a promising method for assessing the physical walking environment. For example, in [12], Baidu street view images were used to measure visual walkability in Shenzhen, creating indicators on "psychological greenery" and "visual crowdedness". In [13] the authors developed and index based on semantic segmentation incorporating factors like greenery and street width. Similarly, in [14] the authors analysed street view images observing variations in object importance at different spatial scales. While these approaches offer new insights into walkability assessment, challenges remain, particularly regarding the accuracy and analysis of semantic segmentation results from SVI and the difficulties in establishing an effective model connecting segmented entities and perceived walkability.

In addition to automatically detecting the specific physical attributes of walkable areas, researchers have also delved into direct modelling of people's perceptions of walkability based on SVI. Starting from the Place Pulse dataset [15–17], created by MIT's Media Lab, which was used as a valuable resource for understanding public perceptions of urban environments through street view images, several deep learning models based on CNNs, have been developed to predict perceptual properties based on training sets of rated SVIs [12,14,18]. For example, in [19] we have developed an approach based on transfer learning in order to build a CNN, pre-trained on the Places dataset with 365 scene categories, capable to classify an input image from GSV into one of the five ordinal categories of walking suitability ranging from 1 to 5, where 1 corresponds to low and 5 to high walkability. The resulting network was trained using a relatively small dataset composed of rated images, with the objective of predicting the score to be assigned to any other image [19,20].

While the CNN-based approaches have proven effective in identifying physical features such as green spaces, street widths, and pedestrian infrastructure, they face limitations. Specifically, approaches based on semantic segmentation often result in a narrow focus on detectable elements without capturing the broader, more synergistic and subjective aspects of walkability. These include overall aes-

thetic appeal of an area, perceived safety, and the ambiance that contributes significantly to pedestrian subjective perception. Furthermore, the CNN-based approaches that aim at modelling the overall perceived walkability lack the ability to provide contextual insights or explanations for their assessments, making it difficult for urban planners and policymakers to understand the underlying "reasoning" and factors that have contributed to or detracted from a machine assessment of walkability.

In recent years, the field of deep learning has seen significant advancements with the introduction of Large Language Models (LLMs) and their subsequent evolution into Multimodal Large Language Models (MLLMs). LLMs have demonstrated remarkable abilities in understanding and generating text, paving the way for applications that require nuanced interpretation of complex data. Building on this, MLLMs extend these capabilities to include the analysis of imagery alongside textual data, offering a more holistic approach to understanding and interacting with the world. These models can integrate and process information across different modalities, providing a richer, more comprehensive understanding of content.

Given these advancements, our study started investigating an application of MLLMs to the domain of urban walkability assessment. This approach aims to overcome the limitations of previous CNN-based methods by not only analysing physical features visible in street view images but also interpreting the broader urban context that influences pedestrian experiences. By leveraging the multimodal processing capabilities of MLLMs, we can extract deeper insights into the factors that contribute to walkability. This methodology could enable a more nuanced analysis that combines quantitative assessments with qualitative evaluations, providing urban planners and researchers with a comprehensive understanding of walkability that aligns more closely with human perceptions and experiences.

Furthermore, our approach emphasizes the importance of explainability in AI-driven assessments. By generating interpretations and explanations for the walkability scores, we aim to provide insights into the specific factors that influence these evaluations. This not only increases the transparency but also enhances the utility of the assessments for informing urban planning and policy decisions. Throu gh this innovative application of MLLMs, we seek to contribute to the development of urban spaces that are not only physically conducive to walking but also enrich the overall quality of life for their inhabitants.

The paper is organized as follows. Next section provides a short background on MLLMs. In Sect. 3, we provide relevant details of the experimental setup, including utilized models, datasets and processes. Section 4 illustrates and discusses the obtained preliminary results. Section 5 concludes the paper with some final considerations and hypotheses on future work.

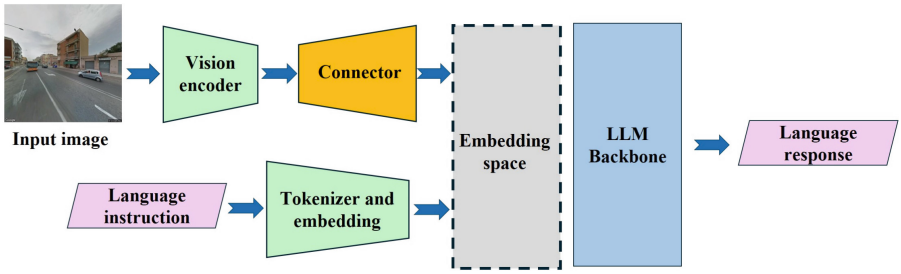


Fig. 1. An example of MLLM architecture.

2 Background

Recently, machine learning technologies featuring LLMs [21] have showcased superior performance in generating coherent and contextually appropriate text, advancing the field of natural language processing (NLP). However, a fundamental constraint of traditional LLMs is their unimodal nature, limiting them to text-based inputs and interactions. This is in contrast to the multimodal approach of human cognition that combines multiple sensory inputs for understanding the environment.

In response to the limitations of unimodal LLMs, recent research has focused on MLLMs, representing a significant enhancement. In general, MLLMs are able to integrate and process information across various data modalities, primarily text and images, expanding the advantages provided by LLMs to include understanding, interpretation, and generation of multimodal content. This advancement is indicative of progress towards artificial general intelligence, with the goal of mimicking the human capability to understand complex concepts through the integration of diverse sensory data.

Generally, MLLMs are characterized by quite complex architectures designed to manage multimodal data. These models include a reliable LLM, which is utilized for its deep linguistic comprehension and generative ability. An example is shown in Fig. 1 where an input image and a language instruction are processed by a vision encoder and a tokenizer respectively, which transform the raw data into a format the system can work with. These processed inputs converge in a connector that fuses the visual and textual information into a single embedding space. This combined data is then fed into the Large Language Model (LLM) backbone, where it is interpreted to generate a coherent language response, encapsulating the insights drawn from both the image and the accompanying text. This architecture permits MLLMs to comprehend multimodal content, namely images and text, producing textual output and represents a substantial step towards AI systems capable of understanding and communicating with the world in a way similar to human cognition [22].

Training approaches for MLLMs have evolved significantly and utilize a combination of techniques to effectively integrate information from various

modalities. These include Multimodal Pre-training (M-PT), Multimodal Instruction Tuning (M-IT), Multimodal In-Context Learning (M-ICL), Multimodal Chain of Thought (M-CoT), and LLM-Aided Visual Reasoning (LAVR) [22]. These methods enable the integration of visual elements with traditional text-based processing in LLMs. For example, M-IT involves fine-tuning pre-trained LLMs with datasets formatted with instructions that include both textual and visual inputs, allowing models to understand and execute tasks requiring information synthesis from multiple modalities. Techniques such as M-ICL and M-CoT further extend LLMs' in-context learning and reasoning capabilities to multimodal domains, facilitating complex reasoning over tasks involving visual components. M-PT utilizes extensive datasets with paired data from different modalities to teach the model the complex relationships between them [22].

The integration of multimodality in LLMs through MLLMs has vast implications across various applications. This includes improvements in content creation and media generation, as well as transforming interactions with digital assistants. In the study described in this paper, we present some results of applying a MLLM able to process images and text to the task of evaluating walkability in urban contexts.

3 Methodology

Using a pre-trained open-source MLLM, we investigated the evaluation of urban walkability from street-level images, aiming to develop an automated tool that can provide insights similar to those of human experts in urban design and walkability. This approach aims at exploiting effectively the new capabilities of machine learning to analyse visual data, potentially transforming how urban spaces are evaluated and improved for pedestrian friendliness. Our methodology encompassed prompt optimization, expert validation, and a scoring system to ensure both accuracy and relevance in our findings.

It is worth noting that one key difference between the approach discussed in this paper and the computer vision methods mentioned in Sect. 1, lies in its ability to provide both a quantitative score and a qualitative explanation for each urban walkability assessment. In fact, unlike standard computer vision techniques that typically output scores or classifications without context, the MLLM-based methodology under investigation generates linguistic explanations alongside scores. This dual output offers insights into the factors influencing each score, thereby enhancing interpretability and trust in the model's evaluations. This aspect is closely aligned with the principles of explainable artificial intelligence (XAI), which emphasizes the importance of making AI decisions transparent and understandable to humans.

In this preliminary study, we relied on prompt optimization without computationally expensive MLLM fine-tuning. This process inherently is based on the model's zero-shot learning capabilities, hypothesizing its potential to assess walkability despite not being explicitly trained on such evaluations. Zero-shot learning represents the model's ability to apply its generalized understanding to

novel tasks, offering a promising avenue for analysing complex urban environments through the lens of walkability without the need for task-specific training data.

Our evaluation of the approach initiated with experts reviewing the model-generated scores and comments for accuracy and relevance. This preliminary phase, involving experts of urban design and walkability, served to validate the model's assessments against professional standards, ensuring the reliability of subsequent automated evaluations.

A subsequent experiment consisted in attributing a walkability score to the streets of Cagliari, a city of about 150,000 inhabitants, and the capital of Sardinia island in Italy. For the automatic walkability assessment we had already collected a dataset of about 30,000 Google Street View photos relative to the city street network. For each edge of the city-wide street network graph we had a number of photos ranging from two to six, depending on the edge's length.

For streets represented by multiple images, we employed a two-stage scoring process. Initially, images from identical viewpoints were aggregated via majority voting, a method that ensured robustness against outliers but might reduce the diversity of perspectives considered. Subsequently, for a comprehensive street evaluation, scores related to different points of the same street were averaged, and a LLM was used to summarize the different comments earlier produced by the MLLM, providing a unified street-level insight. This method balanced detailed analysis and practical scalability.

3.1 The Adopted MLLM: LLaVA

The Language and Vision Assistant (LLaVA) model [23–25] represents a recent and significant advance in the field of MLLMs, primarily focused on visual instruction tuning. LLaVA stands out for its data efficiency and simplicity in architecture. Utilizing a fully-connected vision-language cross-modal connector LLaVA's code and model have been made publicly available, underscoring the team's commitment to open-source advancements in the field.

The LLaVA model's architecture is intuitively simple, consisting of a pre-trained visual backbone to encode visual features, a pre-trained large language model for understanding user instructions and generating responses, and a vision-language cross-modal connector to align the outputs of the vision encoder with the language model. Notably, LLaVA achieves this with minimal architectural complexity and training data, unlike other models that require extensive pretraining on large-scale datasets.

Our approach is based on the latest version named LLaVA-1.6 [25], which introduces enhancements in reasoning, OCR, and world knowledge over its predecessor, LLaVA-1.5. In particular, it features a 4x increase in input image resolution, supporting varied aspect ratios, and improvements in visual instruction tuning data mixture for better OCR and reasoning capabilities.

More in detail, for our experiments we used the available pre-trained LLaVA-1.6 with Mistral-7B LLM backbone [26]. All runs were executed on a consumer-level GPU card Nvidia GTX 4090.

Mistral-7B is a LLM developed by Mistral AI and is equipped with 7.3 billion parameters. According to [26], this model surpasses the performance of Llama 2 with 13 billion parameters, in all tests, and it even outperforms Llama 1 with 34 billion parameters in numerous benchmarks. The model employs techniques like Grouped-query attention (GQA) for quicker processing and Sliding Window Attention (SWA) for better handling of longer text passages.

According to [24], the LLaVA MLLM has been trained using a method that leverages ChatGPT/GPT-4 for multimodal instruction-following data collection from image-caption pairs. The study creates text-based instructions that allow GPT-4 to process visual content in text form. The training dataset, utilizing COCO images [23,24], includes 158K unique samples categorized into three distinct types of instruction-following data: *conversations* that delve into the visual content through a series of questions about objects, actions, and spatial relationships; *detailed descriptions* aimed at generating comprehensive depictions of images; and *complex reasoning tasks* that require in-depth logical analysis. This approach significantly enhances the model's capability in spatial reasoning and understanding visual content through textual descriptions, with GPT-4 consistently outperforming in generating high-quality instruction-following data across these categories. To give an idea of the training data, in Fig. 2 is represented an example of conversation concerning an image from the COCO dataset.

For the experiments we used the source code and pre-trained model provided by the developers of LLaVA 1.6 setting the parameters as in Table 1. In addition, after a few experiments we developed the prompt reported in Fig. 3 which proved effective for obtaining responses easy to process.

Table 1. Main parameters used in LLaVA and their values.

Parameter	Value/Description
temperature	0.0 (Deterministic output, aiming for the most likely outcome)
top_p	None (Default or deterministic behavior in text generation)
num_beams	1 (Beam search used with a single beam, i.e. greedy decoding)
max_new_tokens	1024 (Limits the length of the generated output)

4 Results and Discussion

In this section, we first present a preliminary semi-qualitative analysis of model-generated scores and comments based on a dataset of scored street-view images collected in a previous study [19]. Then, we present a practical experiment to assess the walkability of streets in a city, using a dataset of SVIs covering the city's street network.



Training dialog:
“from": “human", “value": “How many traffic signals are there in the image?"
“from": “gpt", “value": “There are two traffic signals in the image."
“from": “human", “value": "Is there a lot of traffic in the image?"
“from": “gpt", “value": “No, the image shows an empty street with no visible traffic."
“from": “human", “value": “Where are the traffic lights positioned in the image?"
“from": “gpt", “value": “The traffic lights are positioned near the street intersection on the sidewalk or a little island."
“from": “human", “value": “What is the perspective of the image?"
“from": “gpt", “value": “The perspective of the image is a topside view, likely taken from an elevated position."
“from": “human", “value": “What can you infer about the traffic in the area based on the image?"
“from": “gpt", “value": “While it’s not possible to draw definitive conclusions about traffic conditions in the area based on just this single image, a few inferences could be made. The fact that the intersection is empty at the time the photo was taken might suggest that the area experiences low to moderate traffic or that the image was captured during a non-peak traffic period. Having multiple traffic signals and a dedicated island for the signals could indicate that this intersection is usually busier during different times of the day. Increased traffic could be expected, for example, during morning and evening rush hours or on weekends, when more people are likely to be out driving. In summary, the image provides limited information about overall traffic conditions in the area, but the presence of traffic signals suggests that the intersection may experience higher traffic volumes at other times of the day or week."

Fig. 2. Figure 000000393242.jpg from the COCO dataset and the corresponding LLaVA training conversation [24].

4.1 Evaluations of LLaVA Responses

We leveraged the expertise of urban walkability assessments from a prior study [19], where experts were engaged solely to furnish walkability scores to facilitate the training of a CNN-based classifier. This phase had the only purpose to perform a preliminary evaluation of the degree to which LLaVA’s computational assessments align with or diverge from human expert evaluations. Such a com-

Analyze the provided street-level photo to assess the walkability of the urban environment. Focus on the urban context visible in the image.
Scoring Criteria: Assign a walkability score from 1 to 5. A score of 1 indicates extremely poor walkability, and a score of 5 indicates high walkability. Base your score on both visible elements and inferred aspects of the urban environment. Provide a score utilizing contextual clues and inferences as necessary.
Response Template: Adhere strictly to the template below for your analysis. Begin your response with the motivation and conclude with the score. Do not include any additional text outside of this template.
- Motivation: [Provide the reasoning behind your score in no more than 250 characters].
- Score: [Insert the walkability score from 1 to 5].

Fig. 3. Prompt used for the LLaVA-based assessment of urban walkability.

parison is instrumental in validating the efficacy of MLLMs in recognizing and interpreting the complexities inherent in urban walkability metrics. However, a certain level of disagreement was expected being the LLaVA model not explicitly trained on walkability evaluation.

In Figs. 4, 5, 6, 7 and 8 we provide some of the evaluations we performed in terms of comparison between the LLaVA scores with human experts' scores.

In the first example of Fig. 4 both LLaVA and the human expert recognize the high walkability of a residential area. LLaVA noted the mix of urban and natural elements, presence of parked cars indicating pedestrian activity, ample greenery, and well-maintained wide sidewalks. This instance demonstrates LLaVA's capability to accurately assess environments that exhibit clear and strong indicators of walkability, aligning closely with human perception. In the second example, LLaVA assigns a slightly higher score than the human expert to a street scene. While LLaVA acknowledge the presence of parked cars, pedestrians, greenery, perhaps it weighs the positives more heavily than the human expert. The discrepancy could indicate a variation in the prioritization of certain features or a difference in the perceived impact of these features on walkability. This is perfectly understandable in a zero-shot learning setting.

In the examples of Fig. 5, LLaVA rates narrow alleyways and streets higher than human experts do. As mentioned, the motivation of human experts are not available, as they were asked to provide only a score. However, the model recognizes clear paths for pedestrians and the potential for pedestrian-friendliness due to parked cars and a mix of residential or commercial buildings. The higher scores by LLaVA may reflect its valuation of unobstructed paths and mixed-use environments. However, human experts might have considered additional factors such as the narrowness of pathways and the lack of certain amenities, which slightly lower their assessments.

In Fig. 6, in which some intermediate-level walkability scenarios are depicted, both LLaVA and human experts agree on the scores. LLaVA noted the presence



Human expert score: 5; **LLaVA score**: 5
LLaVA comment: "The image depicts a street-level view of a residential area with a mix of urban and natural elements. The presence of parked cars along the sidewalk indicates that there is pedestrian activity in the area. The street is wide and appears to be designed for moderate traffic flow. There is ample greenery with trees lining the street, contributing to a pleasant and walkable environment. The sidewalk is wide and appears to be well-maintained".



Human expert score: 4; **LLaVA score**: 5
LLaVA comment: "The image depicts a street in a European city with a mix of residential and commercial buildings. There are parked cars on both sides of the street, and the presence of pedestrians suggests that the area is walkable. The buildings are close to the sidewalk, and there are trees and greenery, which contribute to a pleasant urban environment. The street appears to be designed for moderate traffic, and there are no highways or motorways visible".

Fig. 4. Two examples of street view images scored 5 by LLaVA.

of sidewalks and urban settings but also recognizing limitations like the primary design for vehicular traffic and lack of pedestrian-focused infrastructure. These examples show a good alignment between LLaVA's assessments and human judgment, suggesting that the model is capable of nuanced analysis where walkability indicators are mixed.

In Fig. 7, both assessments provide a score of 2. LLaVA points out the challenges to walkability posed by a wide street primarily designed for vehicular traffic, with parked cars on both sides and a lack of clear pedestrian infrastructure. This agreement indicates that LLaVA can effectively identify environments where the walkability is compromised by factors prioritizing vehicle movement over pedestrian access.

Finally, in Fig. 8 both LLaVA and the human expert assign the lowest score to an environment characterized by a highway with multiple lanes of traffic and an absence of pedestrian infrastructure. This consensus underscores the model's



Human expert score: 3; **LLaVA score**: 4
LLaVA comment: "The image depicts a narrow alleyway between two buildings, with a clear path for pedestrians. The buildings appear to be residential or commercial, and there are no visible obstructions or hazards. The presence of a mailbox suggests that this area is likely to be maintained and serviced. The alleyway is not wide, but it is clear and unobstructed, which could make it a pleasant and convenient route for walking".



Human expert score: 3; **LLaVA score**: 4
LLaVA comment: "Motivation: The image depicts a narrow street lined with parked cars on both sides. The presence of multiple parked cars suggests that the area is likely to be pedestrian-friendly, as it is designed to accommodate parked vehicles. The buildings appear to be residential or commercial, which could indicate a mix of activities and services available within walking distance. The street itself is not wide enough to accommodate high-speed traffic, and there are no visible signs of highways or motorways. The overall impression is that of a pleasant urban environment with a moderate level of walkability".

Fig. 5. Two examples of street view images scored 4 by LLaVA.

ability to recognize and evaluate scenarios where walkability is severely limited due to the dominance of vehicular traffic and industrial or commercial settings with no pedestrian access.

Overall, we mostly found instances where LLaVA's walkability scores align with those of human experts. This congruence demonstrates LLaVA's potential to accurately interpret and assess the walkability of urban environments. Such consistency is particularly notable in cases where the walkability determinants are visibly apparent, such as the presence of sidewalks, greenery, and pedestrian activity. These findings suggest that LLaVA can effectively identify and weigh these visible factors in its assessments.

However, the analysis also reveals discrepancies between LLaVA scores and those of human experts. These differences may arise from the model's interpretation of subtle urban design features or its valuation of certain walkability



Human expert score: 3; **LLaVA score**: 3
LLaVA comment: "The image depicts a street in a residential area with low-rise buildings, parked cars, and a few pedestrians. The presence of sidewalks and the absence of high-speed traffic suggest a relatively walkable environment. However, the lack of visible shops, cafes, or other amenities might limit the attractiveness of the area for pedestrians".



Human expert score: 3; **LLaVA score**: 3
LLaVA comment: "The image depicts a street scene with a mix of parked and moving vehicles, including cars and a motorcycle. There are pedestrians visible, and the presence of a sidewalk suggests a pedestrian-friendly environment. The street is lined with buildings, indicating an urban setting. However, the street appears to be primarily designed for vehicular traffic, with no clear pedestrian pathways or crossings. The presence of a bicycle lane suggests some consideration for alternative modes of transportation".

Fig. 6. Two examples of street view images scored 3 by LLaVA.

determinants over others. For instance, LLaVA may assign a higher score to a street with ample greenery and pedestrian infrastructure but overlook factors like street connectivity or the presence of amenities, which human experts might consider crucial for walkability. Such discrepancies highlight areas where LLaVA's assessment criteria could be refined to better align with human judgment.

These variations suggest areas for further refinement in LLaVA's evaluation process, particularly in fine-tuning its sensitivity to subtle urban design features and the relative importance of different walkability determinants. Training LLaVA with a dataset explicitly labeled with walkability scores and associated rationale could enhance the model's ability to evaluate complex urban environments similarly to human experts. This process could help the model better understand the nuances of what makes an area walkable, including less tangible aspects like safety perceptions or the social vibrancy of a street, which are often considered by human assessors.



Human expert score: 2; **LLaVA score**: 2
LLaVA comment: "The image shows a wide, tree-lined street with parked cars on both sides, indicating a residential area. There are no visible pedestrian crossings or sidewalks, and the street appears to be designed primarily for vehicular traffic. The presence of parked cars and the absence of clear pedestrian infrastructure suggest that this is not a highly walkable urban environment".

Fig. 7. An example of photo scored 2 by LLaVA.



Human expert score: 1; **LLaVA score**: 1
LLaVA comment: "The image shows a highway with multiple lanes of traffic, including cars and trucks. There is no visible pedestrian infrastructure, such as sidewalks or crosswalks, and the surrounding area appears to be industrial or commercial with no clear pedestrian access. The environment is not pedestrian-friendly, and the presence of high-speed traffic suggests a low walkability score".

Fig. 8. An example of photo scored 1 by LLaVA.

4.2 Example of Application to an Urban Area

To illustrate a practical application of the MLLM-based methodology for walkability assessment, we conducted an experiment focusing on the city of Cagliari, the capital of Sardinia Island, Italy. The urban area has a varied urban fabric, encompassing a mix of historical and modern elements, diverse street layouts, and different levels of pedestrian infrastructure. This provided a rich dataset for evaluating the MLLM's ability to assess urban walkability comprehensively. The experiment utilized a dataset of approximately 30,000 Google Street View images, covering a significant portion of Cagliari's street network. Each street segment in the network was represented by two to six images, depending on the segment's length, to ensure a robust spatial representation. The MLLM, specifically the LLaVA model, was then tasked with analyzing these images to assign walkability scores based on its training and zero-shot learning capabilities.

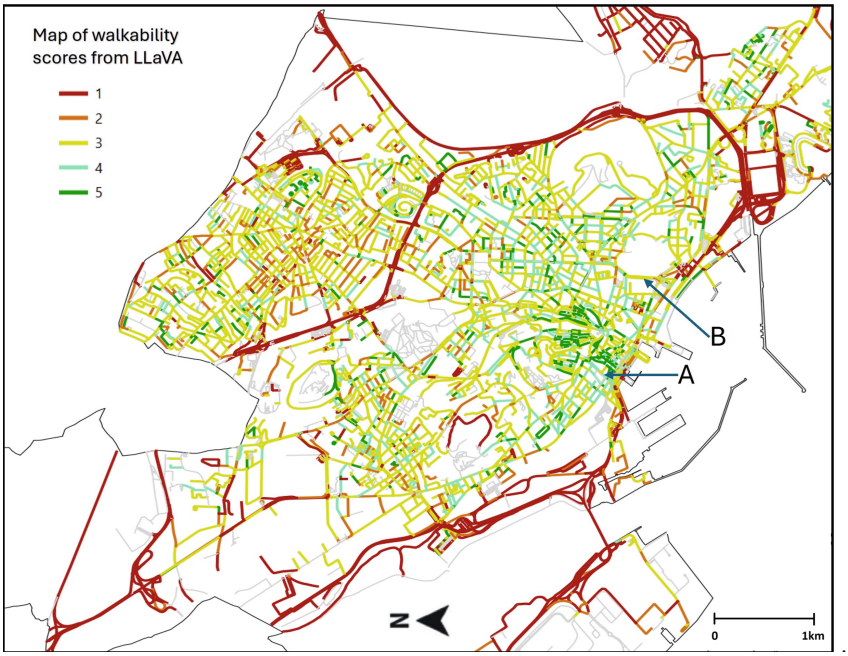
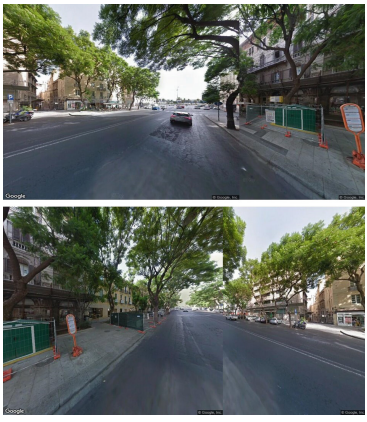


Fig. 9. The map of walkability scores obtained for Cagliari using LLaVA in a zero-shot learning setup. (Grey lines represent natural and administrative boundaries, or borders of areas non accessible to pedestrians.) Details of results obtained from streets labeled as A and B are shown in Figs. 10 and 11.



Overall generated comment: "The photographs capture a city street with a mix of pedestrian and vehicular activity. Lined with trees, the street offers shade, enhancing the walkability and creating a pleasant environment. The presence of parked cars, a motorcycle, and outdoor seating at a café indicates accessibility and use by people. Although designed for moderate traffic, the street is pedestrian-friendly with clear lanes for vehicles and sidewalks. Its location in a low-density urban area suggests a higher walkability score, making it an attractive and inviting space for both pedestrians and vehicles.".

Fig. 10. Details of case A in Fig 9. The overall generated score for the street is 5.



Overall generated comment: "The street in question is a mix of urban elements, with varying degrees of pedestrian-friendliness. The presence of trees and sidewalks in some areas suggests efforts to create a welcoming environment for pedestrians, while the visible vehicular traffic and wide roads in other sections may indicate a design prioritizing cars over pedestrians. The overall walkability of the street appears to be moderate, with potential for pedestrian activity in low-density urban areas, but concerns about pedestrian safety due to the lack of designated crossings or traffic calming measures. ".

Fig. 11. Details of case B in Fig 9. The overall generated score for the street is 3.

The application process involved three steps. First, each image was independently assessed by LLaVA, which provided a walkability score ranging from 1 (low walkability) to 5 (high walkability), accompanied by a qualitative explanation of the score. This step utilized the model's ability to generate insights based on both visible elements and inferred aspects of the urban environment. Subsequently, for streets represented by multiple images, scores were aggregated to derive an overall walkability assessment for each street segment. This aggregation considered the variability and diversity of perspectives within the same street, using majority voting for identical viewpoints and averaging for different viewpoints. Finally, the aggregation of all the available qualitative comments for each street segment was performed using the open source Mistral LLM developed by Mistral AI [26]. This advanced language processing step is necessary for integrating diverse observations and insights into a comprehensive summary that reflects the collective attributes of walkability as depicted across multiple images. This process goes beyond simple averaging or majority voting.

The final result of the process outlined above is a detailed walkability map of the city, with each street segment color-coded according to its overall walkability score and annotated with synthesized comments that highlight specific areas of

strength or needed improvements. The comprehensive street-level walkability map of Cagliari is shown in Fig. 9. This map serves as a practical tool for urban development, providing a clear visual and textual representation of walkability across different parts of the city.

The walkability map generated from LLaVA's assessments revealed reasonable patterns of walkability across the city, highlighting areas with high pedestrian friendliness as well as zones where improvements could enhance pedestrian experiences. Notably, historical centers and residential areas generally received higher walkability scores, attributed to their urban layouts, presence of greenery, and limited vehicle traffic. In contrast, more modern districts characterized by wider roads and vehicle-oriented design tended to receive lower scores.

To gain further insights we focused on two notable cases, labeled as A and B in the walkability map. In Case A (Fig. 10), the street, receiving a high walkability score of 5, was characterized by abundant greenery, pedestrian infrastructure, and a balanced mix of vehicular and pedestrian activity. LLaVA's overall comments underscored the area's pleasant and inviting environment for pedestrians. In Case B (Fig. 11): the street received a moderate walkability score of 3. The model's assessment pointed to the mixed urban elements, with some pedestrian-friendly features offset by areas prioritizing vehicular traffic, indicating room for pedestrian safety and amenity improvements. Overall the generated comments seem well justified by the street view images.

One of the distinctive characteristics of the MLLM-based walkability evaluation investigated in this study, lies in associating aggregated comments to the classical walkability scores. This provides urban planners and policymakers with an easily interpretable overview of the factors affecting walkability, enriched by the model's ability to detect and articulate complex patterns and insights.

5 Conclusions and Future Work

With its capability to produce qualitative comments along the walkability scores, the proposed method is significantly more powerful than conventional ones reliant on CNNs and SVI, which typically yield scores without explanations. This dual-output approach improves the interpretability of walkability data, aiding urban planners and policymakers in making informed decisions. Such transparency fosters trust in AI-driven assessments and encourages public participation by making urban walkability assessments more understandable.

Furthermore, the detailed comments allow for targeted urban development strategies, directly addressing specific walkability factors. This method facilitates a comparative analysis over time or between different areas, enabling a dynamic approach to urban improvement.

In this preliminary study, we relied on prompt optimization without computationally expensive MLLM fine-tuning. This process inherently is based on the model's zero-shot learning capabilities, hypothesizing its potential to assess walkability despite not being explicitly trained on such evaluations. Future work will explore enhancing this approach through specific fine-tuning of the MLLM. This

refinement could further align the model's assessments with expert evaluations and real-world urban walkability considerations, strengthening the methodology's application in creating pedestrian-friendly urban environments.

Another potential direction of improvement is represented by the integration with other data sources: combining SVI with other data types, such as pedestrian flow data or urban infrastructure databases, aerial images could provide LLaVA with a more comprehensive basis for its assessments, potentially reducing discrepancies with human expert scores.

Acknowledgments. This study was carried out within the MOST-Sustainable Mobility National Research Centre and received funding from the European Union Next-GenerationEU (PNRR, Missione 4, Componente 2, Investimento 1.4, D.D. 1033 17 June 2022, CN00000023). The research was also supported by funding from the Collaboration agreement between Autonomous Region of Sardinia (CRP) and Universities of Cagliari (DICAAR) and Sassari (DADU) on "Support to local authorities and the regional administration for the construction and implementation of planning, management and monitoring models on transport systems and services for sustainable mobility" related to Priority 4 of the PR FESR 2021-2027 and to the National Strategy for the Internal Areas 2021-2027 (RAS Convenzione n. 87 Prot. n. 10469 23.12.2022).

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

References

1. Saelens, B.E., Handy, S.L.: Built environment correlates of walking: a review. Med. Sci. Sports Exerc. **40**, 550–566 (2008)
2. Maghelal, P.K., Capp, C.J.: Walkability: a review of existing pedestrian indices. URISA J. **23**, 5–19 (2011)
3. Talen, E., Koschinsky, J.: The walkable neighborhood a literature review. Int. J. Sustain. Land Use Urban Plan **1**, 42–63 (2013)
4. Vale, D.S., Saraiva, M., Pereira, M.: Active accessibility: a review of operational measures of walking and cycling accessibility. J. Transport Land Use **9**, 209–235 (2016)
5. Frank, L., Schmid, T., Sallis, J., Chapman, J., Saelens, B.: Linking objectively measured physical activity with objectively measured urban form: findings from SMARTRAQ. Am. J. Prev. Med. **28**, 117–125 (2005)
6. Lee, K., Ahn, K.: An empirical analysis of neighborhood environment affecting residents' walking: case study of 12 areas in Seoul. J. Architectural Inst. Korea Plann. Des. **24**, 293–302 (2008)
7. Wang, W., Li, P., Wang, W., Namgung, M.: Exploring determinants of pedestrians' satisfaction with sidewalk environments: case study in Korea. J. Urban Plann. Dev. **138**, 166–172 (2012)
8. Mateo-Babiano, I.: Pedestrian's needs matter: examining manila's walking environment. Transp. Policy **45**, 107–115 (2016)
9. Lee, E., Dean, J.: Perceptions of walkability and determinants of walking behavior among urban seniors in Toronto, Canada. J. Transport Health **9**, 309–320 (2018)

10. Biljecki, F., Ito, K.: Street view imagery in urban analytics and GIS: a review. Landscape Urban Plann. **215**, 104217 (2021)
11. Neuhold, G., Ollmann, T., Rota Bulò, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: International Conference on Computer Vision (ICCV) (2017)
12. Zhou, H., He, S., Cai, Y., Wang, M., Su, S.: Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning. Sustain. Cities Soc. **50**, 101605 (2019)
13. Li, Y., Yabuki, N., Fukuda, T., Zhang, J.: A big data evaluation of urban street walkability using deep learning and environmental sensors-a case study around Osaka university Suita campus. In: Proceedings of the 38th eCAADe Conference, Berlin, Germany, eCAADe (2020) 319–328 16–18 September 2020 (2020)
14. Wu, C., Peng, N., Ma, X., Li, S., Rao, J.: Assessing multiscale visual appearance characteristics of neighborhoods using geographically weighted principal component analysis in Shenzhen, China. Comput. Environ. Urban Syst. **84**, 101547 (2020)
15. Ordonez, V., Berg, T.L.: Learning high-level judgments of Urban perception. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 494–510. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_32
16. Porzi, L., Rota Bulò, S., Lepri, B., Ricci, E.: Predicting and understanding urban perception with convolutional neural networks. In: Proceedings of the 23rd ACM International Conference on Multimedia. MM 2015, New York, NY, USA, ACM (2015) 139–148 (2015)
17. Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A.: Deep learning the city: quantifying urban perception at a global scale. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016, pp. 196–212. Springer International Publishing, Cham (2016)
18. Liu, L., Silva, E.A., Wu, C., Wang, H.: A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. Comput. Environ. Urban Syst. **65**, 113–125 (2017)
19. Blečić, I., Cecchini, A., Trunfio, G.A.: Towards automatic assessment of perceived walkability. In: Gervasi, O., Murgante, B., Misra, S., Stankova, E., Torre, C.M., Rocha, A.M.A.C., Taniar, D., Apduhan, B.O., Tarantino, E., Ryu, Y. (eds.) ICCSA 2018. LNCS, vol. 10962, pp. 351–365. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95168-3_24
20. Blečić, I., Santos, A.G., Moura, A.C., Trunfio, G.A.: Multi-criteria evaluation vs perceived Urban quality: an exploratory comparison. In: Misra, S., Gervasi, O., Murgante, B., Stankova, E., Korkhov, V., Torre, C., Rocha, A.M.A.C., Taniar, D., Apduhan, B.O., Tarantino, E. (eds.) ICCSA 2019. LNCS, vol. 11621, pp. 612–627. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-24302-9_44
21. Chang, Y., et al.: A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. **15**(3), 1–45 (2024)
22. Yin, S., et al.: A survey on multimodal large language models. arXiv preprint arXiv:2306.13549 (2023)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
25. Liu, H., et al.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge - https://llava-vl.github.io/blog/2024-01-30-llava-next (2024)
26. Jiang, A.Q., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023)