This paper is peer-reviewed and accepted at the *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.
https://isprs-annals.copernicus.org/articles/X-4-W7-2025/1/2025/
DOI: 10.5194/isprs-annals-X-4-W7-2025-1-2025

Can a Large Language Model Assess Urban Design Quality? Evaluating Walkability Metrics Across Expertise Levels

Chenyi Cai[1], Kosuke Kuriyama[1,2], Youlong Gu[1,3], Filip Biljecki[3,4], Pieter Herthogs[1,3]

[1] Singapore-ETH Centre, Future Cities Lab Global Programme, CREATE campus, 1 Create Way, #06-01 CREATE Tower, 138602, Singapore.
[2] Takenaka Corporation, Project Development Division, 4-1-13 Hommachi, Chuo-Ku, Osaka, 541-0053, Japan.
[3] Department of Architecture, National University of Singapore, 4 Architecture Dr, 117566, Singapore
[4] Department of Real Estate, National University of Singapore, 15 Kent Ridge Drive, 119245, Singapore

Keywords: public space evaluation, walkability assessment, semantic and visual elements, street view images, ontology.

Abstract

Urban street environments are vital to supporting human activity in public spaces. The emergence of big data, such as street view images (SVI) combined with multi-modal large language models (MLLM), is transforming how researchers and practitioners investigate, measure, and evaluate semantic and visual elements of urban environments. Considering the low threshold for creating automated evaluative workflows using MLLM, it is crucial to explore both the risks and opportunities associated with these probabilistic models. In particular, the extent to which the integration of expert knowledge can influence the performance of MLLM in the evaluation of the quality of urban design has not been fully explored. This study set out an initial exploration of how integrating more formal and structured representations of expert urban design knowledge (e.g., formal quantifiers and descriptions from existing methods) into the input prompts of an MLLM (ChatGPT-4) can enhance the model's capability and reliability to evaluate the walkability of built environments using SVIs. We collect walkability metrics through the existing literature and categorise them using relevant ontologies. Then we select a subset of these metrics, used for assessing the subthemes of pedestrian safety and attractiveness, and develop prompts for MLLMs accordingly. We analyse MLLM's abilities to evaluate SVI walkability subthemes through prompts with multiple levels of clarity and specificity about evaluation criteria. Our experiments demonstrate that MLLMs are capable of providing assessments and interpretations based on general knowledge and can support the automation of image-text multimodal evaluations. However, they generally provide more optimistic scores and can make mistakes when interpreting the provided metrics, resulting in incorrect evaluations. By integrating expert knowledge, MLLM's evaluative performance exhibits higher consistency and concentration. Therefore, this paper highlights the importance of formally and effectively integrating domain knowledge into MLLMs for evaluating urban design quality.

1. Introduction

Urban street environments are vital to supporting human activity in public spaces (Jacobs, 2010). Streets are essential connectors within urban networks, allowing for seamless movement of pedestrians, cyclists, and vehicles. Well-designed urban environments can have a range of positive impacts, such as encouraging physical activities, improving mood, strengthening urban identity, and promoting public health (Koohsari et al., 2020; Wedyan and Saeidi-Rizi, 2025). In contrast, poorly designed or unpleasant urban environments can have various negative consequences (de Jong and Fyhri, 2023; Giles-Corti et al., 2016).

Multi-modal large language models (MLLM) such as GPT-4 (OpenAI), with their ability to analyze textual and visual data, hold the potential for providing evidence-based evaluations and quality improvement suggestions for the built environments. Currently, the large availability of street view imagery (SVI) provides a rich data source, along with the rapid development of computer vision techniques, enabling the computational assessment of the visual quality of the street environment. Existing literature on SVI and environmental quality has primarily focused on identifying key correlations between visual street features and travel behaviour, enhancing the accuracy of street quality indicators, and mapping the spatial distribution of environmental attributes within study cities (Biljecki and Ito, 2021; Liu and Sevtsuk, 2024). MLLMs hold significant potential for extracting physical environment features and automating evaluative workflows (Malekzadeh et al., 2025). Although LLMs demonstrate a certain level of knowledge about global cities, their limitations become evident when they encounter unfamiliar tasks, often producing generic or random outputs (Li et al., 2024). A significant challenge lies in the gap between the generalised training data of MLLMs and the specialised knowledge required for evaluating built environment quality. The risks and advantages of MLLM — a black box — in providing evaluation results, and the potential to enhance its performance by integrating expert urban design knowledge, have not been fully

explored.

Formal definitions and representations of expert knowledge enable the automation of environmental quality assessments and the operationalization of urban design. Public space quality assessments, such as walkability metrics, utilise determinants and criteria as representations of expert knowledge (Fonseca et al., 2022; Ariffin et al., 2021). While the key components of walkable urban environments are well-recognised by the research community (Ewing and Handy, 2009; Dragović et al., 2023), translating these principles into interpretable and robust indicators requires formal definitions of domain-specific concepts. In walkability evaluation, there are mixed-use of terms and measurements from different domains, scales and data sources. First, for instance, land use density is linked to attractiveness in one framework (Frank et al., 2010) and to accessibility in another (Pelclová et al., 2014). Additionally, methods for measuring land use density vary across different studies. Second, walkability metrics contain both contextual factors (e.g., entropy index of different land uses in an area) and site-specific factors (e.g., presence of streetlights) (Fonseca et al., 2022). Metrics include elements that are directly measurable (e.g., presence of fixed furniture on streets) and those requiring monitoring (e.g., history of thefts). Some elements in these frameworks are actionable for urban design, while others are not (Dragović et al., 2023; Reisi et al., 2019). In summary, the variable use of terms and measurements across different frameworks causes challenges for interpretability and comparability between cases and between frameworks. In turn, a lack of formal definitions and categorisations for urban design-related metrics also makes it difficult to generate clear design recommendations to improve environments.

Exploring the risks and opportunities of MLLM in urban evaluative frameworks through the integration of formal expertise is crucial. First, it requires translating multiple characteristics into formalised and methodologically practical indicators. Second, integrating expert urban design knowledge (e.g., definitions, categorisations, scoring models) formally and effectively into MLLMs and enhancing their performance in evaluating built environment quality requires more investigations. Hence, the following research questions are raised:

- Can MLLMs provide consistent responses when prompted to evaluate street environment quality?
- To what extent can expert knowledge (more formal definitions and semantic clarity) and a structured evaluation framework enhance MLLMs' ability to evaluate street environment quality (e.g. walkability)?

Our research aims to bridge core urban design variables of street environments with urban design solutions to increase suitability for human activities. We investigate how formal representation of urban design knowledge can improve the MLLM's consistency to assess walkability. In this paper, we sets out the initial explorations and implementations, focusing on: 1) categorising walkability metrics through tangible measurements in the built environment, and 2) investigating how the level of formality in urban design expert knowledge influences MLLM performance, including expert descriptions from the literature review and semantic clarity of walkability metrics. First, through the existing literature, we collect and categorise walkability metrics based on measurements, criteria, methods, and data sources. Second, using example metrics related to pedestrian Safety and Attractiveness, we develop prompts, with varying levels of formalization and identification, to compare MLLMs' produced assessments. Third, along with SVIs from selected locations in Singapore, we examine the performance of MLLMs in evaluating SVI walkability. We apply statistical analysis to the assessment results, focusing on the general score distributions and notable differences between particular metrics. Finally, the paper provides an example of identifying potential urban design interventions for places requiring improvement, focusing on the metrics that are actionable within the scope of urban design.

2. Background

Extensive research explores how urban spatial characteristics influence the suitability of street environment for people and activities. Street view images (SVI) enable the study of the physical environment and its interactions with the socioeconomic environment at various scales (Zhang et al., 2024) and have been widely used for numerous applications – ranging from analysing vegetation and transportation to health and socio-economic studies (Biljecki and Ito, 2021). In walkability-related studies, because of SVIs' convenience, field auditing works have been replaced by the desktop-auditing tools (Larranaga et al., 2019). A wide range of walkability and SVI-related research focuses on uncovering the most significant correlations between visual street features and travel behaviour, as well as mapping the spatial distribution of specific environmental features (Larranaga et al., 2019; Huang et al., 2024).

The advent of MLLMs, such as GPT-4 (Achiam et al., 2023), which combine the textual interaction capabilities of LLMs with image analysis, unlocked new possibilities for applications that demand integrated visual and textual interpretation. Liu et al. (2023) introduced methods using the Multimodal Contrastive Learning Model(CLIP) to assess perceived walkability by analysing both tangible and subjective factors such as safety and attractiveness. Compared with convolutional neural networks (CNN), MLLM has the strength in increasing explainability in AI-driven assessments. It can generate interpretations and explanations for walkability assessments, hence it can provide insights into the specific factors that influence the evaluations (Blečić et al., 2024). However, it is evident that while LLMs possess a certain level of urban knowledge, they have limitations when faced with unfamiliar tasks, often generating generic or random outputs (Li et al., 2024). Given the potential of using MLLMs, they are developed based on generalised training data (Bender et al., 2021). However, evaluating walkability and providing suggestions for improvement requires specialised urban design knowledge. Therefore, whether MLLMs can effectively evaluate walkability and offer professional improvement suggestions and how to integrate expert knowledge remains unexplored.

Various evaluation frameworks and metrics have been developed to represent urban design knowledge, using varied criteria and metrics in different scales. Focusing on walkability, the metrics from the literature examine local elements (e.g. visible amenities, greenery) and contextual factors (e.g. street networks, neighbourhood land use) (Fonseca et al., 2022; Dragović et al., 2023), some of which are actionable for urban designers (e.g. park and green zones, aesthetics of buildings), while others are not (e.g. motorised transport speed, weather conditions). *Greenness* is an example of a criterion name label that is measured in very different ways, such as assessing landscape coverage from satellite images (Fan et al., 2018), calculating the

percentage of trees in street view images (Huang et al., 2024), counting the number of street segments with street trees (Lee et al., 2020), or even defining "green" to include the presence of amenities that seem wholly unrelated to green, such as health services, banks, or auto services (Pereira et al., 2020).

In many urban design-related walkability evaluation frameworks, there are different sets of criteria. Yin (2017) developed the evaluation framework by methodologically interpreting and translating the criteria proposed by Ewing and Handy (2009). Arellana et al. (2020) assess it based on factors such as sidewalk condition, traffic safety, comfort, and attractiveness. Meanwhile, Larranaga et al. (2019) employed a different set of criteria, such as connectivity, topography, sidewalk surface, number of police officers, and number of shops. There is a mix of criteria types. For example, criteria such as safety or attractiveness are dispositions, while criteria such as the number of shops or sidewalk width are directly measurable properties. Metrics often share the same definition but are labelled differently, for instance, the number of parks was used to measure imageability (Yin, 2017) in one study and to measure amenity density (Fonseca et al., 2022) in another. Grisiute et al. (2024) highlighted the need for a structured and shared vocabulary for bike network evaluations to enhance the coherence of evaluation methods within the field of bike network planning — the same is true for walkability. The fact that different studies use different evaluation frameworks and models is to be expected, but the observable lack of formal definitions and semantic accuracy. These inconsistencies also hinders the potential implementations using digital technologies (e.g.LLMs).

3. Methodology

This section introduces our methodology for MLLM-based street environment walkability evaluation. Our methodology was shaped by the following scopes and aims:

Aims In the paper, our goal is to define expert knowledge in walkability assessment in distinct levels based on current literature, and to assess and compare the evaluative results provided by the MLLMs. This paper does not aim to propose a new set of indicators or refine the semantic correctness of indicator names from the literature; these aspects will be addressed in future work.

Scopes We collect metrics that are directly related to public spaces, along with the criteria used to cap them, as documented in the existing literature. We conducted a comparative study designing four prompt sets. The prompt sets have varying levels of semantic clarity to represent different levels of embedding of expert knowledge. We applied inferential statistics to identify differences in MLLM assessments and highlight metrics with statistically significant variations between models, hence identifying the influence of intergating expert knowledge to MLLMs.

3.1 Selecting and structuring walkability metrics

We selected our walkability metrics and identified the criteria based on two review articles about measuring walkability (Dragović et al., 2023; Fonseca et al., 2022). Based on the combined literature sets from both reviews, we conduct a review of studies that develop walkability metrics. This combined set was used to develop a list of indexes, aimed at building upon existing work and reducing bias in the indicator selection process. Our work reveals that varying interpretations of what defines a good walking environment have led to a broad array of interchangeably used terms. For example, inconsistencies were observed where different metric names referred to similar measurements, or conversely, identical metric names were used despite differences in the underlying methods.

From our initial literature set, we collected 124 walkability metrics. We use *Metric* to refer to direct measurables of the street environment, *Methods* to refer to evaluation methods (e.g. a survey or a tool), *Criterion* to refer to specific criteria in the evaluation approaches (e.g. Accessibility, Attractiveness). We introduced *DataSource* class and *ScoringFunction* class. We applied the Triple-A ontology (Herthogs, 2021) to hierarchically structure and describe the final versions of these metrics, similar to the metric structuring described by Grisiute et al. (2024) and Ataman et al. (2022). We set out the categorisation as the first step towards structuring walkability evaluation metrics. The entire metric database is available (Cai, 2025).

In this paper, we take 21 metrics that are used for evaluating *Safety* and *Attractiveness* as examples and feed multi-modal large language models (MLLM). We set a scale from 1 (lowest) to 5 (highest) for each metric, hence the total scoring scale is from 21 to 105. As shown in Table 1, we identified 21 metrics for each *Criterion*. We then adapted these into two sets of metrics: one with vague definitions while naming the metrics and another with more quantifiers in the metric names. These metric lists, developed based on the current literature, are non-exhaustive for evaluating walkability, safety, and attractiveness but are used as comparison sets to test MLLMs' performance in evaluating the street environment.

3.2 Shaping prompts for SVI samples

We selected streets representing various types in Singapore, including locations from the downtown area, countryside, commercial centre, and housing areas (Figure 1). A total of 42 SVIs are sourced from a crowd-sourced platform KartaView.

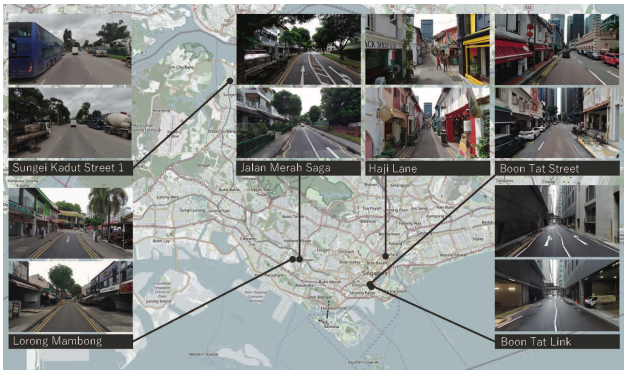

Figure 1. Streets in Singapore were selected for evaluating SVI walkability. (c) KartaView contributors.

The ChatGPT-4 model was used to evaluate the image dataset. Drawing on the metrics outlined in Table 1, we developed four distinct prompt sets, with different levels of defined information (which is one way of representing different levels of expertise), to assess the influence of formal definitions of expert knowledge on the walkability assessment using MLLM. Correspondingly, four MLLMs are developed by using the designed prompt sets representing four levels of expertise. The models demonstrate an increasing expert level of evaluation metric definition and semantic clarity from Model-C1 to Model-C4. Model C1 has

Table 1. Overview of the identified metrics for safety and attractiveness (as used in walkability studies). The vague metrics (Metric-1) exhibit vague definitions in their names and the quantified metrics (Metric-2) have more formal names using explicit quantifiers.

Vague Safety Metric-1	**Quantified Safety Metric-2**	**Vague Attractiveness Metric-1**	**Quantified Attractiveness Metric-2**
CrossingAids	PresenceOfCrossingAids	GreenArea	PresenceOfGreenArea
TrafficSignals	PresenceOfPedestrianSignals	InstitutionalArea	PresenceOfInstitutionalArea
SpeedBumps	PresenceOfTrafficCalmingDevice	ResidentialArea	PresenceOfResidentialArea
PoliceStations	PresenceOfPoliceStations	CommercialArea	PresenceOfCommercialArea
CCTV	PresenceOfSecurityCameras	Parks	PresenceOfParks
CrossRoads	NumberOfCrossingFacilities	Trees	PresenceOfTrees
RiskOfTrafficAccidents	PerceivedRiskOfTrafficAccidents	Attractiveness	PerceivedAttractiveness
VehicleSpeed	MotorisedTransportSpeed	EstheticFeatures	PerceivedNeighborEstheticFeature
VehicleFlow	VehicleFlow	CulturalCentres	NumberOfCulturalCentres
TrafficSafety	PerceivedTrafficSafety	Retails	AreaOfRetailTradeOrGastronomy
TrafficCalmingDevice	NumberOfTrafficCalmingDevices	FixedFurniture	NumberOfFixedFurniture
PoliceOfficers	NumberOfPoliceOfficers	PublicToilets	NumberOfPublicToilets
CrimeRate	PerceptionOfCrimeRate	TransportationStations	NumberOfTransportationStations
CrimeSecurityDuringDay	PerceivedDaytimeCrimeSecurity	Greenness	ProportionOfGreenness
CrimeSecurityAtNight	PerceivedCrimeSecurityAtNight	WalkableSpace	ProportionOfWalkableSpace
Lights	NumberOfLights	DiverseLandscape	LandscapeDiversityIndex
Graffiti	GraffitiOnBuildings	Colorfulness	EnvironmentalColorDiversity
FootTraffic	PerceptionOfPedestrianFlow	Sky	ProportionOfSky
LandmarkVisibility	LandmarkVisibilityIndex	Cleanliness	StreetCleanliness
DiverseLandscape	LandscapeDiversityIndex	LandmarkVisibility	LandmarkVisibilityIndex
Colorfulness	EnvironmentalColorDiversity	Transparency	TransparencyIndex

the lowest expertise level (level 1) and Model C4 has the highest expertise level (level 4). Model-C1 simply asks GPT-4 to assess Safety and Attractiveness without any metric input. Model-C4 prompt has a document of formal descriptions for each metric with more clarified definitions and scoring models. Table 2 shows three examples of metrics and their descriptions. The full document is available (Cai, 2025).

- Level 1 of expertise : Model-C1 uses no metrics, asking GPT-4 to rate pedestrian *Safety* and *Attractiveness* on a scale from 21 (lowest) to 105 (highest) without specifying any evaluation metrics.
- Level 2 of expertise: Model-C2 uses the 21 metrics from literature, but in vague language, incorporating Vague Safety Metric-1 and Vague Attractiveness Metric-1 with ratings on a scale from 1 (lowest) to 5 (highest) for each metric.
- Level 3 of expertise: Model-C3 uses the 21 metrics with quantifiers, incorporating Quantified Safety Metric-2 and Quantified Attractiveness Metric-2 with ratings on a scale from 1 (lowest) to 5 (highest) for each metric.
- Level 4 of expertise: Model-C4 uses quantified metrics and formal descriptions, integrating Quantified Safety Metric-2 and Quantified Attractiveness Metric-2, along with a document containing specified descriptions for each metric.

The models are tested in order from Model-C1 to Model-C4. We deploy Model-C4 with metrics and descriptions as the last one to prevent the language model from learning from the descriptions and influencing the results of other models. All tests are conducted on the same machine by the same user to avoid discrepancies between machines and accounts and to ensure the comparative study's consistency and reliability.

To further evaluate the scoring behaviour of the four language models under different prompts, Levene's Test is conducted to assess the assumption of homogeneity of variances as a prerequisite for variance analysis. Subsequently, Welch's ANOVA and the Games-Howell post hoc test are employed to examine differences in score distributions, both for the overall score and across individual metrics. Among the 21 metrics analyzed, the six exhibiting the most significant statistical differences are identified and further investigated to gain insights into the models' interpretative behaviour. Based on the statistical results, we further investigate the case studies of places with lower scores and discuss the potential urban design interventions.

Table 2. Three examples Model-C4 prompt, including metrics and their descriptions. The descriptions outline the definitions and scoring methods.

Examples
Metric: PresenceOfInstitutionalArea **Description**: Institutional area refers to educational, medical, community and cultural areas. **Scoring**: If one of the above institutional areas is present, score: 5. If not, score: 1.
Metric:NumberOfTrafficCalmingDevices **Description**:The number of traffic calming devices, such as speed bumps, raised crosswalks, and pedestrian islands. **Scoring**:A higher number of traffic calming devices corresponds to a higher score.
Metric: GraffitiOnBuildings **Description**: If graffiti is present on buildings, indicating a sense of unsafety, score: 1.

4. Results and Discussion

4.1 Results of MLLM evaluations on SVIs

Figure 2 shows the average safety and attractiveness scores for each street across the four models. Model-C1 diverges significantly from the other models in both assessments. It shows notable differences in scoring ranges across the six streets, too. For instance, Lorong Mambong scores lowest in safety in Model-C1 but ranks mid-range in other models.

Figure 3 presents the SVIs that received the highest and lowest scores across the four models. Overall, the models similarly

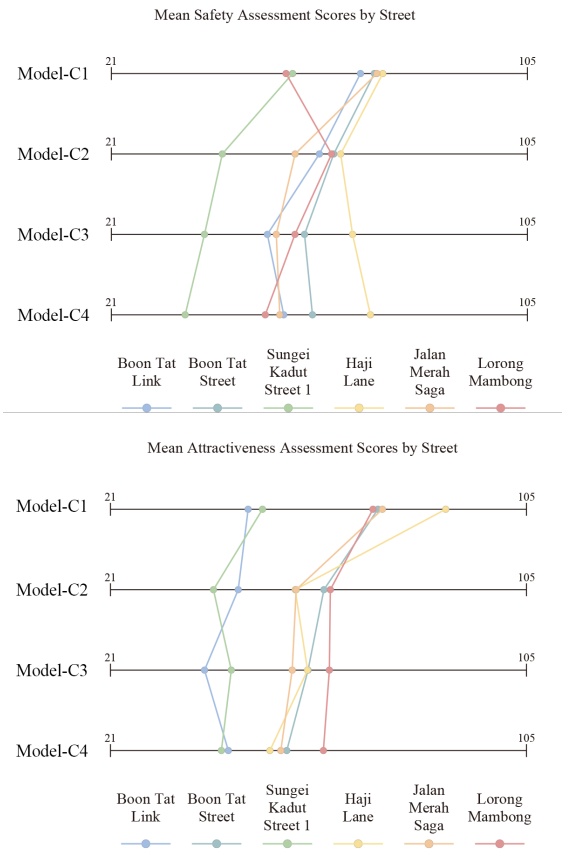

Figure 2. Plot of the streets' average scores in safety and attractiveness assessments, as evaluated by the four MLLMs.

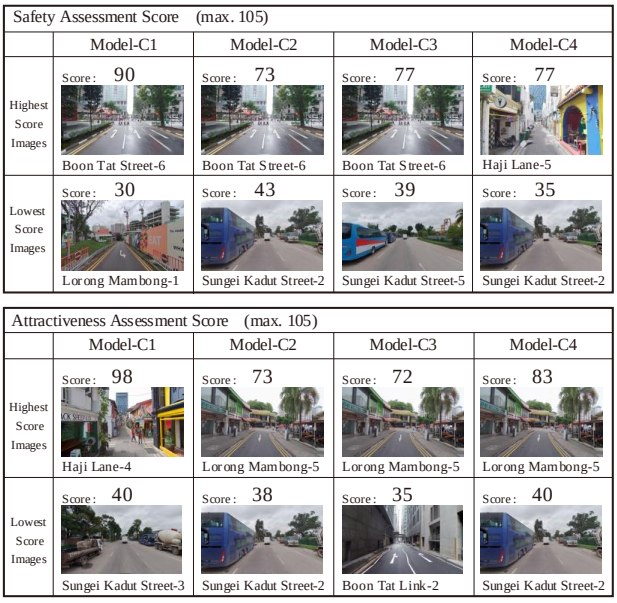

Figure 3. Street view images with the highest and lowest scores for safety (top) and attractiveness (bottom), giving the corresponding scores.

depict the safest and most attractive streets, as well as the least safe and least attractive streets. In the safety assessment, Haji Lane-5 achieved the highest score in Model-C4, while Boon Tat Street-6, which ranked highest in the other models, received the second-highest score (75) in Model-C4, showing the similarity between the four models. These results provide an initial indication that while MLLM can offer assessments based on general knowledge, integrating MLLM with representations of expert knowledge significantly influences its evaluative performance.

To compare the performance differences between the models, we conducted statistical analysis on all the SVI assessment scores generated from all four models. We conducted Levene's Test for the four models applied for *Safety* and *Attractiveness* to assess the equality of variances. In both cases, the null hypothesis was rejected. Consequently, we employed Welch's ANOVA to evaluate the differences in distributions across the four models for each metric. The results revealed statistically significant differences, with p-values for both tests below 0.01. We then performed the Games-Howell post hoc test and visualised the score distributions within 95% confidence intervals.

Table 3. Statistical comparison of MLLMs' scoring distributions for safety and attractiveness using Welch ANOVA. p-val(S) is the p-value from safety scores, and p-val(A) is the p-value of attractiveness scores.

Inter-group Comparisons		**p-val(S)**	**p-val(A)**
Model-C1	Model-C2	<0.01**	<0.001***
Model-C1	Model-C3	<0.001***	<0.001***
Model-C1	Model-C4	<0.001***	<0.001***
Model-C2	Model-C3	0.20	0.99
Model-C2	Model-C4	0.27	0.51
Model-C3	Model-C4	0.99	0.57
Overall Test Statistics		<0.001***	<0.001***

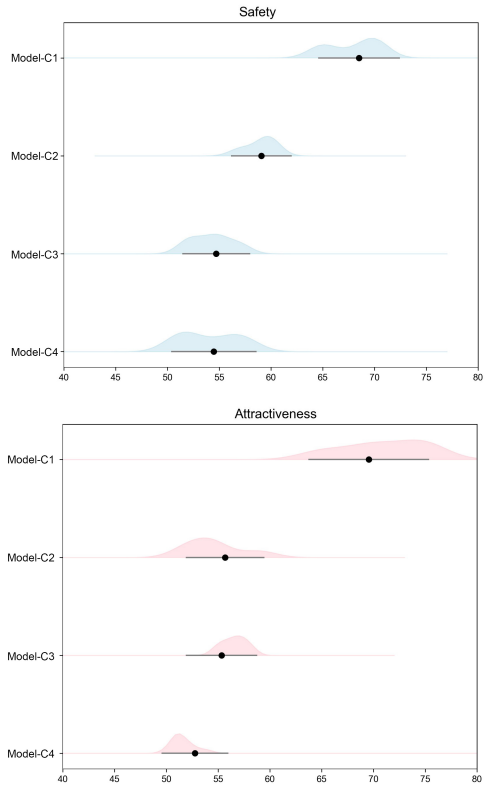

Figure 4. The score distributions of the four MLLMs assessing safety (top) and attractiveness (bottom).

From the score distributions (Figure 4) and post hoc test results (Table 3), Model-C1 produced significantly higher scores,

suggesting that in the absence of detailed instructions, the language model tends to produce more optimistic assessments. In contrast, the other three models show no significant differences in their distributions, while their median scores exhibited slight variations in different directions. This indicates that adding evaluation metrics with specific criteria can significantly influence MLLM's assessments. However, increasing the level of semantic clarity by adding descriptions, has a relatively less pronounced impact on the overall scores.

To further investigate the difference between the models with input metrics, we analyzed the scoring differences for each metric assessed by Model-C2 (with vague metrics), Model-C3 (with quantified metrics), and Model-C4 (with quantified metrics and descriptions). We conducted the Kruskal-Wallis Test for the three models. Figure 5 shows the top six metrics with the largest statistical differences under the *Safety* and *Attractiveness* assessments, ranked by their test statistics.

In most metrics shown in Figure 5, the score distribution for Model-C4 exhibited a higher degree of concentration. This trend might be attributed to the more defined descriptions provided along the metrics in Model-C4, which likely enhance the MLLM's ability to analyze according to the specific definition of each metric. To further investigate the influence of descriptions on MLLM evaluations, we examined two examples shown in Figure 6. In the *Safety* assessment, Model-C4 assigned a low score to the image because the specified traffic calming devices (e.g., speed bumps and raised crosswalks) were not visible. In contrast, Model-C2 inferred the presence of traffic calming devices based on narrow roads and markings, resulting in a more positive score. Similarly, in the attractiveness assessment, Model-C4 assigned scores based on the descriptions of *FixedFurniture*, while Model-C2 mistakenly considered crosswalks as a type of fixed furniture, leading to an incorrect evaluation.

Therefore, the more varied and dispersed score distributions in Model-C3, which uses only quantified metrics, and Model-C2, which uses vague metrics, could be attributed to the ambiguity resulting from the lack of definitions, leaving room for MLLM's interpretations. In contrast, the detailed descriptions in Model-C4 reduce ambiguity in the prompts, thereby guiding the model's interpretations with the intended evaluation criteria, resulting in higher concentration and consistency.

4.2 Discussion

We discuss the findings of the paper in three aspects. First, the results offer preliminary evidence that MLLMs can perform assessments based on the provide evaluation metrics, using general knowledge. The multimodal model can automatically provide interpretations and scoring based on images. For example, it assumed the perceived traffic flow based on the width of the road. Therefore, MLLMs are potentially helpful for large-scale studies of urban design quality assessment. Second, Model-C1 produces more optimistic scores compared to the other three models. The results of its evaluation for the street segments show discrepancies compared to the others. For example, a street that receives the lowest safety score in Model C1 is ranked mid-range in the other models. This suggests that MLLM evaluations without incorporating expert knowledge tend to be overly optimistic and may produce feedback that diverges from expert-informed evaluations. Third, informing MLLMs with expert knowledge (e.g., evaluation metrics) significantly influences their evaluative performance. For example, enhancing the semantic clarity of the walkability metrics described in the prompts helps align the MLLM interpretations with the specified criteria, leading to increased concentration and consistency. Although the addition of descriptions to the metrics has a comparatively less pronounced effect on overall scores, it can prevent MLLM from making mistakes when interpreting the provided metrics.

The paper presents our initial exploration of the risks and advantages of large language models in evaluating environmental quality and the potential to integrate expertise and domain knowledge. MLLMs offer a low entry barrier for generating automated multimodal image-text assessments of walkability and can support large-scale evaluations of the quality of urban design. While this automation holds promise for broad, data-intensive studies, it also presents risks, such as overly optimistic evaluations, misinterpretation of terminology, and inaccurate assessments. Therefore, further research is needed to better understand how MLLM can be effectively integrated into urban design quality surveys and to establish standardised guidelines for prompt formulation.

5. Conclusion

This study explores how integrating more formal and clarified representations of expert urban design knowledge into the input prompts of an MLLM (ChatGPT) can enhance its capability to evaluate walkability using SVIs. Walkability metrics were collected and categorised through the existing literature. A comparative study was conducted for MLLMs fed with prompts with varying levels of clarity and specificity, using the metrics for assessing pedestrian safety and attractiveness. The findings demonstrate that MLLMs' evaluative performance can be enhanced by integrating expert knowledge. Furthermore, increasing the semantic clarity of expert knowledge representations improves the consistency of MLLMs' evaluative outputs.

This paper has limitations of the small size of the SVI database and the limited set of metrics for walkability evaluation. These challenges mark the starting points for future enhancements and developments in this area. Our future work will be extended from the following aspects. First, we will engage urban design practitioners to evaluate the SVIs, and compare their assessment with the MLLMs, hence to investigate the difference between MLLM evaluators and human evaluators, and enhance the reliability of the MLLM assessments. Second, a more rigorous process of defining walkability metrics will be established based on a more comprehensive review of literature, through the lens of urban design assessment characteristics. Third, as currently the test dataset is limited in size, we will scale up the dataset by adding more SVI variations and provide a more rigorous selection process for SVIs in terms of the built environment characteristics. Fourth, the automated evaluation workflow could be developed to provide urban designers with practical guidelines.

Acknowledgements Part of this research was conducted at the Future Cities Lab Global at Singapore-ETH Centre. Future Cities Lab Global is supported and funded by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme and ETH Zurich (ETHZ).

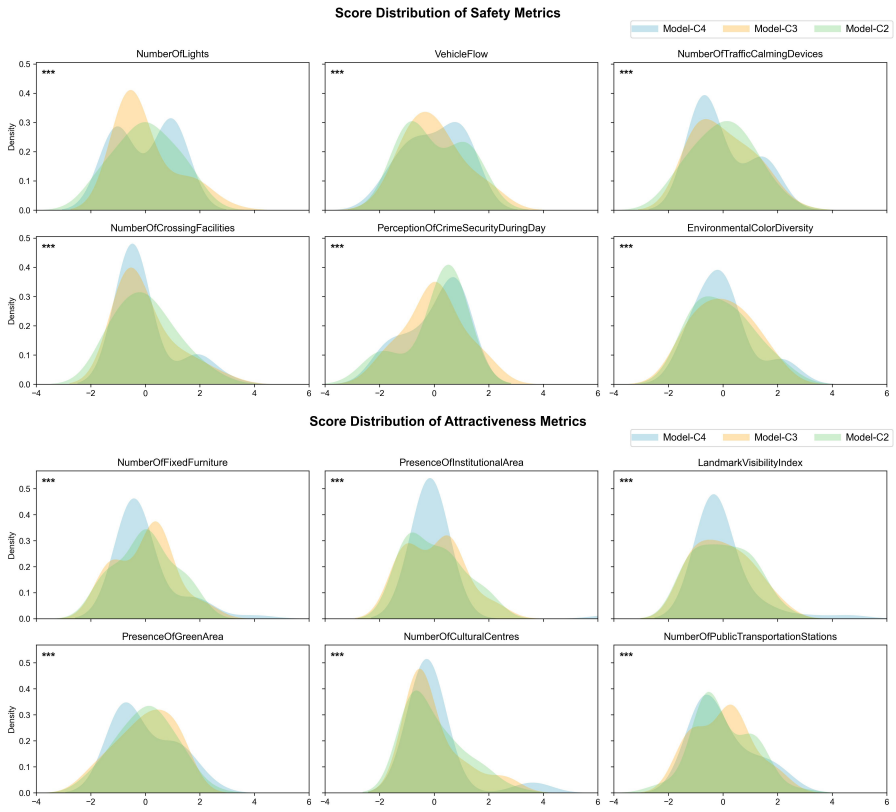

Figure 5. The top six metrics with the largest statistical differences according to safety scores (top) and attractiveness scores (bottom), ranked by their test statistics. The density plots show the level of concentration in each MLLM when measuring the particular metric.

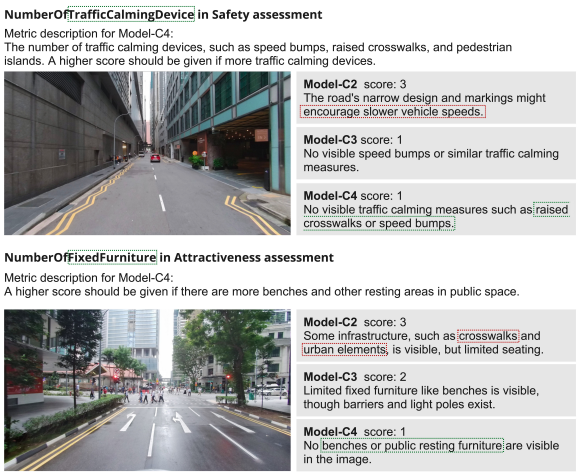

Figure 6. Two example responses from MLLMs for the metrics that differ significantly among the three models, assessing safety (top) and attractiveness (bottom). This shows the influence of descriptions to MLLM evaluations.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al., 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arellana, J., Saltarín, M., Larrañaga, A. M., Alvarez, V., Henao, C. A., 2020. Urban walkability considering pedestrians' perceptions of the built environment: a 10-year review and a case study in a medium-sized city in Latin America. *Transport reviews*, 40(2), 183–203.

Ariffin, R. N. R., Rahman, N. H. A., Zahari, R. K., 2021. Systematic literature review of walkability and the build environment. *J. Pol'y & Governance*, 1, 1.

Ataman, C., Herthogs, P., Tunçer, B., Perrault, S., 2022. Multi-Criteria Decision Making in Digital Participation.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

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

Blečić, I., Saiu, V., A. Trunfio, G., 2024. Enhancing urban walkability assessment with multimodal large language models. *International Conference on Computational Science and Its Applications*, Springer, 394–411.

Cai, C., 2025. MLLM assessments for walkability. doi.org/10.6084/m9.figshare.28236869.v1.

de Jong, T., Fyhri, A., 2023. Spatial characteristics of unpleasant cycling experiences. *Journal of transport geography*, 112, 103646.

Dragović, D., Krklješ, M., Slavković, B., Aleksić, J., Radaković, A., Zećirović, L., Alcan, M., Hasanbegović, E., 2023. A literature review of parameter-based models for walkability evaluation. *Applied Sciences*, 13(7), 4408.

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

Fan, P., Wan, G., Xu, L., Park, H., Xie, Y., Liu, Y., Yue, W., Chen, J., 2018. Walkability in urban landscapes: A comparative study of four large cities in China. *Landscape ecology*, 33, 323–340.

Fonseca, F., Ribeiro, P. J., Conticelli, E., Jabbari, M., Papageorgiou, G., Tondelli, S., Ramos, R. A., 2022. Built environment attributes and their influence on walkability. *International Journal of Sustainable Transportation*, 16(7), 660–679.

Frank, L. D., Sallis, J. F., Saelens, B. E., Leary, L., Cain, K., Conway, T. L., Hess, P. M., 2010. The development of a walkability index: application to the Neighborhood Quality of Life Study. *British journal of sports medicine*, 44(13), 924–933.

Giles-Corti, B., Vernez-Moudon, A., Reis, R., Turrell, G., Dannenberg, A. L., Badland, H., Foster, S., Lowe, M., Sallis, J. F., Stevenson, M. et al., 2016. City planning and population health: a global challenge. *The lancet*, 388(10062), 2912–2924.

Grisiute, A., Wiedemann, N., Herthogs, P., Raubal, M., 2024. An ontology-based approach for harmonizing metrics in bike network evaluations. *Computers, Environment and Urban Systems*, 113, 102178.

Herthogs, P., 2021. Triple-a design: a mid-level ontology for design goals and design evaluation. 2021.

Huang, G., Yu, Y., Lyu, M., Sun, D., Dewancker, B., Gao, W., 2024. Impact of Physical Features on Visual Walkability Perception in Urban Commercial Streets by Using Street-View Images and Deep Learning. *Buildings*, 15(1), 113.

Jacobs, J., 2010. *Dark age ahead: Author of the death and life of great American cities*. Vintage Canada.

Koohsari, M. J., Nakaya, T., Hanibuchi, T., Shibata, A., Ishii, K., Sugiyama, T., Owen, N., Oka, K., 2020. Local-area walkability and socioeconomic disparities of cardiovascular disease mortality in Japan. *Journal of the American Heart Association*, 9(12), e016152.

Larranaga, A. M., Arellana, J., Rizzi, L. I., Strambi, O., Cybis, H. B. B., 2019. Using best–worst scaling to identify barriers to walkability: A study of Porto Alegre, Brazil. *Transportation*, 46, 2347–2379.

Lee, S., Lee, C., Nam, J. W., Abbey-Lambertz, M., Mendoza, J. A., 2020. School walkability index: Application of environmental audit tool and GIS. *Journal of transport & health*, 18, 100880.

Li, Z., Wang, Y., Song, Z., Huang, Y., Bao, R., Zheng, G., Li, Z. J., 2024. What can LLM tell us about cities? *arXiv preprint arXiv:2411.16791*.

Liu, L., Sevtsuk, A., 2024. Clarity or confusion: A review of computer vision street attributes in urban studies and planning. *Cities*, 150, 105022.

Liu, X., Haworth, J., Wang, M., 2023. A new approach to assessing perceived walkability: Combining street view imagery with multimodal contrastive learning model. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications*, 16–21.

Malekzadeh, M., Willberg, E., Torkko, J., Toivonen, T., 2025. Urban attractiveness according to ChatGPT: Contrasting AI and human insights. *Computers, Environment and Urban Systems*, 117, 102243.

Pelclová, J., Frömel, K., Cuberek, R., 2014. Gender-specific associations between perceived neighbourhood walkability and meeting walking recommendations when walking for transport and recreation for Czech inhabitants over 50 years of age. *International journal of environmental research and public health*, 11(1), 527–536.

Pereira, M. F., Almendra, R., Vale, D. S., Santana, P., 2020. The relationship between built environment and health in the Lisbon Metropolitan area–can walkability explain diabetes' hospital admissions? *Journal of Transport & Health*, 18, 100893.

Reisi, M., Nadoushan, M. A., Aye, L., 2019. Local walkability index: Assessing built environment influence on walking. *Bulletin of Geography. Socio-economic Series*, 7–21.

Wedyan, M., Saeidi-Rizi, F., 2025. Assessing the impact of walkability indicators on health outcomes using machine learning algorithms: A case study of Michigan. *Travel Behaviour and Society*, 39, 100983.

Yin, L., 2017. Street level urban design qualities for walkability: Combining 2D and 3D GIS measures. *Computers, Environment and Urban Systems*, 64, 288–296.

Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, M., Liu, Y., Batty, M., Ratti, C., 2024. Urban Visual Intelligence: Studying Cities with Artificial Intelligence and Street-Level Imagery. *Annals of the American Association of Geographers*, 114(5), 876–897.