Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39246

Multimodal Artificial Intelligence Based Evaluation Model for Smart Public Space Experience

Shijia Zhang

College of Art and Design, Glasgow School of Art, Glasgow, UK
zhangshijia0418@163.com

Abstract. Smart public spaces combine sensing infrastructures, digital services, responsive facilities, and physical spatial elements, producing experience evidence that cannot be adequately represented by questionnaires or isolated environmental indicators. This study develops a multimodal artificial intelligence model that evaluates comfort, perceived safety, spatial engagement, accessibility, and an integrated experience score from synchronized behavioral video, UWB trajectories, environmental sensing, and textual feedback. Data are collected from 120 participants completing four eight-minute spatial conditions in an instrumented campus plaza. Modality-specific encoders generate 128-dimensional representations, which are temporally aligned and integrated by a four-layer cross-modal Transformer with hierarchical regression heads. On participant-independent testing, the proposed model achieved an overall MAE of 0.044 ± 0.004, RMSE of 0.061 ± 0.005, R^2 of 0.905 ± 0.012, and Spearman ρ of 0.951 ± 0.008, significantly outperforming early-fusion and single-modality baselines ($p < .001$). Removing trajectory or textual information increased overall MAE to 0.061 ± 0.006 and 0.059 ± 0.005, respectively. The model also retained stable performance across crowd-density and lighting conditions, while SHAP analysis identified distinct modality contributions to each experience dimension.

Keywords: Multimodal Artificial Intelligence, Smart Public Space, Experience Evaluation, Cross-Modal Fusion, Spatial Behavior

1. Introduction

The latest trend towards smart public spaces includes digital information, adaptive lighting, environmental sensing, interactive facilities, and traditional spatial design. Assessment of such public spaces should not rely solely on physical measures, as users constantly alter their paths, dwelling patterns, interpersonal distance, attention, and interaction patterns. Current literature on the assessment of public open spaces demonstrates the use of computer vision along with multi-source urban data to augment the traditional audit and subjective assessments [1]. Systematic reviews suggest that the application of machine learning becomes more and more important to assess user behavior, experience and opinions in public spaces [2]. User experience and behavior in public spaces depend also on environmental factors and affective experience, which is associated with the differences in activity distribution and engagement [3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39246

Urban sensing can contribute to the body of evidence. Urban sensing enables the collection of data on pedestrian movement and analysis of flows and trajectories [4] and crowdsourced movement data can show significant temporal and spatial differences in walking behavior [5]. However, sensing alone cannot provide a full picture of experience. Research on interpretability of machine learning models demonstrates that urban visual features can predict several perceptual dimensions, and also provides an explanation why black-box models need to be explained in terms of input features [6]. Literature on safety research shows that visual AI perception and questionnaire-based experience measure distinct aspects of environment evaluation [7]. Explanations of deep learning models trained on street imagery helped researchers to correlate environmental features and pedestrians' safety perception [8].

Human-centered urban analytics also need models that preserve person-environment interactions. Spatiotemporal GeoAI demonstrated the advantages of combining computer vision, sequential movement and contextual information to predict outdoor comfort [9], and digital participatory evaluation highlighted the importance of including subjective user evidence in public space evaluation [10]. What remains as a methodological issue is the separation of behavior, trajectories, environmental measurement, and text perception. This study solves this problem through the use of modality-specific encoding, temporal alignment, cross-modal attention, hierarchy of dimensions prediction, and attributions.

2. Literature review

Public-space quality is commonly evaluated through dimensions such as comfort, safety, accessibility, activity, and satisfaction. Automated public-space assessment has demonstrated that imagery and computer vision can supplement conventional geographic indicators [1], whereas recent reviews identify user perception, demographic variation, and data acquisition as persistent challenges for machine-learning-based public-space research [2]. Behavioral evidence also matters because environmental attributes and affective responses influence how long and where pedestrians choose to engage in public-space activities [3].

Contemporary urban sensing provides continuous evidence that questionnaires cannot capture alone. Pedestrian-monitoring research shows that sensor data can quantify flows and movement patterns [4], and large-scale movement datasets can reveal temporal variability in route use [5]. Visual machine learning additionally supports multidimensional perception prediction and interpretable analysis of environmental features [6]. Yet perception inferred from visual appearance does not necessarily coincide with lived evaluation, particularly for safety [7]. Explainable AI methods therefore become important when model outputs are intended to support planning decisions [8].

A multimodal framework should consequently preserve the structure of each evidence source before integration. Human-centered GeoAI demonstrates that sequential movements and environmental context can jointly represent outdoor comfort [9], while participatory digital evaluation emphasizes that subjective feedback remains necessary for interpreting public-space quality [10]. For smart public spaces, video behavior, trajectories, sensor conditions, and textual feedback represent complementary rather than interchangeable evidence. Their timestamps must first be aligned, then their dependencies learned across modalities. A hierarchical output is preferable to one undifferentiated score because comfort, safety, engagement, and accessibility may respond to different features. The resulting model therefore combines cross-modal fusion with dimension-level prediction, aggregate scoring, missing-modality robustness, and interpretable attribution.

DOI: 10.54254/2755-2721/2026.39246

3. Methodology

3.1. Multidimensional experience evaluation structure

The evaluation objective encompasses comfort, safety perception, spatial engagement, and accessibility, all on a continuous 0 to 1 scale. Comfort encompasses behavioral consistency and environmental exposure and physical ease. Safety is a combination of route hesitation, interpersonal distance, scanning behavior, and questionnaire evaluation. Spatial engagement includes interaction frequency, spatial exploration extent, visual-object attention, and activity duration, but accessibility encompasses route efficiency, navigation mistakes, destination attainment, and perceived ease of movement. Supervised ground truth is provided by normalizing the questionnaire scores, but observable metrics are used as explanatory measures.

Predictions for dimensions are generated independently before the final aggregation. Four latent embeddings for dimensions are obtained after multimodal fusion, and an aggregation layer estimates the contributions of each of the four dimensions towards overall experience. Scores are clipped using sigmoid activation to keep the aggregation measure tied to the four dimension-level predictions.

3.2. Multimodal data acquisition and temporal alignment

Behavior is recorded with four fixed RGB cameras at 25 frames per second. MediaPipe Pose detects 33 landmarks, and OpenCV calculates orientation of the torso, intensity of the locomotion, interaction events, and interpersonal distance. Landmarks are removed if the visibility is less than 0.60, and the gaps less than 400 milliseconds are linearly interpolated. Two-dimensional position with Sewio UWB tag is detected at 10 Hz, and speed, acceleration, curvature, dwell time, coverage, route efficiency, and repeated path ratio are measured. Environmental sensing is performed with BME688, BH1750, and INMP441 modules for temperature, humidity, illumination, and sound pressure at 1 Hz. Crowd density detection uses synchronous sensing, and participants provide feedback in 50-100 words after each condition. The recordings are timestamped using Network Time Protocol and divided into overlapping 20-second windows (10-second overlap). Each condition of eight minutes results in 47 windows, which provides 22,560 multimodal samples per 120 participants and four conditions.

3.3. Modality-specific feature encoding

Pose maps and RGB crops with respect to the person-of-interest define behavioral sequences. Swin Transformer Tiny operates on eight frames uniformly sampled within each window, producing a 128-dimensional embedding using attention pooling. UWB trajectories are downsampled to 10 Hz and fed into a Temporal Convolutional Network with three layers of kernel size 3 and dilations 1, 2, and 4.

Environmental variables are standardized and encoded by a two-layer MLP with 64 and 32 hidden units, GELU activation, and 0.15 dropout. Text is truncated to 128 tokens and encoded by RoBERTa-base; its contextual [CLS] representation is projected to 128 dimensions. Each modality receives a learned modality token and temporal-position encoding. Layer normalization places all embeddings on a comparable scale without collapsing their modality-specific temporal or semantic structures.

DOI: 10.54254/2755-2721/2026.39246

3.4. Cross-modal fusion and hierarchical score prediction

The four modality tokens enter a four-layer cross-modal Transformer with eight attention heads, model width 128, and feed-forward width 512. Cross-attention allows contextual evidence to modify another modality, such as interpreting route hesitation differently under high and low crowd density. Fusion is calculated as shown in Equation (1)-(4):

$$\mathbf{Z}^{(l)} = LayerNorm[\mathbf{Z}^{(l-1)} + Concat_{h=1}^{H}(softmax(\frac{\mathbf{Q}_h\mathbf{K}_h^T}{\sqrt{d_h}} + \mathbf{A} + \mathbf{M})\mathbf{V}_h)\mathbf{W}_O] \tag{1}$$

$$\mathbf{Q}_h = \mathbf{Z}^{(l-1)}\mathbf{W}_h^Q \tag{2}$$

$$\mathbf{K}_h = \mathbf{Z}^{(l-1)}\mathbf{W}_h^K \tag{3}$$

$$\mathbf{V}_h = \mathbf{Z}^{(l-1)}\mathbf{W}_h^V \tag{4}$$

Here, $\mathbf{A}$ is the temporal-alignment bias and $\mathbf{M}$ masks unavailable modalities. One modality is randomly masked with probability 0.15 during training to improve incomplete-record robustness. As shown in Figure 1, the proposed framework integrates multimodal data acquisition, modality-specific encoding, temporal alignment, cross-modal Transformer fusion, hierarchical dimension prediction, and interpretable overall experience evaluation within a unified architecture.

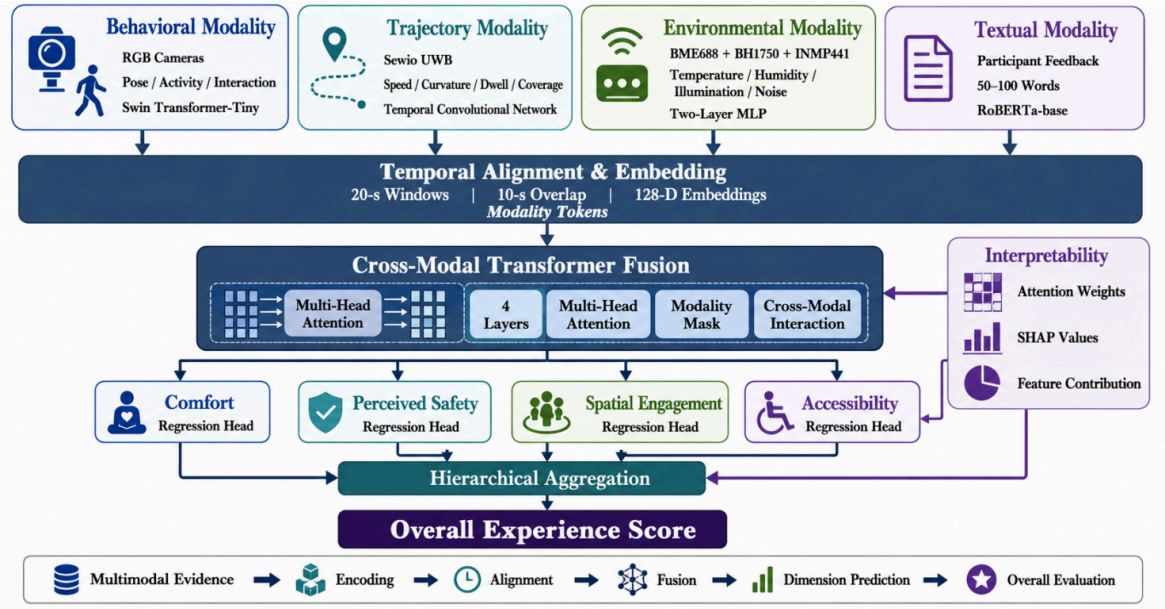


Figure 1. Multimodal AI evaluation architecture for smart public space experience

DOI: 10.54254/2755-2721/2026.39246

4. Experimental design

4.1. Experimental site and spatial conditions

Testing is carried out in a 40 m x 30 m instrumented campus plaza equipped with seating, walking paths, digital information screens, and lighting. Four different experimental conditions, structured as 2 × 2, will be created by varying crowd density (0.05 and 0.20 persons per m^2) and lighting mode (static and responsive). All other variables—geometry, task destinations, display messages, and seating—will be the same across conditions.

4.2. Participants and experimental procedure

A total of 120 subjects will participate in all four conditions in counterbalancing sequence. The four-minute experiment condition will require that they locate an appropriate facility, interact with the information screen, and choose a place to rest. Participants will subsequently answer the experience questionnaire and provide textual comments prior to a five-minute interlude. The data will be split according to participants: 84 participants will form the training sample, 18 the validation sample, and 18 the test sample. It means there will be 15,792 training windows and 3,384 windows in each evaluation set.

4.3. Model comparison and evaluation indicators

The proposed model is compared with support vector regression, XGBoost, a trajectory-only LSTM, and an early-fusion MLP under identical participant splits. MAE, RMSE, (R^2), and Spearman correlation evaluate prediction. Neural models are trained with five random seeds and reported as mean ± SD. Ablation removes one modality at a time. Cross-condition testing holds out each crowd-lighting combination. Repeated-measures ANOVA and Bonferroni-adjusted comparisons test model differences, while SHAP quantifies modality and feature contributions.

5. Results

5.1. Multidimensional experience score prediction performance

The proposed model achieved dimension-level MAE values of 0.041 ± 0.004 for comfort, 0.046 ± 0.005 for safety, 0.043 ± 0.004 for engagement, and 0.039 ± 0.003 for accessibility (see Table 1). Overall model differences were significant, $F(4,85)=39.72$, $p<.001$.

Table 1. Overall experience prediction performance

Model	MAE	RMSE	R^2	Spearman
SVR	0.083 ± 0.007	0.109 ± 0.009	0.713 ± 0.031	0.842 ± 0.018
XGBoost	0.071 ± 0.006	0.094 ± 0.008	0.778 ± 0.027	0.879 ± 0.015
Trajectory LSTM	0.069 ± 0.005	0.092 ± 0.007	0.786 ± 0.023	0.884 ± 0.014
Early-fusion MLP	0.061 ± 0.005	0.083 ± 0.006	0.825 ± 0.019	0.902 ± 0.012
Proposed model	0.044 ± 0.004*	0.061 ± 0.005*	0.905 ± 0.012*	0.951 ± 0.008*

*Note: ***(p<.001) versus early-fusion MLP.*

DOI: 10.54254/2755-2721/2026.39246

5.2. Multimodal fusion and modality ablation results

Removing any modality degraded prediction. Trajectory removal produced the largest overall error, whereas environmental removal primarily affected comfort and behavioral removal most strongly affected engagement (see Table2).

Table 2. Dimension-level MAE under modality ablation

Configuration	Comfort	Safety	Engagement	Accessibility	Overall
Full model	0.041 ± 0.004	0.046 ± 0.005	0.043 ± 0.004	0.039 ± 0.003	0.044 ± 0.004
Without behavior	0.049 ± 0.004**	0.054 ± 0.005**	0.068 ± 0.007***	0.052 ± 0.005**	0.057 ± 0.005***
Without trajectory	0.053 ± 0.005**	0.061 ± 0.006***	0.059 ± 0.006***	0.071 ± 0.007***	0.061 ± 0.006***
Without environment	0.066 ± 0.006***	0.052 ± 0.005*	0.048 ± 0.004*	0.047 ± 0.004*	0.053 ± 0.004***
Without text	0.052 ± 0.005**	0.069 ± 0.007***	0.061 ± 0.006***	0.054 ± 0.005**	0.059 ± 0.005***

*Note: *(p<.05), **(p<.01), ***(p<.001) relative to the full model.*

5.3. Cross-condition generalization and model interpretability

Held-out-condition MAE of the model was 0.054 ± 0.006 for the low crowd/static lighting environment, 0.058 ± 0.006 for the high crowd/static lighting environment, 0.051 ± 0.005 for the low crowd/responsive lighting environment, and 0.063 ± 0.007 for the high crowd/responsive lighting environment. The mean absolute SHAP values determined that the environmental factor contributed the most to the comfort level (0.024 ± 0.005), the text contributed the most to the safety level (0.021 ± 0.004), the behavior contributed the most to the engagement level (0.026 ± 0.006). Route efficiency contributed 0.018 ± 0.004 to accessibility, while local crowd density contributed 0.017 ± 0.004 to safety, indicating dimension-specific rather than uniform multimodal weighting.

6. Conclusion

This study develops a multimodal AI model for evaluating smart public-space experience from synchronized behavior, trajectory, environment, and textual evidence. Modality-specific encoding and temporal alignment preserve heterogeneous information before cross-modal Transformer fusion, while hierarchical prediction maintains separate comfort, safety, engagement, and accessibility scores before overall aggregation. The proposed model produced lower prediction error than conventional regression, trajectory-only learning, and early fusion. Ablation and SHAP results further showed that each modality contributed differently across experience dimensions. The participant-independent split and cross-condition tests also support robustness beyond repeated observations from the same individuals or one crowd-lighting setting. The main methodological contribution is therefore the transformation of heterogeneous spatial evidence into temporally corresponding, modality-aware, and interpretable experience predictions.

DOI: 10.54254/2755-2721/2026.39246

References

[1] Chen, S., & Biljecki, F. (2023). Automatic assessment of public open spaces using street view imagery. *Cities, 137*, 104329.
[2] Zhu, Y., Zhang, Y., & Biljecki, F. (2025). Understanding the user perspective on urban public spaces: A systematic review and opportunities for machine learning. *Cities, 156*, 105535.
[3] Yang, S., Dane, G. Z., & Arentze, T. A. (2024). Modeling pedestrians' activity time-use choices in a virtual urban public space: The influences of the environment and affective experience. *Cities, 154*, 105399.
[4] Han, C., Seshadri, P., Ding, Y., Posner, N., Koo, B. W., Agrawal, A., Lerch, A., & Guhathakurta, S. (2024). Understanding pedestrian movement using urban sensing technologies: The promise of audio-based sensors. *Urban Informatics, 3*, 22.
[5] Angel, A., & Plaut, P. (2024). Tempo-spatial analysis of pedestrian movement in the built environment based on crowdsourced big data. *Cities, 149*, 104917.
[6] Liu, Y., Chen, M., Wang, M., Huang, J., Thomas, F., Rahimi, K., & Mamouei, M. (2023). An interpretable machine learning framework for measuring urban perceptions from panoramic street view images. *iScience, 26*(3), 106132.
[7] Kang, Y., Abraham, J., Ceccato, V., Duarte, F., Gao, S., Ljungqvist, L., Zhang, F., Näsman, P., & Ratti, C. (2023). Assessing differences in safety perceptions using GeoAI and survey across neighbourhoods in Stockholm, Sweden. *Landscape and Urban Planning, 236*, 104768.
[8] 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, 195*, 107400.
[9] Liu, P., Zhao, T., Luo, J., Lei, B., Frei, M., Miller, C., & Biljecki, F. (2023). Towards human-centric digital twins: Leveraging computer vision and graph models to predict outdoor comfort. *Sustainable Cities and Society, 93*, 104480.
[10] Alvarado Vazquez, S., Ostermann, F. O., Pinto Soares Madureira, A. M., & Pfeffer, K. (2025). Using a digital participatory platform to evaluate public space quality. *GeoJournal, 90*(3), 145.