A Care-Centered Vision–Language Model Framework for Evaluating Urban Under-Bridge Spaces for Older Adults

Zian Huang[1], Siwei Zhang[2], and Xingyu Tao[3]*

[1] The Hong Kong Polytechnic University, Hong Kong, China
[2] The University of Hong Kong, Hong Kong, China
[3] The Hong Kong Polytechnic University, Hong Kong, China
zi-an.huang@connect.polyu.hk, siweiz@connect.hku.hk, xingyu.tao@polyu.edu.hk

Abstract

In many urban areas, under-bridge spaces have become common gathering spots for older adults, offering venues for leisure, social interaction, and physical activity. Various policies and community initiatives have been introduced to improve the quality and usability of these spaces. However, efficiently and automatically assessing the elder-friendliness of such environments remains an open challenge, particularly for supporting data-informed spatial design and policy decisions. To this end, this study proposes a Care-Centered Vision–Language (CC-VLM) framework with three contributions: (i) Develop a care-theory-centered evaluation metric model, using this as the basis for subsequent labeling and assessment; (ii) Design a multi-elder-agent workflow for automatically generating labels to solve the problem of data scarcity from older adults; and (iii) Build a XGboost-based classification model for automated under-bridge spaces scene assessment. The proposed framework is applied to 5 under-bridge spaces in Hong Kong. Experimental results demonstrate an accuracy of 65.5% and a recall rate of 71.4%. This framework offers a scalable and interpretable approach to support the inclusive and age-friendly transformation of overlooked urban spaces.

Keywords –

Care Theory; Older Adults; VLM; Under-Bridge;

1 Introduction

Globally, under-bridge spaces in high-density cities have emerged as vital gathering spots for older adults, providing venues for leisure, social interaction, and physical activity. Many regions face a pressing contradiction: a growing and aging population drives sustained demand for public recreational spaces, while existing provision remains insufficient and is projected to fall further behind. For instance, Hong Kong—representative of such high-density urban contexts—has seen its elderly population proportion projected to rise from 18.4% in 2029 to 36.3% in 2064, alongside a long-standing shortfall in per capita open space (failing to meet its own 2m ²standard since 2008) [1-3]. The current use of under-bridge spaces is subject to government regulations [4] that limit permissible uses. However, official policy exhibits a duality of control and activation: while protecting safety, traffic, environmental, and structural concerns, it also actively encourages greening, beautification, and adaptive reuse. Notable examples include the "Fly the Flyover Operation," organised by the Development Bureau's Energising Kowloon East Office [5] and statements in the Official Record of Proceedings from the Legislative Council [6]. This situation indicates the value of utilizing remaining urban spaces, such as under-bridge areas. But because they are so numerous, automated assessment tools become particularly important.

Existing research on evaluating urban spaces has followed two main paradigms: qualitative approaches and quantitative data-driven methods. Qualitative studies provide rich contextual insights into user experiences, urban space design theory, and elderly-specific needs, emphasising Inclusive, Pleasurable, Safe, Comfortable, and Meaningful Activities [7]. There are numerous qualitative assessment tools for public space quality, among which the UN-Habitat and the Good Public Space Index (GPSI) are considered the most comprehensive; most of these tools rely on on-site observation, questionnaire surveys, and site visits [8]. Those methods are subjective, labour-intensive, and limited in scale, and these drawbacks are amplified when studying older adults due to their physical, cognitive, and communication challenges. Quantitative methods, largely enabled by street-view imagery and computer vision, can partially compensate for the shortcomings of qualitative research. Common quantitative methods employ semantic segmentation (e.g., PSPNet, SegNet, Swin Transformer) to extract features, such as using a transformer-based framework for urban physical disorder detection [9]. Quantitative methods also include using statistical analysis, machine learning, or deep learning to predict some socioeconomic profiles, like using the spatial distribution of notable urban features identified through the street view images to predict poverty status, crime, and health behaviour [10]. However, these quantitative studies typically require extremely large sample sizes—often involving datasets with over a million street-view images sourced from platforms like Google Street View or Baidu Maps [9, 10]. Such data demands pose significant barriers for niche, overlooked spaces like under-bridges, where imagery is sparse, site-specific, and difficult to collect at scale due to accessibility constraints, variable lighting, and limited coverage in mapping services. Furthermore, the qualitative method cannot elicit people's subjective perceptions of space. Although qualitative analysis can achieve this, its limitations have been discussed previously.

Vision-Language Models (VLMs) represent a paradigm shift in multimodal AI. By jointly encoding images and text through aligned embeddings (typically involving an image encoder, text encoder, fusion mechanism, and generative language component), VLMs enable advanced tasks such as image captioning, visual question answering (VQA), and zero/few-shot reasoning on novel concepts [11, 12]. VLMs can provide language-guided interpretability. VLMs can also serve as simulated "elderly agents" through targeted prompting, automating subjective evaluations from an older adult's perspective and mitigating the inherent drawbacks of qualitative methods noted earlier. Recent applications show the effectiveness of VLMs in urban scene recognition and street-view analytics. Systematic reviews emphasize VLMs' role in assessing street-level perceptions, including safety, walkability, greenery, and socioeconomic indicators from imagery sources like Google Street View [13]. The spatial reasoning of VLM can be further enhanced by tuning particularly for challenging tasks involving negation or counterfactuals [14]. In addition, VLM can be integrated with broader frameworks for urban renewal diagnostics, combining perceptual scoring with spatial autocorrelation to inform planning decisions [15]. However, off-the-shelf VLMs are not directly suited to evaluating elderly suitability in under-bridge spaces. Pre-training datasets under-represent neglected residual areas, leading to inconsistent performance on atypical layouts and fine details [13, 14]. General-purpose VLMs also lack grounding in elderly-specific care concepts and remain largely vision-centric, with limited handling of specific factors critical for older adults.

This paper, therefore, proposes a Care-Centered Vision–Language Model (CC-VLM) for comprehensive, interpretable, scalable, and elderly-specific assessment of under-bridge spaces. Three sub-objectives are:

1. Develop a care-theory-centered evaluation metric model through a literature review on care and urban design theory, combined with on-site observation, to provide theoretical grounding and a novel perspective on under-bridge spaces assessment.

2. Design a multi-elder-agent workflow to automatically generate diverse visual labels from an elderly perspective, aiming to reflect older adults' perception of the environment. This can overcome subjectivity and labour intensity, particularly acute when elderly participants face physical, cognitive, or communication barriers.

3. Build an XGboost-based classification model for automated scene assessment, integrating care-theory metrics and multi-elder-agent labels to deliver quantitative elderly-friendliness scores and explanations.

2 Methodology

CC-VLM consists of three stages:

(1) Care-Theory-Centered Evaluation Metric Development and Feature Extraction. A hierarchical indicator system is built from care ethics literature and field observation, yielding 27 measurable spatial indicators grouped into three layers (Physical Shelter, Psychological Belonging, Social Activity). Then, a Detectron2 model with a ResNet-101-FPN backbone pre-trained on COCO will be established to perform Panoptic segmentation. According to the Evaluation Metric, Custom class merging and post-processing yield pixel-level ratios, instance densities, and morphological metrics for the 27 indicators.

(2) Multi-Elder-Agent Scoring and Demographic Weighting. Based on the metric developed in step 1, four distinct elderly personas (differing in age, mobility, and sensory profile) are simulated via structured prompts to MiniCPM. Each agent scores the three care layers on a 1–100 scale. Final image scores are computed as a weighted average based on Hong Kong's official elderly age distribution, yielding a continuous elderly-friendliness score.

(3) Assessment Model Building and Interpretation. The extracted features and weighted scores are used to train an XGBoost binary classifier. The CC-VLM framework is shown in Figure 1.

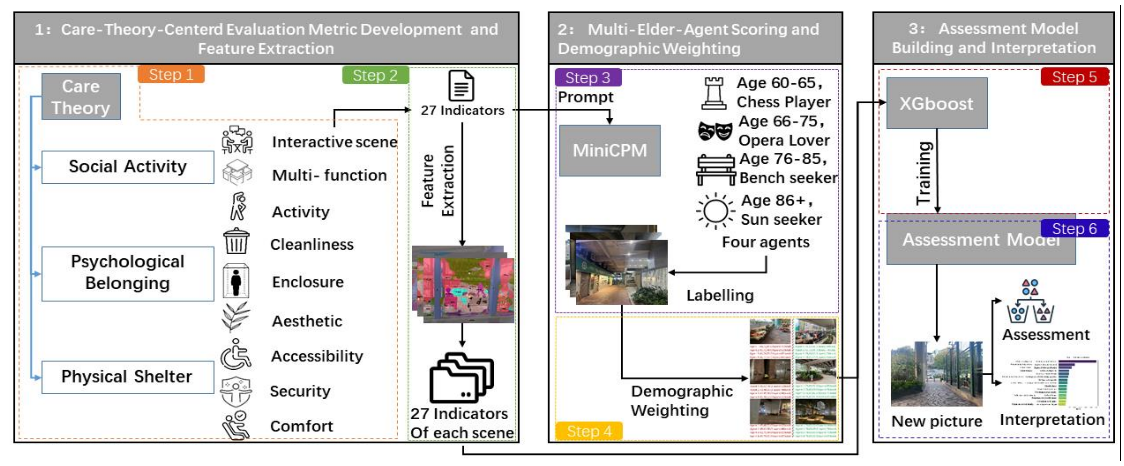

Figure 1. CC-VLM framework.

2.1 Care-Theory-Centered Evaluation Metric Development and Feature Extraction

2.1.1 Care-Theory-Centered Evaluation Metric Development

To address the limitations of existing public space evaluation tools, this study constructs a dedicated Care-Theory-Centered Evaluation Metric through theoretical and methodological synthesis. Drawing on the care ethics in urban contexts, the framework integrates three progressive phases: (1) political ethics, rooted in foundational manifestos emphasizing interdependence and universal care as a societal organizing principle [16]. (2) empirical studies, informed by care infrastructures in high-density cities, such as Boston's labor-focused activations and relational interventions [17], framework for a revised approach to intergenerational public space design, pocket park in Los Angeles [18]. and (3) design practice, exemplified by relational and caring city models in London that translate ethics into actionable placemaking [19]. This evolutionary trajectory is complemented by established urban design standards, including environmental psychology principles, WHO Age-Friendly Cities guidelines, and Hong Kong-specific urban design directives, which collectively emphasize multidimensional human needs in public realms [7, 8]. The structure of the metric is shown in Table 1.

2.1.2 Feature Extraction

This step contains two parts that transform raw street-view images into quantifiable indicators.

Part 1: Visual Analysis Leverages Detectron2's panoptic segmentation model with a ResNet-101 Feature Pyramid Network (FPN) backbone ShuffleNet), ResNet demonstrated superior overall performance [20]. Besides, it can also cover common item categories.

Table 1. Care-Theory-Centered Evaluation Metric

Primary	Secondary	Tertiary
Physical Shelter	Comfort	Seat, Green rate, Microclimate.
	Security	Physical disorder, Visual transparency, Safety protection facilities.
	Accessibility	Accessible space, Accessibility facilities, pedestrian paths.
Psychological Belonging	Aesthetic experience	Disorderliness, Color richness, Green rate,
	Enclosure sense	semi-enclosed spaces, Visibility of sky, Boundary closure.
	Cleanliness	Garbage coverage rate, Cleanliness, Ground pavement.
	Activity Support	Activity space area, Social furniture, Pedestrian activity.
Social Activity	Multiple function	Convenience facilities, Non-motorized Transport, Convenience of Access.
	Interactive scene	Recreational facilities, Stay friendliness, Openness of Public Spaces.

Part 2: Logical Extension Transform visual outputs into the indicators through systematic post-processing and mapping. Three kinds of representative indicators across distinct calculation categories are as follows:

Greenery Coverage Rate (turn pixel data into ratio indicators). This indicator reflects the proportion of vegetation in spaces and serves as a core metric for evaluating ecological comfort for older adults. Its calculation formula is shown as equation 1:

$$G = 100 \times \frac{M_{tree} \cup M_{grass}}{N} \tag{1}$$

Where $M_{tree} \cup M_{grass}$ represents the total pixel count of the merged "tree " and "grass " categories in the panoptic segmentation mask. N represents the total pixel count of the image.

Bench Density (turn count into density indicators). This indicator measures the supply of elderly-friendly resting facilities. It can be shown as equation 2:

$$P = 100 \times \frac{count(bench)}{A}, A = N \times 0.01^2 \tag{2}$$

Where P represents the bench density in the space per 100 m²; $count(bench)$ represents the total instance count of the "bench" category in the panoptic segmentation result. A is physical area of space. 0.01 is the conversion coefficient from pixel to meter.

Semi-Enclosed Space Quantity (turn spatial features into morphological indicators). This indicator characterizes the scale of enclosed local areas in under-bridge spaces. It can be shown as equation 3:

$$E = num\{S | S \geq ENCLOSED\_AREA\_THRESH\} \tag{3}$$

Where E is the number of semi-enclosed spaces. S is the area of a single enclosed region in the under-bridge space. $ENCLOSED\_AREA\_THRESH$ is the area threshold for defining a semi-enclosed space and was set to 1.5 m². According to field surveys, enclosed areas smaller than 1.5 m² are mostly structural gaps with no actual rest value, so this threshold filters invalid regions. OpenCV is used as the core computer vision tool, it extracts the outer contours of discrete spatial regions and computes the pixel area of each region.

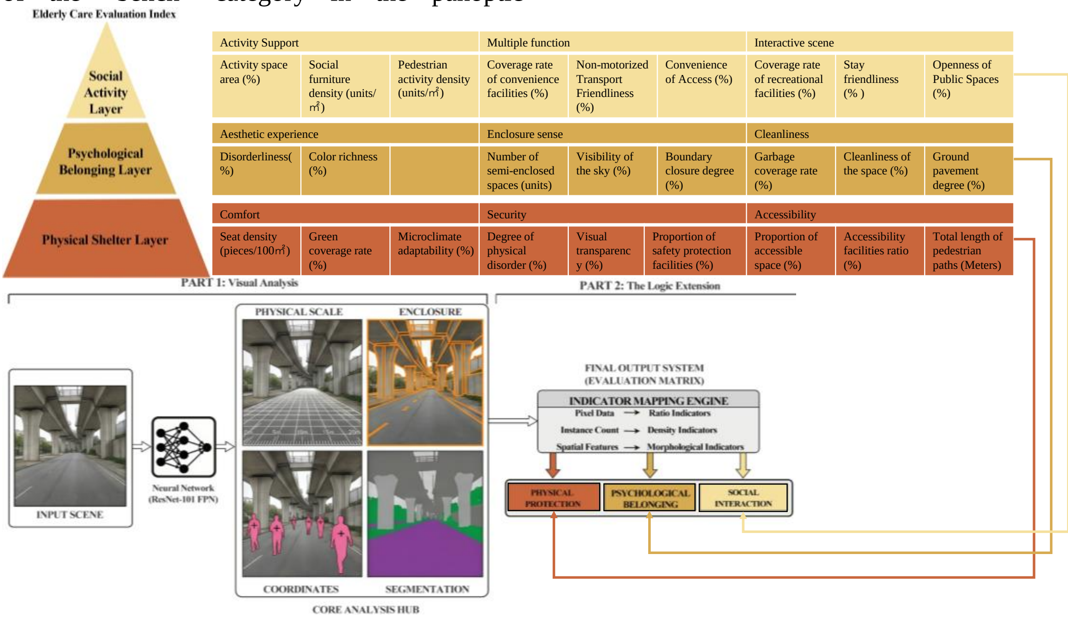

Figure 2. Feature extraction based on care-theory-centered evaluation metric

2.2 Multi-Elderly-Agent Scoring and Demographic Weighting

Existing research has realized task splitting and collaborative execution of BIM spatial queries through an LLM-based agent framework [21]. Drawing on this idea, this study designs multi-elderly agents to simulate the subjective evaluation perspectives of elderly groups.

This step addresses the subjectivity and small-scale limitations of traditional qualitative evaluations by simulating diverse elderly perspectives through AI agents and ensuring that scores reflect the needs of the target population via demographic weighting.

2.2.1 Design of Multi-Elderly Agents

To comprehensively cover the heterogeneous needs of older adults across different age groups and physical conditions, 4 elderly agents were designed through a prompt engineering approach.

Prompt Engineering: Each persona receives a structured prompt template, assigned a realistic background and key concerns derived from care theory, to ensure the agents' evaluations mirrored real-world elderly needs:

'You are a [age]-year-old elderly person with [mobility/sensory profile]. Evaluate the attached under-bridge space image from your perspective according to three dimensions: (1) Physical Shelter (accessibility, safety, comfort), (2) Psychological Belonging (aesthetics, enclosure, cleanliness), (3) Social Activity (activity support, functional diversity, interaction affordance). Provide three scores from 1–100.'

Model Deployment and Optimization: The MiniCPM model was adopted for agent reasoning, optimized for batch processing with 8-bit NF4 quantization and image resizing to balance efficiency and accuracy. Each agent processed 302 images, generating three-dimensional scores and an average score per image.

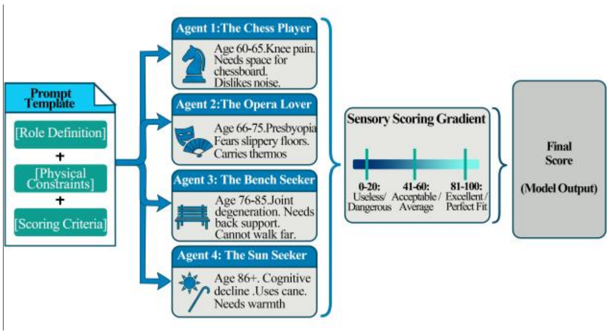

Figure 3. Multi-Elderly-Agent Scoring Framework.

2.2.2 Demographic Weighting

To reflect the actual age structure of Hong Kong's elderly population, weighted aggregation was performed on the four agents' total scores. The weighted total score for each under-bridge space was calculated as equation 4:

$$Weighted\ Score = \sum_{i=1}^{4} (S_i \times W_i) \tag{4}$$

In which S_i is the average total score of the i-th elderly agent, and W_i is the corresponding demographic weight. According to the statistics from the Census and Statistics Department of Hong Kong, the age structure of the elderly population in Hong Kong is shown in Table 2.

Table 2. The age structure of the elderly population in Hong Kong

Ages	Ratio (%)
60-65	26
66-75	43.7
76-85	19.8
Above 85	10.5

2.3 XGboost-Based Classification Model And Model Performance Analysis

This section integrates objective spatial features and demographic-weighted elderly satisfaction scores to construct a robust, interpretable, and scalable automated assessment model. The core goal is to learn the mapping relationship between under-bridge space attributes and elderly-friendly degrees, while providing data-driven decision support for age-friendly urban design or renewal.

As shown in equation 5, given the input feature matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ (where $n = 302$, the number of under-bridge image samples, $m = 27$, the quantifiable spatial indicators) and the demographic-weighted satisfaction score $S_\omega \in [1, 100]$ for each sample, the task is formulated as a binary classification problem to predict elderly satisfaction levels. The binary label y_i is defined using a calibrated threshold $\tau = 68$ (the average score determined by elderly agent). Similar urban application employs XGBoost on street-view features to predict public-space vitality scores and then reference the dataset mean to derive high or low classifications for age-inclusive urban design interventions [23].

$$y_i = \begin{cases} 0, S_{\omega,i} < \tau (Low\ Elderly - Care) \\ 1, S_{\omega,i} \geq \tau\ (Low\ Elderly - Care) \end{cases} \tag{5}$$

2.3.1 Model Selection and Building

Extreme Gradient Boosting (XGBoost) is selected as the base model due to its superior performance in handling high-dimensional tabular data, nonlinear feature interactions, and class imbalance. Its regularized objective function balances fitting capability and generalization. It can be expressed as equation 6.

$$L(\varphi) = \sum_{i=1}^{n} l(y_i , \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

Where $\sum_{i=1}^{n} \boldsymbol{l}(\boldsymbol{y_i} , \boldsymbol{\hat{y}_i})$ stands for the empirical loss of the predicting model. It measures the prediction error between the model's output ($\hat{y}_i$, for under-bridge space samples) and the true binary label (y_i , 0=low elderly-care, 1=high elderly-care), using logistic loss to fit the binary classification task. $\sum_{k=1}^{K} \boldsymbol{\Omega}(\boldsymbol{f_k})$ stands for regularization loss.

When constructing the model, we encountered three other issues: (1) Class imbalance in labeled samples, (2) Noise sensitivity of spatial features due to diverse under-bridge layouts, and (3) Need for interpretable predictions to guide urban planning decisions. So we added four interconnected modules: (1) Feature preprocessing with outlier-resilient normalization; (2) Dual-stage data augmentation for class imbalance mitigation; (3) Bayesian optimized hyperparameter tuning; (4) Interpretability enhancement.

1. Feature preprocessing with outlier-resilient normalization. To mitigate the impact of outliers in spatial feature measurements, we employ RobustScaler for normalization to ensure stability. It is expressed as equation 7.

$$X_{ij}^{scaled} = \frac{X_{ij} - median(X_j)}{IQR(X_j)} \tag{7}$$

Where $\mathrm{IQR}(X_j) = Q_3\,(X_j) - Q1(X_j)$ (25th and 75th percentiles of feature j).

2. Dual-stage data augmentation for class imbalance mitigation. To address class imbalance without distorting feature semantics, we propose a dual-stage augmentation strategy, including SMOTE Oversampling and Gaussian Noise Augmentation.

3. Given the dataset imbalance and the high-stakes urban renewal context in Hong Kong, the model is optimized using a weighted multi-metric score via Optuna. This process can be expressed as equation 8.

$$\text{Score} = 0.45\text{Acc} + 0.30\text{F1} + 0.125\text{Prec} + 0.125\text{Rec} \tag{8}$$

Where equation 9

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{9}$$

is the correctness of the model. It assesses overall classification correctness, weighted most heavily. Because bridges, as basic city infrastructure, are very difficult to change or renew. A high accuracy ensures the model's baseline credibility and avoids resource misallocation. Equation 10

$$\text{F1} = 2 \times \frac{Prec \times Rec}{Prec + Rec} \tag{10}$$

Serves as the primary balance metric for our imbalanced dataset. F1 mitigates the accuracy bias toward the majority class, ensuring that the model neither over-misses high-potential regions nor over-labels low-quality regions. Equation 11

$$\text{Prec} = \frac{TP}{TP + FP} \tag{11}$$

Measures the reliability of "high elderly-friendliness" predictions. A high precision avoids unnecessary renovation costs, for misclassifying a low-friendly space would waste urban renewal budgets. Equation 12

$$\text{Rec} = \frac{TP}{TP + FN} \tag{12}$$

Evaluate the ability to identify all high-potential spaces. A high recall rate can prevent missed opportunities for optimization; failing to detect spaces with high potential will result in neglecting available space for the elderly in underutilized urban areas.

To further enhance model robustness and reduce generalization error, we construct a 9-model soft-voting ensemble based on the Bayesian-optimized parameter set.

4. To bridge the gap between model predictions and urban planning practice, we integrate complementary interpretability tool, Gain-Based Feature Importance, to quantify the contribution of each spatial indicator by cumulative loss reduction during tree splitting. This process can be expressed as equation 13.

$$Gain_j = \frac{\sum_{k=1}^{K} Gain_{jk}}{\sum_{l=1}^{m}\sum_{k=1}^{K} Gain_{lk}} \tag{13}$$

Where $Gain_{jk}$ represents the loss reduction from splitting feature j in tree k. The numerator $\sum_{k=1}^{K} Gain_{jk}$ accumulates the total loss reduction contributed by indicator j across all K decision trees in the XGBoost ensemble; the denominator $\sum_{l=1}^{m}\sum_{k=1}^{K} Gain_{lk}$ sums the total loss reductions of all m spatial indicators, ensuring $0 \ll Gain_j \ll 1$ and $\sum_{j=1}^{m} Gain_{jk} = 1$. Then we select the top 15 indicators with the largest normalized importance scores.

3 Proof of Concept

3.1 Experiment Environment

For the selection of the survey sites for the underpass space in Hong Kong, five research sites were selected, namely: Tsuen Wan-MTR Station Flyover, Sham Shui Po-Tung Chau Street Flyover, Tai Po-Tung Lo Wan Road Flyover, Causeway Bay-Canal Road East, To Kwa Wan- Flyover. Their locations are shown in Figure 4.

The Experiments were conducted on a high-performance computing server equipped with dual NVIDIA GeForce RTX 3090 GPUs (each with 24 GB GDDR6X memory, totaling 48 GB VRAM), an Intel Xeon Gold 6248R CPU @ 3.00 GHz (24 cores/48 threads), 256 GB DDR4 RAM, and 2 TB NVMe SSD storage. The operating model Ubuntu 22.04 LTS, with CUDA 12.1 and cuDNN 8.9 installed for GPU acceleration. All codes were implemented in Python 3.10.

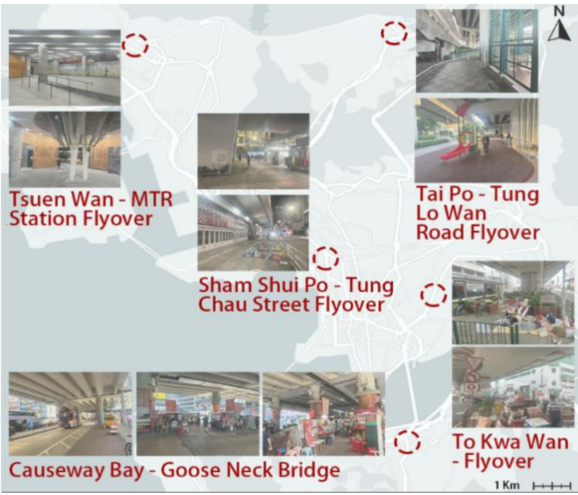

Figure 4. Map of the locations for on-site research.

3.2 Results of CC-VLM

As shown in Figure 5, high scoring samples are more likely from spaces with high green coverage rates, complete resting facilities, and moderate social vitality. These attributes align with the core needs of older adults emphasized in care theory and in on-site observations. In contrast, low-scoring samples are mostly concentrated in spaces characterized by inadequate safety facilities, high garbage coverage rates, and limited semi-enclosed spaces.

Figure 5. Multi-Elderly-Agent Scoring example.

3.3 Results Evaluation

To assess the performance of the classification model, we employ a set of metrics, including accuracy, recall, F1 Score, area under the receiver operating characteristic curve. These indicators are selected for their ability to provide a balanced evaluation. Besides, these metrics are consistent with the research objective, as they not only quantify predictive efficacy but also emphasizing sensitivity to adverse conditions. For instance, high recall ensures that the model reliably flags spaces requiring intervention, supporting equitable transformations of overlooked urban areas for aging populations. Baselines for these indicators are established relative to random guessing and prior benchmarks in analogous street-view-based urban analytics. For binary classifier, random performance yields accuracy, F1 and AUC-ROC are all of 0.5. When comparing with similar study, Hu et al. reported 79.9% accuracy and 75.51% mean average precision in urban physical disorder detection [9]. In our experiments, recall and AUC-ROC values exceeding 0.7 and 0.6 respectively are considered adequate. As shown in Table 3, overall performance metrics confirm balanced yet modest results across standard measures.

Table 3. Binary Classification Model Performance Metrics

Indicator Name	Score
AUC	0.615
Recall	0.714
Precision	0.536
F1-Score	0.612
Accuracy	0.655

As shown in Figure 6, the model demonstrated moderate discriminative ability, with an Area Under the ROC Curve (AUC) of 0.615. This AUC value, while above random chance, reflects the limited variability in the 302-image dataset, indicating potential for improvement with expanded data or refined features.

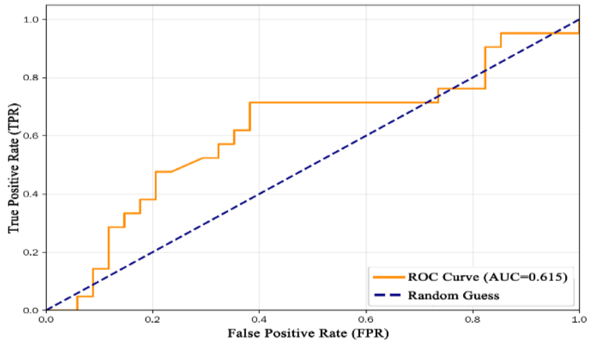

Figure 6. ROC Curve.

As shown in Figure 7, feature importance rankings offer interpretable insights directly tied to the care theory. Indicators from the Physical Shelter Layer predominated, led by green coverage rate and facility completeness, emphasizing greenery and basic infrastructure as primary drivers of elderly satisfaction. Psychological Belonging Layer features ranked prominently, supporting the influence of emotional comfort. Social Activity Layer metrics also contributed substantially.

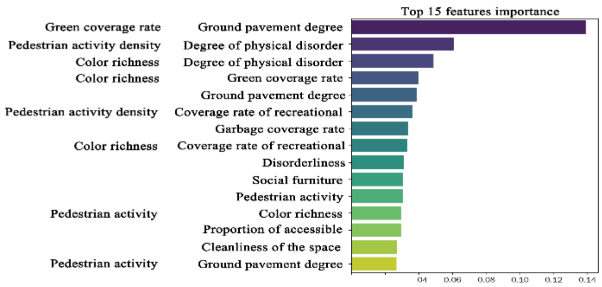

Figure 7. The importance of the top 15 features.

4 Conclusion

This study addresses the critical challenge of efficiently assessing urban under-bridge spaces. By proposing the CC-VLM framework, we achieve three core objectives: (1) a theory-grounded hierarchical metric model synthesizing care ethics into practice; (2) an automated workflow that simulates diverse elderly perspectives to generate subjective visual labels; and (3) a multi-model-enhanced VLM integrated with XGBoost for scene assessment and prediction. Experimental results validate the framework's effectiveness, confirming that greenery, facility completeness, and activity density are dominant drivers.

Limitations include modest predictive performance due to the small dataset size, overreliance on visual imagery alone, and reliance on simulated agents. In particular, the XGBoost classifier is trained using scores generated by the VLM agents as target labels, therefore it is not directly reflecting real-world elderly experiences.

Future work could expand to larger datasets, integrate more data from real older adults, and incorporate multi-modal environmental sensors for evaluations.

5 Acknowledgement

This research is supported by the Hong Kong PolyU (UGC) Start-up Fund for New Recruits (P0056204) and Zhejiang Provincial "Pioneer" and "Leading Goose" Science and Technology Plan Project (2026C04040)

Reference

[1] Census and Statistics Department The Government of the Hong Kong Special Administrative Region. "Table 110-01001 : Population by Sex and Age Group."https://www.censtatd.gov.hk/en/web_table.html?id=110-01001

[2] Development Bureau of Hong Kong. "Hong Kong 2030+."https://www.pland.gov.hk/file/planning_studies/comp_s/hk2030plus/TC/index.htm

[3] Planning Department of Hong Kong. "Hong Kong Planning Standards and Guidelines." https://www.pland.gov.hk/file/tech_doc/hkpsg/full/pdf/ch4_en.pdf

[4] R. ROHINIKUMAR, "Design Guidelines for urban Public Spaces for the Elderly," Tesis de maestr á en Arquitectura. Uttarakhand: Indian Institute of …, 2017.

[5] Kowloon East Office. "Promoting the "Spirit of Creation" Fly the Flyover Operation." https://www.ekeo.gov.hk/en/the-spirit-of-creation/innovative-placemaking-urban-design/fly-the-flyover-operation/index.html

[6] LEGISLATIVE COUNCIL OF HONG KONG. "OFFICIAL RECORD OF PROCEEDINGS." https://www.legco.gov.hk/yr2025/chinese/counmtg/floor/cm20250521-confirm-ec.pdf#nameddest=wrq11

[7] V. Mehta, "Evaluating Public Space," *Journal of Urban Design*, vol. 19, no. 1, pp. 53–88, 2013, doi: 10.1080/13574809.2013.854698.

[8] S. Tawfik Mohamed, A. Mandour, and H. Baker, "A critical review of quality assessment tools for public spaces," *Engineering Research Journal*, vol. 177, pp. 255–274, 2023.

[9] C. Hu *et al.*, "UPDExplainer: An interpretable transformer-based framework for urban physical disorder detection using street view imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 204, pp. 209–222, 2023.

[10] Z. Fan, F. Zhang, B. P. Loo, and C. Ratti, "Urban visual intelligence: Uncovering hidden city profiles with street view images," *Proceedings of the National Academy of Sciences*, vol. 120, no. 27, p. e2220417120, 2023.

[11] A. Ghosh, A. Acharya, S. Saha, V. Jain, and A. Chadha, "Exploring the frontier of vision-language models: A survey of current methodologies and future directions," *arXiv preprint arXiv:2404.07214*, 2024.

[12] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.

[13] Z. Peng *et al.*, "Vision language model (VLM)-enabled street view analytics: a systematic literature review," *Engineering, Construction and Architectural Management*, pp. 1–19, 2025.

[14] J. Ro, N. Kim, and Y. Yoon, "How Well Do Vision-Language Models Understand Cities? A Comparative Study on Spatial Reasoning from Street-View Images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 6476–6485.

[15] Y. Yao, D. Ò. Giuliano, and F. Lu, "Urban Street-Scene Perception and Renewal Strategies Powered by Vision–Language Models," 2025.

[16] A. Chatzidakis, J. Hakim, J. Litter, C. Rottenberg, and C. Collective, *The care manifesto: The politics of interdependence*. Verso Books, 2020.

[17] A. Binet *et al.*, "The urban infrastructure of care: Planning for equitable social reproduction," *Journal of the American Planning Association*, vol. 89, no. 3, pp. 282–294, 2023.

[18] C. Nelischer, A. Loukaitou-Sideris, and G. Wendel, "Caring public space: Advancing justice through intergenerational public space design and planning," *Journal of Urban Affairs*, vol. 47, no. 8, pp. 2789–2808, 2025.

[19] A. Gabauer *et al.*, *Care and the city: encounters with urban studies*. Taylor & Francis, 2022.

[20] S. Zhang, X. Yu, and J. Ma, "Vision AI reveals waste-safety perception relationships in urban environments," *npj Urban Sustainability*, vol. 5, no. 1, p. 83, 2025.

[21] C.-F. Chan *et al.*, "Context-aware vision-language model agent enriched with domain-specific ontology for construction site safety monitoring," *Automation in Construction*, vol. 177, p. 106305, 2025.

[22] A. Li, P. K.-Y. Wong, X. Tao, J. Ma, and J. C. Cheng, "An interactive system for 3D spatial relationship query by integrating tree-based element indexing and LLM-based agent," *Advanced Engineering Informatics*, vol. 66, p. 103375, 2025.

[23] Ruan Y, Zhang X, Wang S, Chen X, Chen Q. Street View-Enabled Explainable Machine Learning for Spatial Optimization of Non-Motorized Transportation-Oriented Urban Design. Land. 2025.