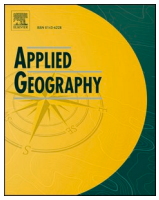
Applied Geography 198 (2027) 104230



Contents lists available at ScienceDirect

Applied Geography

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



Interpretable vision-language reasoning for diagnosing estate-scale residential renewal urgency

Xiaoqin Yan [a,b,c], Leyi Su [a,b,c], Mengfan Tang [a,b,c], Shuliang Ren [d,e], Yitian Liu [a,b,c], Junnan Qi [a,b,c], Jiangpeng Zheng [a,b,c], Yi Bao [f], Zhou Huang [a,b,c,*]

[a] Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing, China
[b] Beijing Key Laboratory of Spatio-temporal Perception and Urban Resilience, Peking University, Beijing, China
[c] Engineering Research Center of Earth Observation and Navigation, Ministry of Education, Beijing, China
[d] The College of Geoscience and Surveying Engineering, China University of Mining and Technology-Beijing, Beijing, China
[e] Beijing Key Laboratory of Urban Spatial Intelligent Sensing and Digital Governance, Beijing, China
[f] Institute of Space Earth Science, Nanjing University, Suzhou, China

ARTICLE INFO

Keywords:
Urban renewal urgency
Vision-language model
Chain-of-thought reasoning
Housing estate imagery

ABSTRACT

Assessing the urban renewal urgency enables planners to prioritize interventions and allocate resources effectively. Current visual-based evaluation methods rely primarily on pattern recognition and fail to provide explicit diagnostic reasoning about observed conditions. Vision-language models (VLMs) offer a promising approach by incorporating intrinsic semantic knowledge and enabling interpretable visual reasoning. Therefore, this study combines a VLM with chain-of-thought (CoT) reasoning to assess renewal urgency from observable physical environmental conditions at the housing estate level. Using Beijing as a case study, we develop multidimensional assessment criteria and fine-tune Qwen3-VL-8B through low-rank adaptation (LoRA) with CoT supervision. The adapted model improves performance across most assessment dimensions. For overall renewal urgency, it achieves an F1-score of 0.760 and a quadratic weighted kappa (QWK) of 0.785. These represent relative improvements of 13.1% and 6.5% over the baseline without CoT. The trained model classifies 2947 housing estates (56.6%) as having moderate overall renewal urgency. Renewal urgency decreases from the city center toward outer ring road zones, reflecting the concentration of aging infrastructure in older inner-city areas. Furthermore, the model provides human-readable reasoning outputs that articulate dimension-specific visual evidence, improving transparency and practical utility. The framework provides a scalable and evidence-based approach to interpretable image-based assessment of urban renewal urgency.

1. Introduction

Global urban development has shifted from rapid expansion to the renewal and improvement of existing urban areas (Elmqvist et al., 2021; Xia et al., 2020). Urban renewal involves maintaining, reconstructing, and systematically upgrading the built environments (Roberts, 2000; Zheng et al., 2014). It is a vital strategy for improving spatial quality and human livability (Wu et al., 2020). In China, a large proportion of housing estates were built before 2000, and many now face pressing issues of physical deterioration, inadequate infrastructure, and functional obsolescence (He et al., 2025). In this context, comprehensive and cost-effective assessments of renewal urgency are essential for informing resource allocation and prioritizing interventions.

Traditional approaches to assessing urban renewal urgency rely on field surveys and administrative records (Han et al., 2021; Liu et al., 2018; Zhuang et al., 2019), which are labor-intensive or provide limited support for fine-grained analysis. The advancement of spatio-temporal big data and artificial intelligence enables more efficient, data-driven assessment (Wang et al., 2020; Zhao, Wang, Ouyang, et al., 2024). By integrating multi-source data and advanced models, effective

* Corresponding author. Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing, China.
E-mail addresses: yanxiaoqin@stu.pku.edu.cn (X. Yan), lysu25@stu.pku.edu.cn (L. Su), mftang@stu.pku.edu.cn (M. Tang), renshuliang@cumtb.edu.cn (S. Ren), liuyitian@stu.pku.edu.cn (Y. Liu), qijunnan@stu.pku.edu.cn (J. Qi), zhengjp@stu.pku.edu.cn (J. Zheng), baoyi@nju.edu.cn (Y. Bao), huangzhou@pku.edu.cn (Z. Huang).

https://doi.org/10.1016/j.apgeog.2026.104230
Received 16 June 2026; Received in revised form 16 October 2026; Accepted 16 October 2026
0143-6228/© 2026 Published by Elsevier Ltd.

analysis of urban functions and human activities can be conducted to assess urban renewal urgency. For example, tree-based models predict urban renewal urgency using building condition, population density, and socioeconomic data (Sun et al., 2024; Zhao et al., 2025). Moreover, Yan et al. (2026) assess urban renewal potential by constructing hierarchical graph networks to capture complex building and spatial relationships. Although these approaches effectively characterize structural features and socioeconomic patterns, they fail to incorporate visual information within the physical environment. This limits their ability to assess building facade conditions, environmental maintenance, and spatial aesthetics that also influence urban renewal urgency (Ma et al., 2021).

Urban imagery, such as street view imagery (SVI), provides direct visual evidence for assessing built environment conditions (Fan et al., 2023; Zhang, Salazar-Miranda, et al., 2024). These visual datasets are widely used in environmental assessment (Lu & Chen, 2024), sentiment analysis (Yang & Zhang, 2026), and building instance classification (Kang et al., 2018). For example, Liu et al. (2017) assess urban environmental quality by measuring building facade and street wall quality from SVI. He et al. (2023) further demonstrate the effectiveness of SVI for the assessment of urban renewal by combining emotional and visual features extracted from the imagery. However, SVI mainly captures street-facing facades, leaving internal housing estate spaces where residents carry out daily activities largely unobserved (Lu et al., 2020). This constrains observations of internal leisure spaces, parking conditions, and courtyard environments that affect residents' quality of life. In contrast, crowdsourced imagery from real estate platforms such as Anjuke captures these internal spaces and provides an alternative source of visual evidence for housing estate assessment (Gao et al., 2024).

With advances in deep learning, convolutional neural networks (CNNs) have become the primary method for extracting image features (Zhao, Wang, Zhang, et al., 2024). Through convolutional operations and hierarchical feature extraction, CNNs effectively identify visual elements such as buildings, vegetation, and road surfaces from urban imagery (Huang et al., 2023; Yao et al., 2019). Despite their strong performance in object detection, these models rely mainly on pattern recognition and lack contextual understanding. For instance, CNNs can reliably identify vehicles and paved surfaces but struggle to distinguish organized parking from unauthorized encroachment. Furthermore, these black-box models also fail to meet the interpretability requirements of urban planning because they produce predictions without explaining the underlying reasoning (Xing & Sieber, 2023). Attention-based methods can highlight specific image regions that influence model decisions (Zhao, Lu, & Lin, 2024), but provide limited explanations of what those regions imply for planning and policy. As a complex diagnostic task, renewal urgency assessment requires an analytical tool that can classify visual conditions and conduct explicit reasoning based on spatial evidence (Nachmany & Hananel, 2023).

Vision-language models (VLMs) offer a promising solution by coupling visual perception with semantic reasoning (Alayrac et al., 2022; Feng et al., 2025). Pretraining on large-scale image-text datasets equips these models with semantic knowledge and reduces reliance on extensive manual knowledge engineering (Wang, Xuan, et al., 2026). Recent studies have applied VLMs to urban studies such as neighborhood sustainability assessment (Howell et al., 2026), building facade assessment (Raghu et al., 2026) and urban perception (Perez & Fusco, 2025; Zheng et al., 2026). These models can further generate reasoning outputs that link visual evidence to assessment results, helping address the interpretability gap in automated urban analysis (Chen et al., 2024). For example, multimodal large language models (e.g. GPT) can score perceived urban safety from SVI with results that align with human judgments (Zhang et al., 2025). These studies show that VLMs can capture the semantic context of urban scenes, supporting explainable urban assessment. However, general-purpose VLMs are trained across diverse domains and may not fully meet the requirements of specific tasks (Zhang, Huang, et al., 2024). Domain adaptation informed by domain knowledge is therefore necessary for assessing urban renewal urgency.

Domain adaptation improves task performance by aligning general VLM capabilities with domain knowledge and task requirements. Existing VLMs have been adapted through domain-specific datasets, improving their ability to perform urban socioeconomic assessment and geospatial reasoning (Cai et al., 2026; Huang et al., 2026; Wang, Xuan, et al., 2026). For example, PlanGPT-VL integrates planning knowledge with structured verification to interpret planning maps (Zhu et al., 2025). Bai et al. (2026) further develop a VLM guided by expert knowledge to identify and interpret micro-renewal interventions from paired historical SVI, demonstrating the feasibility of using adapted VLMs to analyze urban renewal. However, the study focused on the retrospective analysis of historical changes while overlooking current renewal requirements. Further research is needed on how VLMs can provide interpretable assessments of renewal urgency from current urban conditions. These evaluations are crucial for identifying priority areas and informing future interventions (Acevedo-De-los-Rios et al., 2025).

To address this gap, this study develops an interpretable VLM-based framework for assessing urban renewal urgency at the housing estate level. Assessment criteria are defined for five dimensions, including building condition, nature environment, public space, spatial order, and sanitation. These criteria guide the construction of a dataset with chain-of-thought (CoT) annotations linking visual observations, reasoning, and urgency scores. Qwen3-VL-8B is fine-tuned on this dataset using low-rank adaptation (LoRA), followed by model performance comparison. Predictions from imagery are then aggregated for each housing estate to map renewal urgency in Beijing. Overall, this study makes two main contributions:

- We develop a vision-language framework that combines CoT reasoning with LoRA adaptation, improving model performance in the multidimensional assessment of urban renewal urgency.
- The framework generates interpretable reasoning that links visual evidence to renewal urgency judgments, making the assessment rationale explicit and providing a basis for targeted urban renewal interventions.

The remainder of this paper is organized as follows. Section 2 describes the study area, data sources, and methodology. Section 3 presents the model evaluation results, analysis of reasoning outputs, and spatial patterns of renewal urgency. Section 4 discusses the findings and limitations. Section 5 summarizes the main conclusions.

2. Data and methodology

The framework shown in Fig. 1 combines housing estate imagery with CoT reasoning and LoRA-based VLM adaptation for interpretable renewal urgency assessment at the housing estate level. It comprises three main components: (1) data preparation, including imagery preprocessing and instruction dataset construction; (2) LoRA-based VLM adaptation for renewal urgency assessment at the housing estate level, with structured reasoning outputs; and (3) model evaluation and application, including model comparison, reasoning analysis, and spatial mapping of renewal urgency.

2.1. Data preparation and preprocessing

The study area covers Beijing's 5th ring road and its surrounding areas, which integrate diverse urban textures and exemplify the urban renewal challenges faced by megacities (Ren et al., 2025). To capture detailed conditions within housing estates, this study uses imagery from Anjuke (https://www.anjuke.com), a major real estate platform in China. Most housing estates in China are gated, preventing SVI from capturing internal courtyards, public facilities, and shared spaces. In

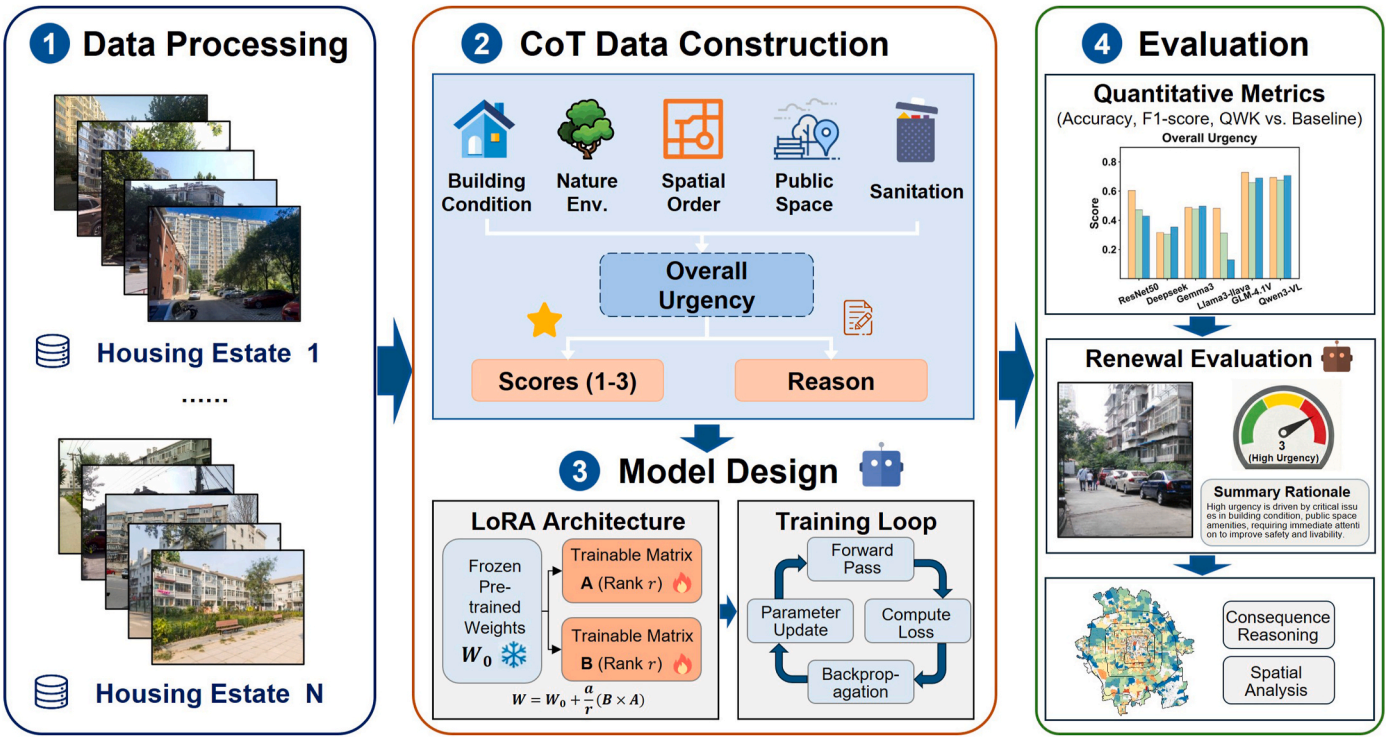


Fig. 1. Overall framework of the proposed study includes housing estate imagery preprocessing, chain-of-thought-based instruction data construction, vision-language model adaptation, and urban renewal urgency evaluation.

contrast, crowdsourced images on Anjuke document these internal environments, including courtyards, green spaces, internal roads, and

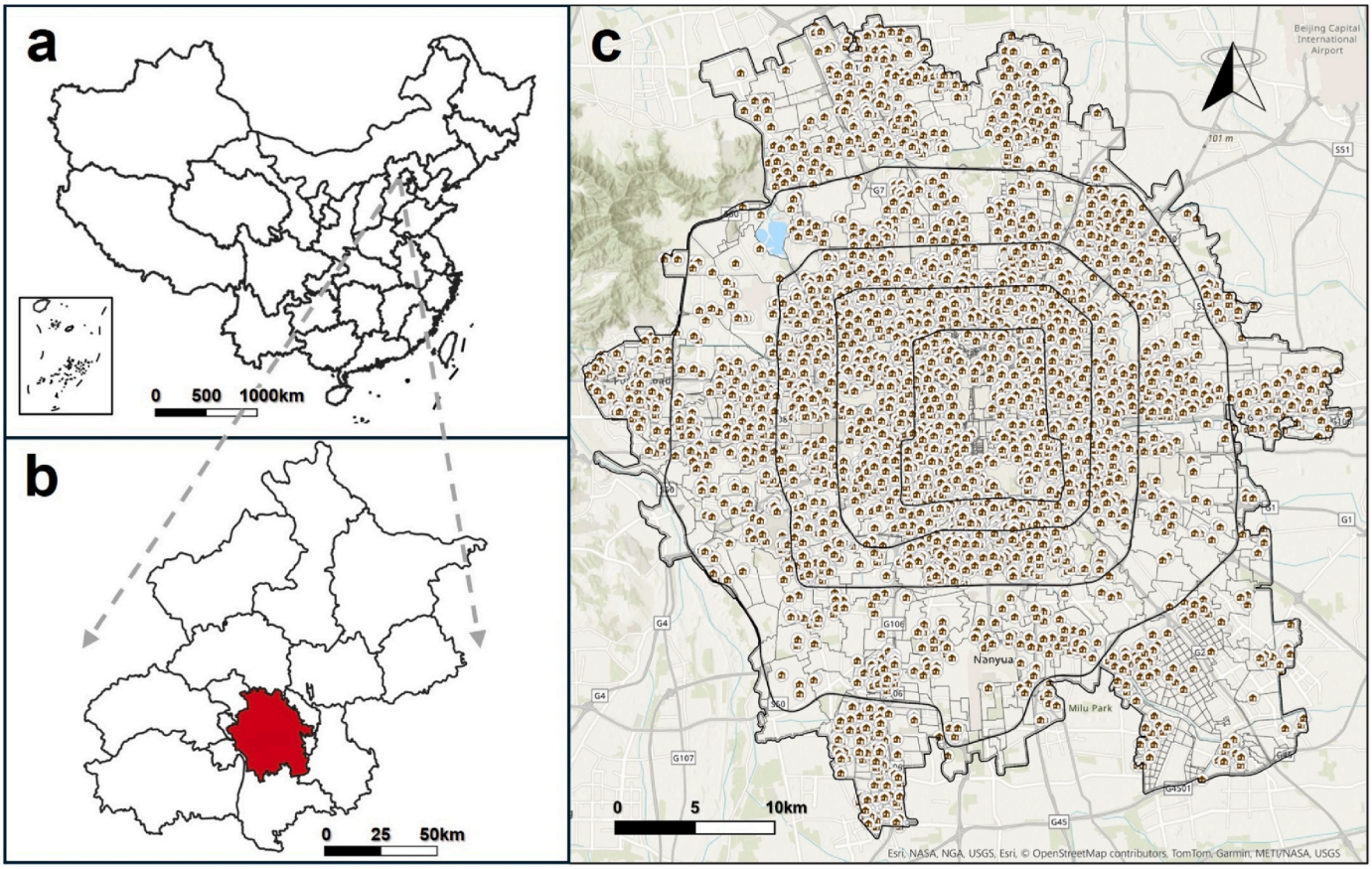


Fig. 2. Location of Beijing in China (a), the study area within Beijing (b), and the distribution of sampled housing estates (c).

building facades. Previous studies demonstrate the effectiveness of this data source for assessing enclosed residential environments in China (Gao et al., 2024; Hong et al., 2026). These images capture conditions relevant to residents' daily lives and provide visual evidence for diagnosing housing estate renewal needs.

The original imagery of Anjuke varies in size and viewpoint, requiring manual screening to select images containing sufficient information for evaluation. Images with severely oblique or obstructed views or abnormal dimensions are excluded, along with interior decoration renderings and single-object close-ups. Five representative images are retained for each estate to balance image counts across estates while minimizing visual redundancy and computational costs (Huang et al., 2023; Li et al., 2015). Images from different positions collectively capture building facades, greenery, public spaces, internal circulation routes, and sanitation-related conditions, reducing bias from any single viewpoint. The images range from 244 to 1024 pixels in width and from 213 to 640 pixels in height. The most common size is 800 × 600 pixels (26.2%), as shown in Fig. S2 of the Supplementary Information. The VLM's dynamic-resolution processor accommodates these variations without requiring a fixed input size. The final dataset contains 26,030 images from 5206 housing estates (Fig. 2).

2.2. Dataset construction with CoT annotations

2.2.1. Dimensions for assessing urban renewal urgency

Urban renewal is a comprehensive process addressing interrelated physical, economic, social, and environmental conditions in existing urban areas (Roberts, 2000). At the housing estate level, evaluating the renewal needs of the physical environment is crucial as it directly affects residents' daily lives. Therefore, this study defines renewal urgency as the extent to which intervention is needed to maintain and improve the visible physical and environmental conditions of housing estates (Gao et al., 2025; Zheng et al., 2014). Specifically, the assessment covers five dimensions, including building condition, nature environment, public space, sanitation, and spatial order. These dimensions can be observed through imagery and reflect the visible physical deterioration, environmental quality, and spatial management conditions relevant to housing estate renewal.

Table 1 summarizes the five assessment dimensions. Building condition describes the visible physical state and maintenance of residential buildings. Nature environment concerns the quality and management of greenery and landscape elements. Public space reflects the condition and functionality of shared outdoor areas. Sanitation captures visible cleanliness and routine environmental maintenance. Spatial order describes the visual and physical organization of the housing estate environment. These dimensions collectively provide a structured description of major observable housing estate conditions that are closely related to renewal diagnosis.

Each dimension is scored on a three-level ordinal scale, with 1 indicating well-maintained conditions and low urgency, 2 indicating moderate degradation, and 3 indicating severe deterioration and high urgency. Table 1 describes the observable conditions associated with the low and high urgency categories. For building condition and nature environment, structurally intact facades and managed greenery correspond to score 1, whereas visible structural damage and bare soil correspond to score 3. Similarly, public space, sanitation, and spatial order distinguish clean, pedestrian-friendly, and visually coherent environments from those with encroached pathways, overflowing waste, or chaotic overhead wiring (score 3). The overall urgency score is calculated by rounding the mean of the five dimension scores, providing a quantitative basis for targeted renewal interventions.

2.2.2. Instruction dataset construction

An instruction dataset is constructed to support multidimensional assessment of housing estate renewal urgency. Each sample follows an image-question-answer format and consists of a housing estate image, a task instruction, and a structured answer. The answer includes dimension-specific judgments, an overall urgency label, and supporting reasoning, enabling the model to learn both the assessment outcomes and the evaluation process.

The structured answers incorporate CoT reasoning, which makes intermediate judgments explicit through step-by-step reasoning before reaching a conclusion (Wei et al., 2022). As shown in Fig. 3, CoT organizes the assessment process across the five dimensions of building condition, nature environment, public space, sanitation, and spatial order. For each image, the annotation identifies relevant visual evidence, assigns a score to each dimension, and then provides an overall rationale and urgency score. All annotations follow a unified template to improve consistency and support VLM instruction tuning. The complete prompt template and an example of the model output are provided in Note S1 and Fig. S1. This standardized format helps the model learn multistep reasoning patterns and improves the interpretability of model outputs. Unlike conventional image classification, this design explicitly represents the reasoning that connects visual observations to renewal urgency judgments.

To ensure the quality of the annotation and mitigate subjective bias, the annotation team comprises ten researchers with domain expertise in urban planning and the built environment. Following a consensus-based approach (Wang, Ye, & Weng, 2026), the team first conducts a pilot annotation exercise on a random subset of images to align interpretations of the scoring criteria (Section 2.2.1) and refines the CoT template. During formal annotation, each image is independently annotated by one researcher and verified by another, with any disagreement adjudicated by a third according to the predefined criteria. The resulting instruction dataset contains 1200 samples covering diverse locations and varying dimensional conditions. Details of the annotations are provided in Fig. S3.

2.3. CoT-guided VLM adaptation for renewal urgency assessment

Based on the instruction dataset, this study defines the urban renewal urgency assessment as a structured visual reasoning task with an "observation-reasoning-scoring" structure. Given an image I_i and a task prompt p_i, the model generates a structured response containing

Table 1
Assessment dimensions and scoring rubric for housing estate renewal urgency.

Dimension	Assessment focus	Score 1: (Low urgency, good condition)	Score 3: (High urgency, severe deterioration)
Building Condition	Visible physical condition and maintenance of residential buildings	Buildings with no visible exterior damage, clean facades, and intact windows	Facade damage or exterior deterioration and makeshift additions
Nature Environment	Quality and management of greenery and landscape elements	Well-maintained greenery, landscaped and manicured areas	Bare soil, overgrown weeds, dead plants, or complete lack of greenery
Public Space	Usability of shared outdoor areas	Defined social areas, paving in good condition, and pedestrian-friendly layouts	Encroached pathways, broken pavement, and disorganized parking that obstructs movement
Sanitation	Visible cleanliness and routine environmental maintenance	Clean streets, organized bins and adequate routine maintenance	Visible litter piles, stained ground, overflowing waste, and signs of sewage
Spatial Order	Visual and physical organization of the housing estate environment	Orderly skyline, organized cables, and clear visual boundaries	Severe visual disorder, chaotic overhead wiring, disorderly signage, and clutter

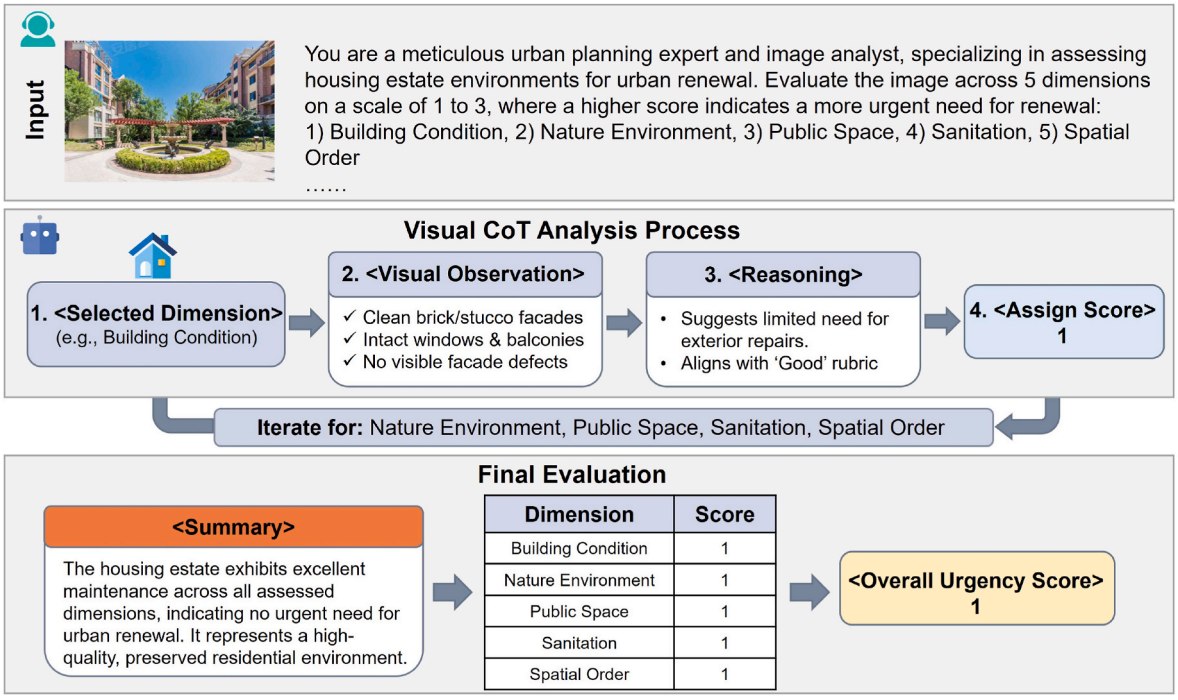


Fig. 3. Structure of the CoT prompt for dimension-specific observation, reasoning, and scoring, followed by overall urgency assessment.

criterion-based observations, reasoning, five dimension-level scores, and an overall urgency score. Let $s_{i,d} \in \{1,2,3\}$ denote the score for dimension d of image i. The prompt specifies that the overall urgency score U_i is calculated as the rounded mean of the five dimension scores:

$$U_i = round\left(\frac{1}{5}\sum_{d=1}^{5} s_{i,d}\right) \tag{1}$$

In Eq. (1), each dimension score of 1 to 3 represents low to high renewal urgency. The five dimensions are equally weighted to comprehensively reflect the renewal urgency. The averaged score is then rounded to assign it to the nearest overall urgency level. The sensitivity to alternative thresholds and weights is reported in Fig. S4.

Qwen3-VL-8B is adopted as the base vision-language model in this study. The model integrates visual and textual representations through a Transformer-based multimodal architecture and performs well on general visual understanding benchmarks, including MMBench and RealWorldQA (Bai et al., 2025). VLMs can process images of varying resolutions (Guo et al., 2024). Images are divided into local patches and mapped to visual tokens for integration with textual representations. Dynamic-resolution processing accommodates different image sizes and aspect ratios by varying the number of visual tokens. This mechanism allows the model to handle variations in the resolution and composition of crowdsourced imagery. The input image is encoded into visual tokens $v \in R^{N_v \times d}$, and the prompt is embedded as $x \in R^{N_x \times d}$, where N_v and N_x denote the number of visual and text tokens, respectively, and d is the token embedding dimension. These multimodal tokens are jointly processed by the language model to produce hidden representations h:

$$h = LLM([v;x]) \tag{2}$$

These hidden representations support the generation of dimension-specific judgments and reasoning from housing estate imagery. The response y is then generated autoregressively over T steps as:

$$P(y|I,p) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, I, p) \tag{3}$$

Each token y_t is conditioned on all previously generated tokens $y_{<t}$, the input image I, and the prompt p. The response y includes urgency assessments across the five dimensions and the corresponding reasoning. Rather than supervising only the score labels, the complete observation-reasoning-scoring response defined by the CoT template is used as the training target. This sequence-level supervision enables the model to learn how observable environmental conditions relate to the assessment criteria and their urgency levels.

The model is fine-tuned using LoRA, which introduces trainable low-rank matrices into selected layers while keeping the pretrained weights frozen. For a pretrained weight matrix $W_0 \in \mathrm{R}^{m\times n}$, the task-specific update and adapted weight matrix are defined as follows:

$$\Delta W = (\alpha/r)BA \tag{4}$$

$$W = W_0 + (\alpha/r)BA \tag{5}$$

Here, $A \in \mathrm{R}^{r\times n}$ and $B \in \mathrm{R}^{m\times r}$ are trainable matrices with rank $r \ll \min(m,n)$, and α is a scaling hyperparameter that controls the magnitude of the low-rank update. LoRA is applied to selected attention and feed-forward projection layers within the Transformer blocks. This parameter-efficient approach reduces training costs and preserves the base model's general multimodal capabilities. Training minimizes the standard autoregressive language modeling loss:

$$\mathscr{L}_{LM} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} \log \mathrm{P}\left(y_{i,t} \middle| y_{i,<t}, I_i, p_i; \theta\right) \tag{6}$$

Here, N denotes the number of samples in the training batch and T_i denotes the number of tokens in the target response for sample i. The trainable parameters θ consist solely of the LoRA matrices introduced in Eq. (4) and Eq. (5), while all other model weights remain fixed. Through LoRA adaptation, the sequence-level supervision aligns the general-purpose VLM with both the five-dimensional assessment rubric and the required output structure. The adapted model therefore produces machine-readable dimension scores together with human-readable diagnostic reasoning, rather than a single classification result. During inference, the model generates a structured response from the input image and prompt. The dimension-specific and overall urgency scores are extracted from the response and verified. For each housing estate,

predictions from its five retained images are averaged and rounded to obtain the corresponding estate-level scores.

2.4. Model configuration and evaluation

Model performance is evaluated using accuracy, macro F1-score, and quadratic weighted kappa (QWK). Accuracy measures the proportion of correctly predicted samples:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i = y_i) \tag{7}$$

where N is the number of samples, y_i and $\hat{y}_i$ are the ground-truth and predicted labels, respectively, and $1(\cdot)$ is the indicator function. Macro F1-score gives equal weight to each class, making it suitable for evaluating performance when class frequencies are imbalanced. For each class, precision and recall are defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{8}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{9}$$

Here, TP_c, FP_c and FN_c denote the numbers of true positives, false positives, and false negatives for class c, respectively. The macro F1-score is then calculated as:

$$Macro\ F1 - score = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c} \tag{10}$$

Quadratic weighted kappa (QWK) measures agreement between predicted and ground-truth labels while accounting for the ordinal distance between categories. It penalizes disagreements between more distant urgency levels more heavily, making it appropriate for the ordinal assessment in this study. QWK is defined as:

$$k = 1 - \frac{\sum_{i=1}^{C}\sum_{j=1}^{C} w_{ij}O_{ij}}{\sum_{i=1}^{C}\sum_{j=1}^{C} w_{ij}E_{ij}} \tag{11}$$

In Eq. (11), O_{ij} and E_{ij} are the entries of the observed and expected confusion matrices for true class i and predicted class j, respectively, and w_{ij} is the quadratic disagreement weight. Higher QWK values indicate closer agreement between predictions and ground-truth labels after accounting for their ordinal distance.

The proposed framework is compared with baselines representing conventional visual learning and multimodal foundation models. ResNet50 serves as the convolutional neural network (CNN) baseline, as its residual architecture is widely used for image classification (He et al., 2016). The VLM baselines include DeepSeek-VL-7B, Gemma-3-12B, Llama3-LLaVA-8B, and GLM-4.1V-9B. DeepSeek-VL uses a hybrid visual encoder to capture both semantic information and fine visual details (Lu et al., 2024). Llama3-LLaVA combines a CLIP-based visual encoder with the Llama3-8B language model within the LLaVA framework (Liu et al., 2024). Gemma-3 and GLM-4.1V represent more recent open-weight multimodal model families with visual instruction-following and reasoning capabilities (Gemma Team, 2025; GLM-V Team, 2025). These VLMs, including Qwen3-VL-8B, encompass diverse visual encoders, multimodal alignment strategies, and model series within a parameter range of 7 to 12 billion. The comparison examines the conventional image classification, zero-shot vision-language inference, and task-specific VLM adaptation.

The annotated dataset is divided at the housing estate level into training and test sets containing 80% and 20% of the images, respectively. All images from the same housing estate are assigned to one of the datasets to prevent potential data leakage. All experiments are conducted on four NVIDIA RTX 4090 GPUs (each with 24 GB of VRAM), running Ubuntu 22.04 with Python 3.9, PyTorch 2.6, and CUDA 12.8. LoRA is applied to the main attention and feed-forward projection layers, with a rank of 128, $\alpha = 64$ and a dropout rate of 0.05. The trainable parameters are optimized with AdamW using a learning rate of 1×10^{-4} and a warmup ratio of 0.1. Training runs for 8 epochs with an effective batch size of 8 and a maximum sequence length of 8192 tokens.

3. Results

3.1. Model performance comparison

Fig. 4 compares the zero-shot performance of general-purpose VLMs with the performance of pretrained ResNet50. Qwen3-VL and GLM-4.1V generally outperform ResNet50 across the assessment dimensions, while the performance of the other VLMs varies by dimension. For overall urgency, ResNet50 achieves an F1-score of 0.471 and a QWK of 0.428. In contrast, Qwen3-VL achieves corresponding scores of 0.675 and 0.706, while GLM-4.1V reaches 0.657 and 0.689. These results indicate that general-purpose VLMs can outperform the traditional classifiers even in zero-shot settings, providing a foundation for subsequent task-specific adaptation.

These performance advantages also extend across various dimensions. Qwen3-VL and GLM-4.1V generally achieve the strongest performance across the five dimensions, with their best scores in each dimension averaging 0.729 for accuracy, 0.632 for F1-score, and 0.683 for QWK, compared with 0.608, 0.466, and 0.355 for ResNet50, respectively. Specifically, Qwen3-VL shows strength when renewal judgments require contextual interpretation of visual conditions. For example, distinguishing well-maintained greenery from unmanaged overgrowth depends on spatial context rather than pixel-level features alone. In the nature environment dimension, Qwen3-VL achieves a QWK of 0.747, compared with 0.488 for ResNet50. Disparity also appears in the sanitation dimension, where QWK increases from 0.090 for ResNet50 to 0.599 for Qwen3-VL. The results indicate that VLMs can capture scene semantics more effectively than traditional models, supporting more reliable urgency predictions.

3.2. LoRA adaptation and the role of chain-of-thought reasoning

Qwen3-VL-8B is further adapted using LoRA with CoT supervision. For overall renewal urgency, the adapted model achieves an accuracy of 0.792 and a QWK of 0.785 (Table 2), compared with 0.693 and 0.706, respectively, under zero-shot inference. Performance also improves across most dimension-specific metrics. For example, building condition accuracy increases from 0.738 to 0.775, while QWK for nature environment rises from 0.747 to 0.821. These results indicate that LoRA adaptation with CoT supervision helps align the model's general semantic understanding with urban renewal urgency assessment.

The contribution of CoT supervision is evaluated by comparing the two LoRA-adapted variants. Across the assessment dimensions and overall urgency, the CoT-enhanced model attains higher values for 14 of the 18 reported metrics. For overall urgency, accuracy, F1-score, and QWK increase from 0.700, 0.672, and 0.737 to 0.792, 0.760, and 0.785, representing relative improvements of 13.1%, 13.1%, and 6.5%, respectively. Most performance metrics also improve across the five dimensions. For example, F1-score of nature environment rises from 0.737 to 0.773 and QWK of public space increases from 0.745 to 0.786. These results suggest that CoT supervision facilitates the integration of visual evidence to support more reliable renewal urgency assessment.

Fig. 5 presents the confusion matrices for Qwen3-VL-8B adapted with LoRA and CoT supervision. The model generally recognizes low and high urgency cases more accurately than moderate urgency cases. Recall for low urgency is around 0.80 or higher across the assessment dimensions. Across the five assessment dimensions and overall urgency, recall averages approximately 0.76 for high urgency and 0.58 for moderate urgency. Recall for high urgency reaches 0.90 in spatial order.

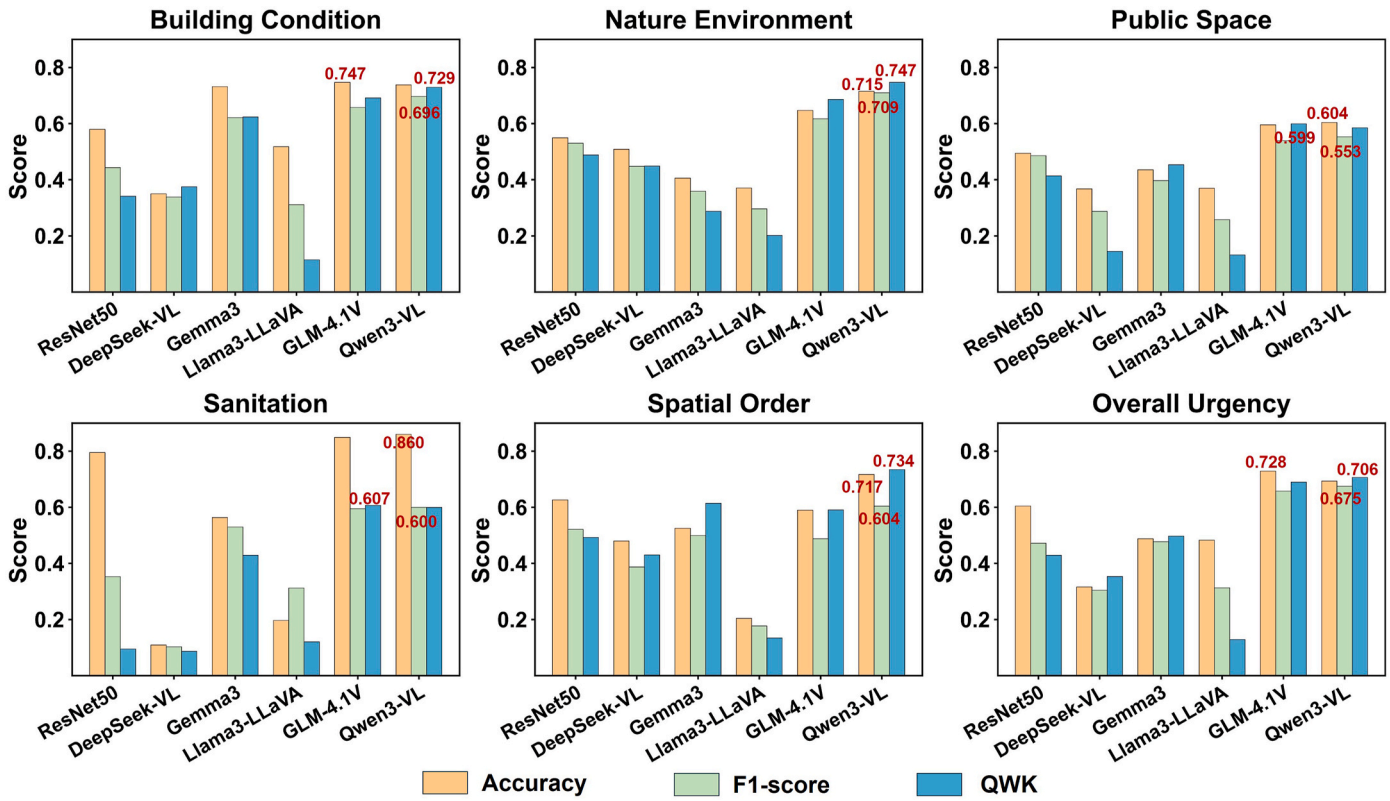


Fig. 4. Performance comparison of ResNet50 and zero-shot VLM baselines across the five assessment dimensions and overall urgency. Red values indicate the highest score for each metric in each panel. For brevity, DeepSeek-VL-7B, Gemma-3-12B, Llama3-LLaVA-8B, GLM-4.1V-9B, and Qwen3-VL-8B are abbreviated as DeepSeek-VL, Gemma3, Llama3-LLaVA, GLM-4.1V, and Qwen3-VL, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Performance of LoRA-adapted Qwen3-VL-8B with and without CoT supervision. Bold values indicate the highest score for each metric.

Dimension	Qwen3-VL-8B w/o CoT			Qwen3-VL-8B w/ CoT		
	Accuracy	F1-score	QWK	Accuracy	F1-score	QWK
Building Condition	0.750	0.732	0.752	**0.775**	**0.756**	**0.766**
Nature Environment	0.762	0.737	0.812	**0.775**	**0.773**	**0.821**
Public Space	0.717	0.694	0.745	**0.729**	**0.722**	**0.786**
Sanitation	**0.875**	0.603	**0.768**	0.829	**0.615**	0.644
Spatial Order	**0.817**	**0.730**	0.797	0.788	0.671	**0.815**
Overall Urgency	0.700	0.672	0.737	**0.792**	**0.760**	**0.785**

For moderate urgency, recall varies considerably, reaching 0.78 for building condition and 0.80 for overall urgency but falling to approximately 0.30 for sanitation and spatial order. This may reflect that moderate urgency represents an intermediate condition whose visual features are less distinctive than those of the low and high urgency categories and partly overlap with adjacent levels, making its classification boundaries harder to delineate. It may also stem from the label distribution, as the smaller samples of moderate cases in sanitation and spatial order limit the model's ability to learn the category characteristics.

3.3. Analysis of VLM reasoning outputs

The model generates textual explanations with urgency scores. Fig. 6 summarizes the vocabulary across each urgency level. Low urgency explanations frequently include clean, intact, and organized, whereas high urgency explanations include poor, deterioration, and blocked. Moderate urgency outputs emphasize continued function alongside a need for improvement. The vocabulary also reflects the focus of each assessment dimension (Fig. 7). In the nature environment dimension, healthy and maintained appear in favorable cases, while sparse and unmanaged occur in unfavorable cases. Other dimensions use terms related to building condition (e.g. clean or paint), public space (e.g. accessible or obstructed), sanitation (e.g. organized or accumulated), and spatial organization (e.g. orderly or chaotic). These patterns show that the explanations reference contextual visual cues and articulate dimension-specific reasons for urgency judgments.

The cases in Fig. 8 illustrate how the model integrates visual observations across dimensions to form renewal urgency judgments. Favorable conditions across the five dimensions support a low urgency score in Fig. 8a, while accumulated deterioration supports a high urgency score in Fig. 8d. In the moderate urgency case shown in Fig. 8b, the model identifies aging facades and functional but informally arranged vegetation and recommends small-scale improvements. The model also recognizes spatial conflicts, such as parking encroachment on pedestrian areas in Fig. 8c, and suggests spatial reorganization. These cases show how the explanations identify conditions requiring attention and make the reasons for the assigned urgency levels explicit.

3.4. Spatial distribution patterns of renewal urgency

The trained model assesses 26,030 images from 5206 housing estates. Scores for each estate are obtained by averaging and rounding the predictions from its five images. Fig. 9 presents the resulting distributions of renewal urgency. Sanitation is predominantly classified as low urgency, with 4800 estates (92.2%) receiving a score of 1. Moderate urgency (score 2) is the largest category in building condition, nature

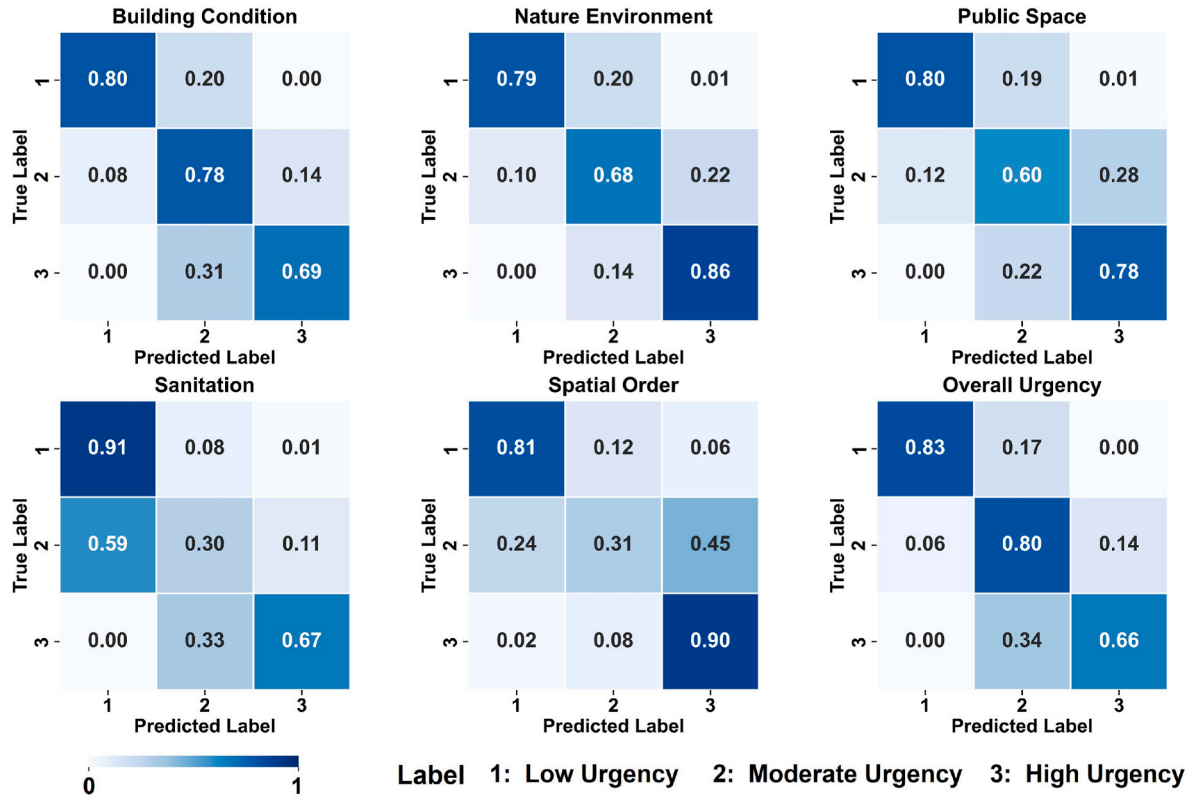


Fig. 5. Confusion matrices for LoRA-adapted Qwen3-VL-8B with CoT supervision across the five assessment dimensions and overall renewal urgency.

environment, and public space. Building condition has the highest number of estates in this category (3456), indicating a widespread need for targeted building maintenance. Spatial order has the largest number of estates with high urgency (score 3), at 1312, highlighting spatial reorganization and regulation as pressing renewal priorities. After averaging across five dimensions, overall renewal urgency is primarily moderate (56.6%) in Beijing, indicating most estates require targeted renewal. Meanwhile, 118 estates (2.3%) exhibit high overall urgency and may require comprehensive renewal interventions.

For spatial visualization, predictions for individual housing estates are aggregated to the community level (Fig. 10). We further examine correlations among building age, overall urgency, and ring road zone to assess the spatial distribution pattern. Mean overall urgency decreases from 1.82 inside the 2nd ring road to 1.45 outside the 5th ring road, while mean building age declines from 28.51 to 18.61 years (Fig. 11a and b). Spearman correlations show that ring road zone rank is negatively correlated with overall urgency ($\rho = -0.276$) and building age ($\rho = -0.349$), while building age is positively correlated with overall urgency ($\rho = 0.568$; all $p < 0.001$). These results indicate that renewal urgency decreases from central areas toward the outer ring road zones and suggest that higher urgency in central areas is associated with the concentration of older estates.

Dimension-specific distributions reveal further differences across ring road zones (Fig. 11c). For building condition, the proportion of estates with moderate urgency decreases from approximately 81% inside the 2nd ring road to 51% outside the 5th ring road. Across the same zones, the proportion with high urgency declines from 16% to 5% for public space and from 37% to 12% for spatial order. The proportion with low urgency in spatial order increases from approximately 20% to 62%. Sanitation remains broadly favorable across all zones, which may reflect routine sanitation management and environmental maintenance by property managers and residents. These patterns indicate higher demand for physical maintenance and spatial reorganization in older central areas.

Fig. 12 illustrates how the model assesses renewal urgency in representative housing estates. Case 1 is a newly constructed estate near the 5th ring road. All five images receive low urgency scores across all dimensions, indicating low renewal needs under the visual assessment criteria. Case 2 shows consistent moderate aging across the images. Dense clusters of overhead cables are the most prominent issue, leading to high urgency scores for spatial order and identifying cable management as an intervention priority. Case 3 represents an internally heterogeneous estate, where images 3 and 4 show parking congestion limiting pedestrian access, resulting in moderate or high urgency scores for public space. Overall, the estate shows mild aging and sparse but non-critical greenery.

Moreover, case 4 is in the city center and receives high urgency scores across all five dimensions at the housing estate level. The images show severe building deterioration, spatial disorder, and degraded public space, suggesting that the estate warrants priority for renewal. These cases illustrate how the model organizes observable visual cues into criterion-based, human-readable assessments. It is noteworthy that when assessing the nature environment, the VLM also captures seasonal factors, observing that leafless trees might indicate the winter season. Nevertheless, the model strictly adheres to the evaluation criteria, stating that "seasonal dormancy cannot serve as an excuse for the lack of planned green infrastructure," and assigns a high urgency score. This indicates that the framework is based on static images and predefined criteria, lacking the capability to make dynamic adjustments, such as accounting for seasonal ecological recovery.

4. Discussion

4.1. Domain adaptation and predictive performance

General-purpose VLMs provide a useful foundation for image-based assessment of renewal urgency. Across the five dimensions, zero-shot Qwen3-VL achieves a mean F1-score of 0.632, compared with 0.466

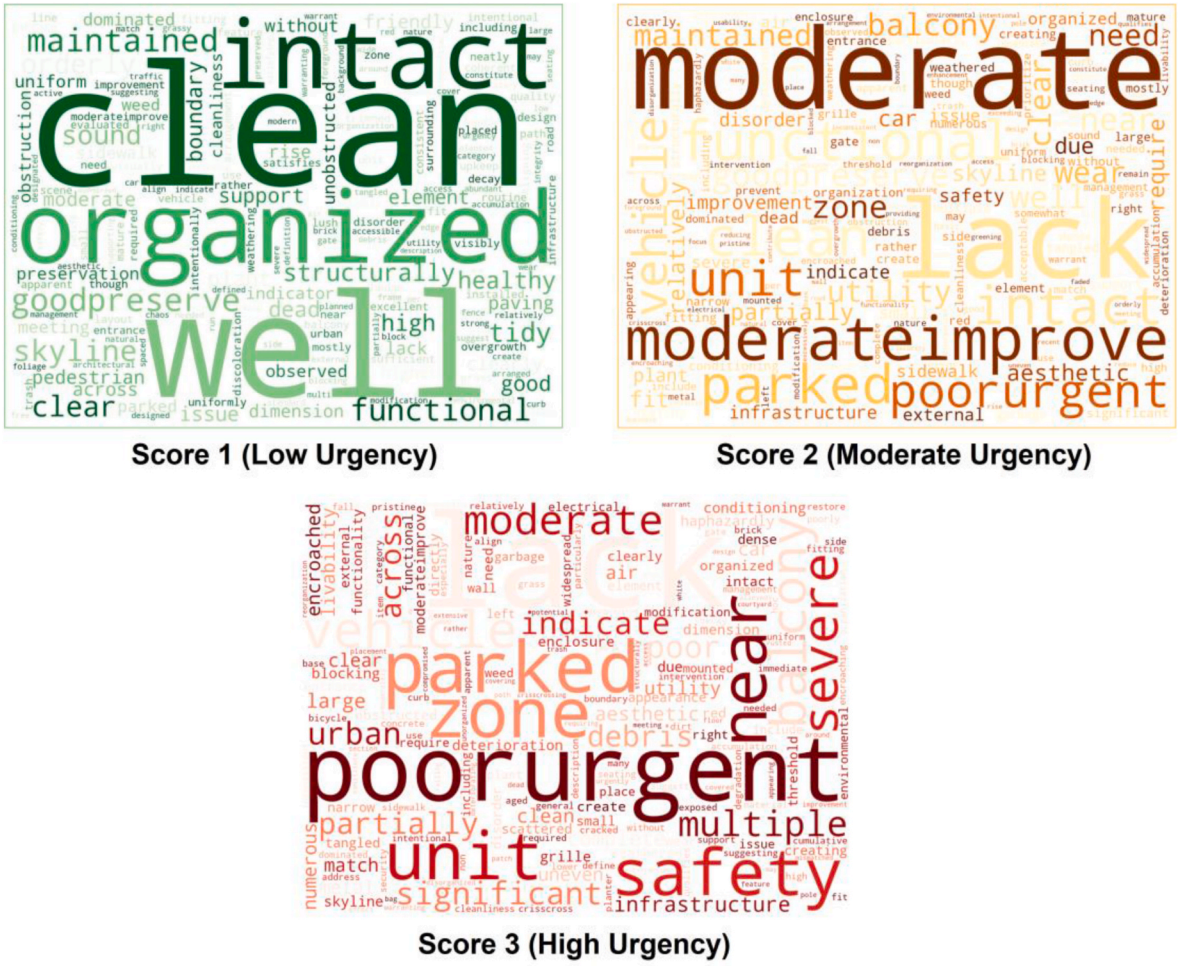


Fig. 6. Word clouds of model reasoning outputs for the three renewal urgency levels.

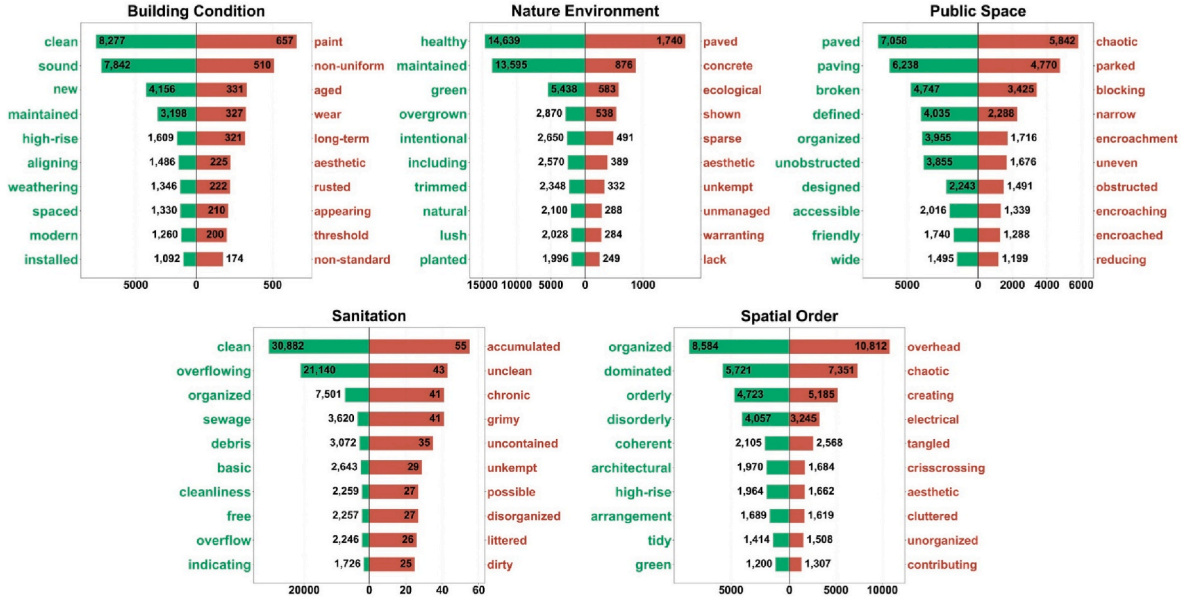


Fig. 7. Dimension-specific high-frequency words in VLM reasoning outputs. Green (left) and red (right) represent words extracted from cases with low and high renewal urgency, respectively. Values indicate word frequencies. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

for ResNet50. This performance suggests that pretrained vision-language knowledge is transferable to the interpretation of urban

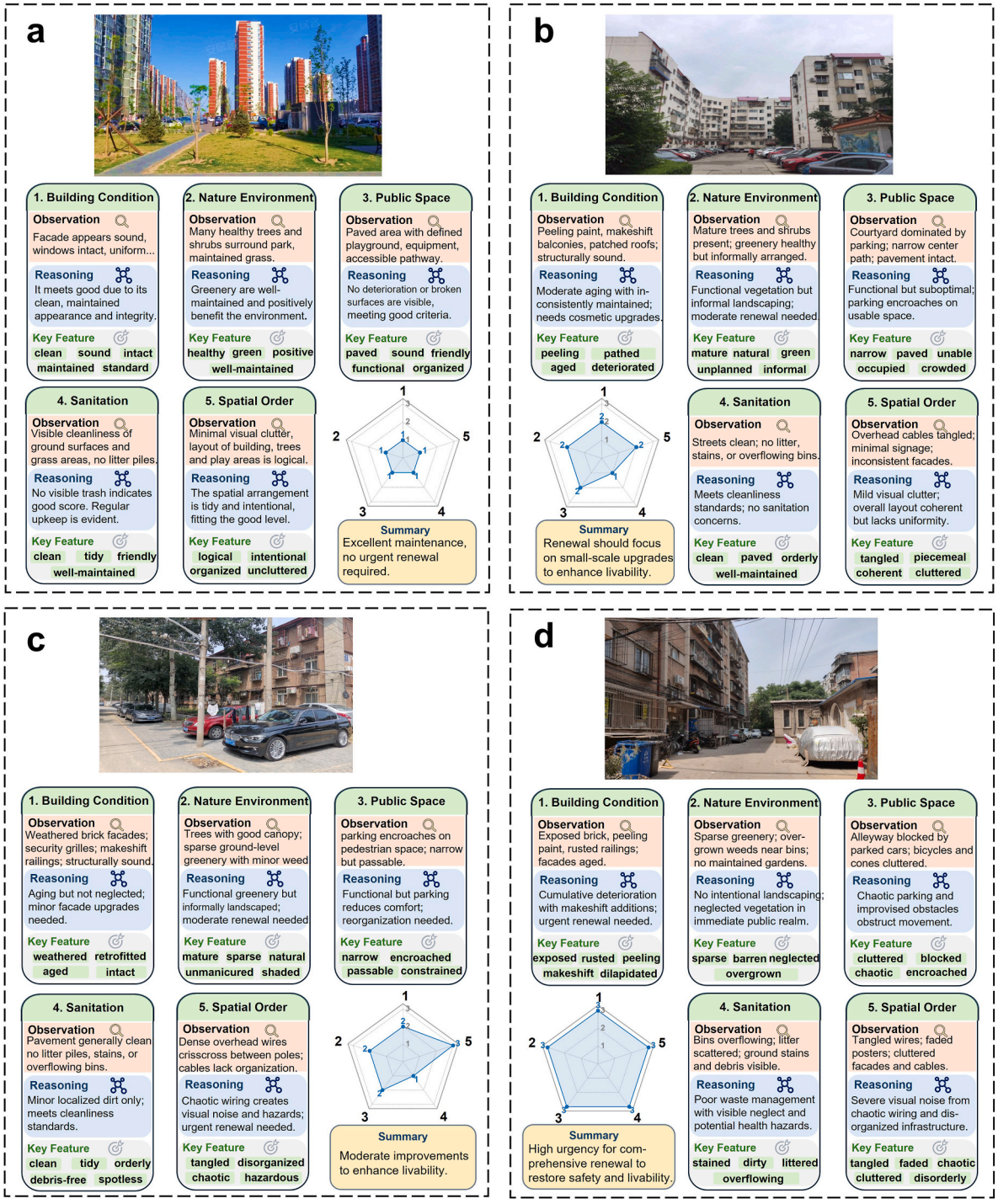


Fig. 8. Examples of CoT reasoning outputs for housing estates with different physical environment conditions.

environment conditions. Unlike traditional object recognition, renewal assessment requires judgments about physical condition, environmental maintenance, and spatial occupation. Similar visual elements can indicate different renewal needs depending on their condition and arrangement. By interpreting imagery alongside textual assessment criteria, VLMs can connect visual observations to these judgments. These findings support VLMs as a practical tool for interpreting the built environment in urban research.

LoRA-based domain adaptation further improves performance on the renewal assessment task. LoRA learns task-specific updates through low-rank matrices while keeping the pretrained backbone weights fixed, enabling adaptation without full-model fine-tuning. The LoRA-adapted model without CoT supervision shows improved performance across most assessment dimensions compared with zero-shot Qwen3-VL. This suggests that task-specific annotations help align the model's general vision-language capabilities with the defined scoring criteria. By linking visible residential conditions to urgency levels, these annotations provide supervision that general scene descriptions alone do not offer. Therefore, parameter-efficient fine-tuning offers a practical strategy for adapting pretrained VLMs to urban assessment tasks under constrained annotation and computational resources.

CoT supervision supports renewal urgency assessment by incorporating intermediate visual observations and assessment logic into model training. Compared with the variant trained without CoT annotations,

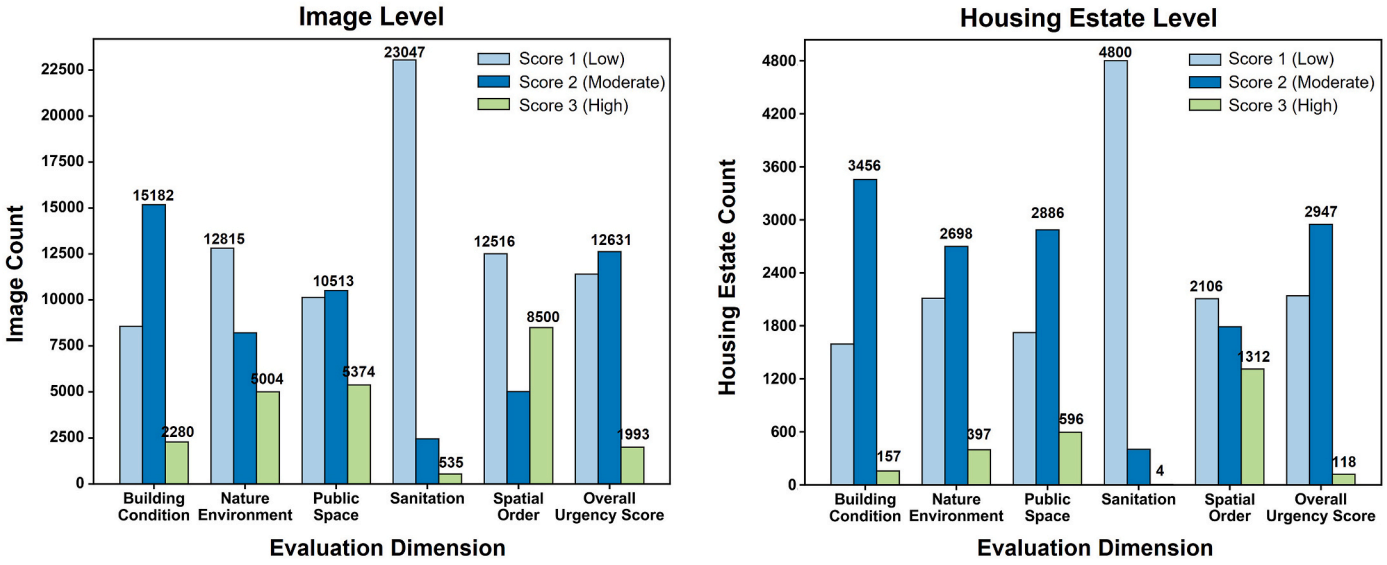


Fig. 9. Distribution of predicted urgency scores at the image level and the aggregated housing estate level.

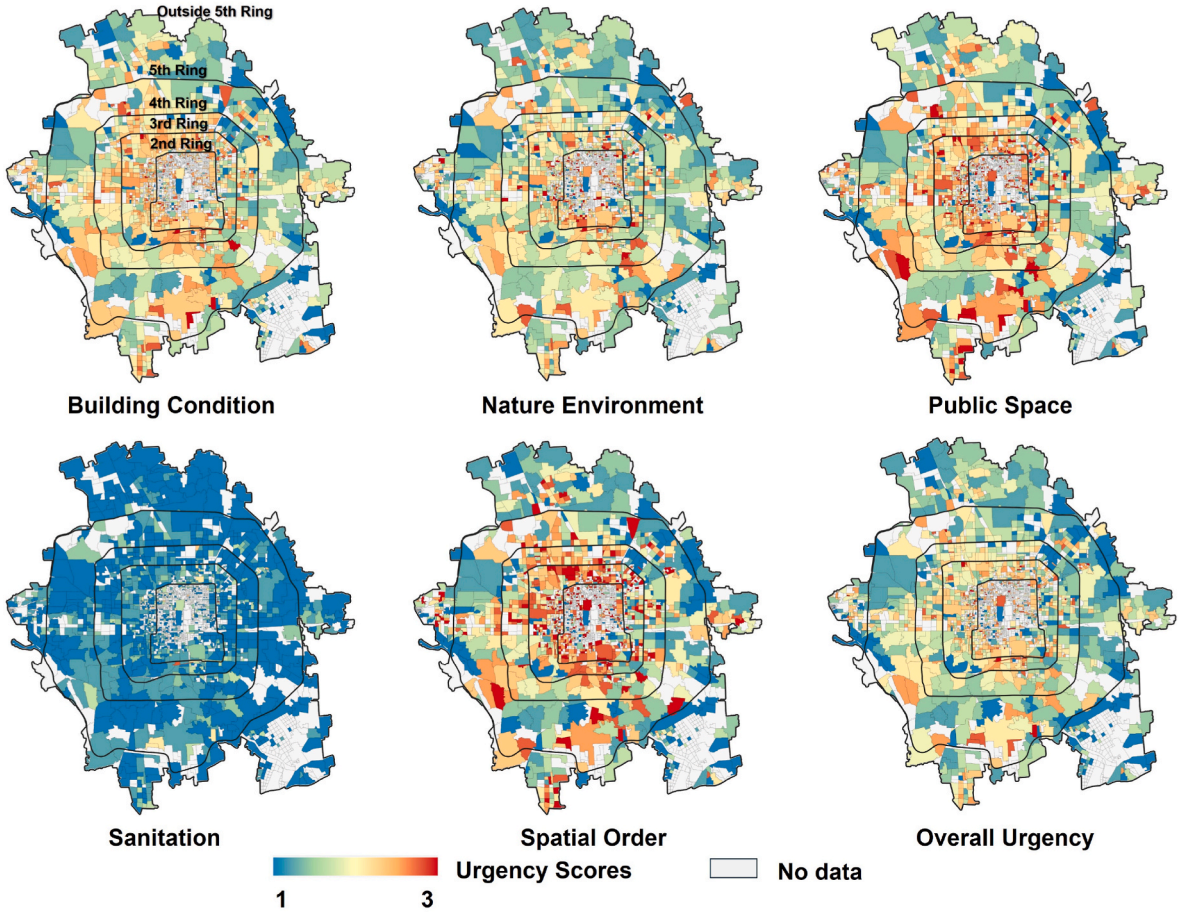


Fig. 10. Spatial distribution of five dimensions and overall renewal urgency at the community level in Beijing.

the model trained with CoT achieves higher scores on 14 of the 18 reported metrics across the five dimensions and overall urgency. For overall urgency, accuracy, F1-score, and QWK reach 0.792, 0.760, and 0.785, representing relative improvements of 13.1%, 13.1%, and 6.5%, respectively. These improvements suggest that CoT supervision helps the model interpret visual evidence according to the criteria for each dimension. It guides the model to identify residential conditions, evaluate renewal needs, and then assign urgency scores. Structuring the assessment into these steps can improve predictive performance and provide an explicit basis for interpreting the resulting judgments.

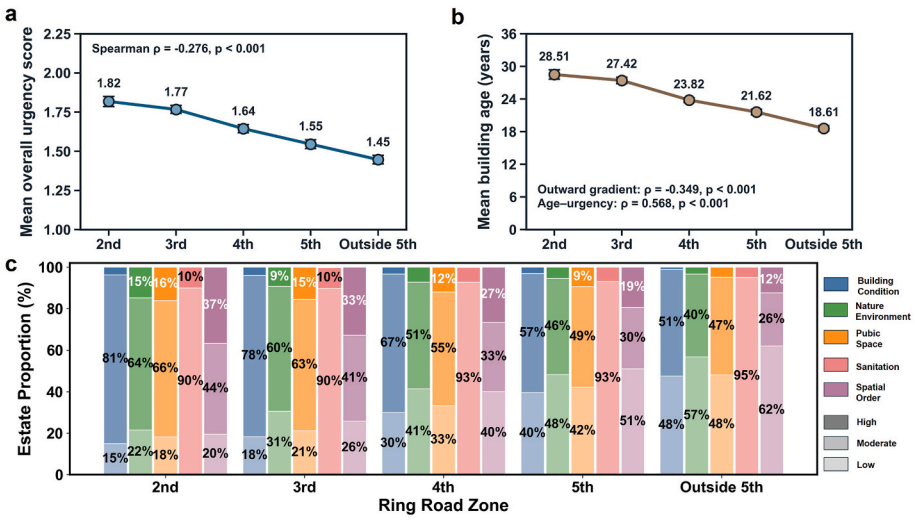


Fig. 11. Renewal urgency and building age across Beijing's ring road zones. (a, b) Mean overall urgency and building age, respectively, with Spearman correlations among urgency, building age, and ring road zone rank. Building age data are obtained from Anjuke. (c) Distributions of urgency levels across the five assessment dimensions.

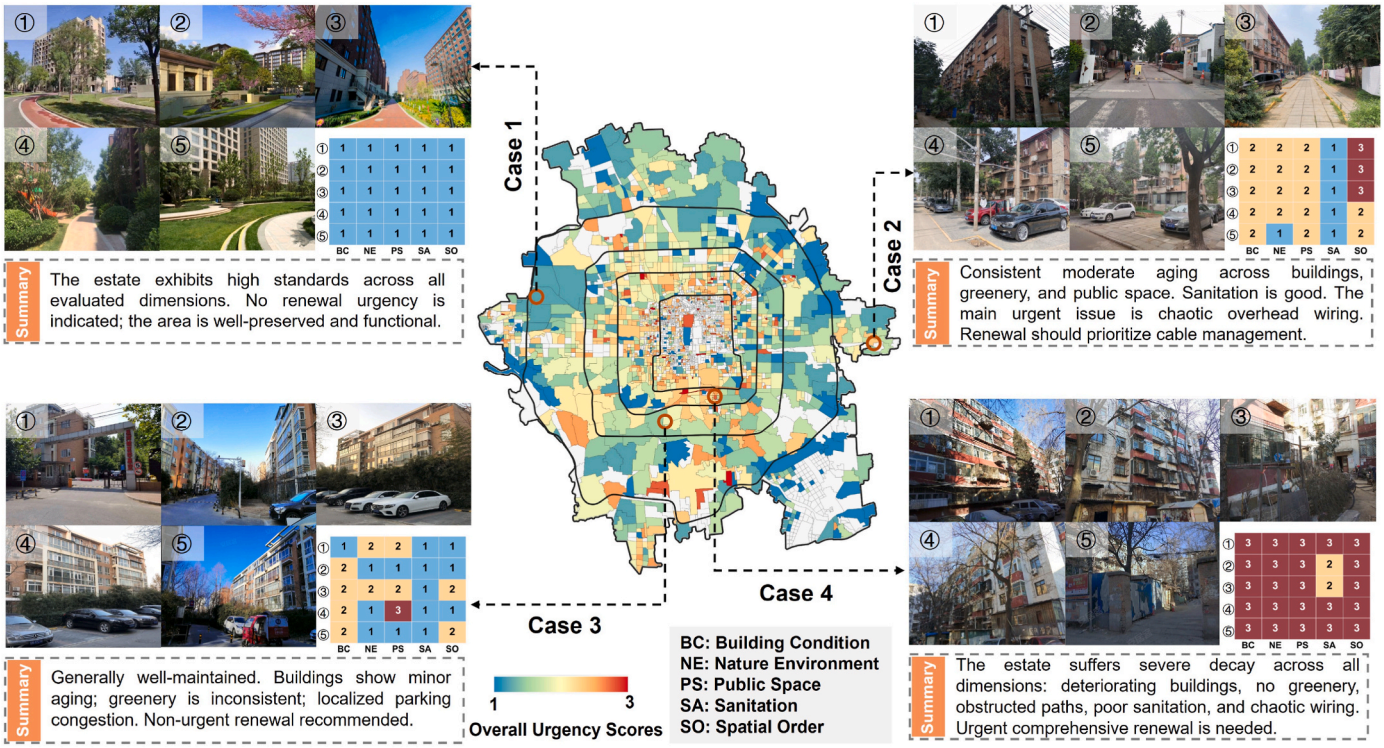


Fig. 12. Assessment of renewal urgency in representative housing estates using the LoRA-adapted VLM.

4.2. Vision-language reasoning for urban renewal decision-making

Vision-language reasoning provides interpretability by explaining how observed conditions support the assigned urgency scores. As shown in Fig. 6, explanations for low urgency emphasize maintenance and intact conditions, while those for high urgency describe deterioration and obstruction. Explanations for moderate urgency refer to continued function alongside a need for improvement. These differences help make the scoring rationale explicit, including in cases with mixed conditions. For example, in Fig. 8b, the model identifies aging facades that require improvement alongside functional greenery and well-maintained sanitation that can be preserved. This distinction provides a rationale for small-scale interventions targeting specific deficiencies. These explanations help planners interpret the specific requirements and scope of the renewal needs without only focusing on urgency scores.

Moreover, the spatial interpretations generated by VLMs offer a useful basis for context-aware urban assessment. In Fig. 8c, the model identifies vehicles occupying pedestrian space and interprets this arrangement as a spatial conflict requiring intervention. Vehicles in designated parking areas and those encroaching on pedestrian routes belong to the same object category but have different implications for accessibility. Assessing these differences requires interpreting spatial relationships and their implications for spatial management. CNNs can learn visual and spatial patterns, but their classification outputs do not

themselves explain why a particular arrangement warrants intervention. This ability to infer spatial relationships advances VLMs toward more spatially context-aware urban computing.

Dimension-specific explanations further clarify the sources of renewal need. The reasoning outputs closely align with planning terminology and provide targeted diagnostic information. Explanations for building condition describe visible wear and maintenance requirements, while those for nature environment address vegetation condition and management. Public space assessments concern accessibility and encroachment, while sanitation assessments address cleanliness and waste accumulation. Spatial order explanations describe how infrastructure and shared space are organized. Conditions can vary across these dimensions, as an estate may remain clean while showing facade aging or restricted access to public space. Retaining separate observations and scores for each dimension distinguishes physical maintenance needs from environmental and spatial management concerns within the overall assessment.

Retaining multidimensional scores also allows the assessment to accommodate different planning priorities. Local objectives may place different emphasis on building maintenance, environmental quality, and spatial organization. For example, higher weight can be assigned to building condition when policy prioritizes building renovation. Overall urgency can then be recalculated using weights and classification thresholds tailored to these objectives, while the predictions for each dimension remain unchanged. Comparing urgency classifications under different weighting and threshold settings shows which estates are prioritized under each planning scenario. This flexibility connects assessment results with policy objectives, providing a basis for targeted interventions and the allocation of renewal resources according to local needs.

4.3. Spatial patterns and policy implications

The center-to-periphery decline in renewal urgency is consistent with the concentration of older housing estates in central Beijing, where building deterioration and spatial encroachment are more pronounced. Estates in the outer zones are generally newer, with low and moderate urgency predominating. Across the study area, moderate overall urgency accounts for the largest share of estates (56.6%). Its prevalence also reflects the aging of Beijing's residential areas, most of which were built in the 1980s (Hu & Kaplan, 2001) and may lack systematic upgrades. These findings suggest that incremental improvements are appropriate for much of Beijing's residential stock, rather than comprehensive demolition and reconstruction.

Renewal priorities also differ across the five assessment dimensions. Spatial order has the highest proportion of estates with high urgency, with 1312 estates (approximately 25%) receiving a score of 3. This pattern may reflect cumulative spatial encroachment and inadequate coordination in managing overhead cables, stored objects, and parked vehicles. Building condition has the largest number of estates with moderate urgency (3456), indicating widespread but non-severe facade deterioration that calls for targeted maintenance rather than reconstruction. By comparison, approximately 4800 estates show low urgency for sanitation, suggesting relatively favorable sanitary conditions in most assessed estates. The higher urgency for public space and spatial order in inner ring road zones further highlights problems with the use and organization of shared space in older central areas. These differences support tailoring interventions to the specific combination of renewal needs identified in each housing estate.

This study supports the development of a differentiated policy framework. Planners can analyze spatial patterns to identify concentrations of renewal need and allocate screening and survey resources accordingly. At the housing estate level, dimension scores and reasoning outputs provide a diagnostic basis for selecting interventions. Estates with high overall urgency and deterioration across multiple dimensions can be prioritized for comprehensive renewal, while targeted measures are more appropriate and resource-efficient for specific deficiencies. The prevalence of high urgency for spatial order calls for particular attention to cable management, parking regulation, and the organization of shared space. Deficits in nature environment and public space can be addressed through landscape restoration, pavement rehabilitation, and courtyard redesign. Moderate urgency for building condition generally calls for targeted facade repair and maintenance, whereas high urgency may require structural assessment and more extensive rehabilitation. This framework provides a basis for decision-making through observable conditions. Practical implementation should also incorporate structural surveys, administrative records, resident needs, and implementation constraints.

4.4. Limitations and future work

Although this study demonstrates the potential of VLMs for image-based urban assessment, several limitations suggest directions for future research. As crowdsourced images vary in capture location and viewing angle, the selected images may not cover all areas within a housing estate even after manual screening. Although multiple images are selected to capture its physical environment, conditions outside the recorded views remain unassessed. Future work could combine imagery with building registration records, resident surveys, and remote sensing products to strengthen the evidence base for prioritizing renewal projects. Moreover, recall for moderate urgency remains relatively low in some dimensions, such as sanitation and spatial order, potentially reflecting insufficient samples and semantic confusion with adjacent levels. Data augmentation and class-balanced training strategies could be adopted to improve model performance in the future. In addition, while the annotation followed a criterion-based procedure, incorporating authoritative sources such as official inspection records or field verification could further enhance sample reliability.

The framework also needs to account more effectively for temporal variation in environmental conditions. In case 4 of Fig. 12, the VLM recognizes that leafless vegetation may reflect seasonal dormancy. However, the predefined criteria assign high urgency to a “complete lack of greenery” without accounting for potential seasonal recovery. The framework executes scoring rules based on the current visual state, lacking the contextual leniency required to adapt to complex variations. Future studies could develop adaptive scoring criteria that incorporate seasonal information or integrate multitemporal imagery to improve assessment reliability across seasons. Integrating generative AI models to visually simulate final renewal scenarios also represents a promising step for actionable planning.

5. Conclusion

This study develops an interpretable framework for assessing urban renewal urgency from housing estate imagery by integrating CoT reasoning with LoRA adaptation of Qwen3-VL-8B. Applied to 26,030 images from 5206 housing estates in Beijing, the framework assesses renewal urgency across five dimensions of the observable physical environment. CoT supervision improves model performance across most assessment dimensions. For overall renewal urgency, the model achieves an F1-score of 0.760 and a QWK of 0.785. These represent relative improvements of 13.1% and 6.5%, respectively, over the same model trained without CoT. Both metrics also exceed those obtained under zero-shot inference. Furthermore, the model generates human-readable explanations that make its assessment rationale explicit. Its spatial semantic understanding also enables it to interpret how space is organized and used, recognizing conflicts such as vehicles encroaching on pedestrian space and their implications for renewal. The assessment results reveal that Beijing's urban renewal mainly involves moderate renewal urgency, illustrating a spatial pattern where renewal urgency decreases gradually from the city center toward the outer ring road zones. By combining structured urgency scores with explicit diagnostic reasoning,

the framework provides a scalable, evidence-based approach to identifying renewal priorities and informing targeted interventions.

CRediT authorship contribution statement

Xiaoqin Yan: Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **Leyi Su:** Visualization, Investigation, Data curation. **Mengfan Tang:** Visualization, Methodology, Data curation. **Shuliang Ren:** Writing – original draft, Methodology, Funding acquisition, Conceptualization. **Yitian Liu:** Resources, Project administration. **Junnan Qi:** Project administration, Investigation. **Jiangpeng Zheng:** Validation, Methodology. **Yi Bao:** Validation, Supervision, Funding acquisition. **Zhou Huang:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge funding support from National Financial Fund Project of China (Grant No. KG2026174KYRW14), National Natural Science Foundation of China (U2344216, 42271471, 42401559) and Beijing Key Laboratory of Urban Spatial Intelligent Sensing and Digital Governance (NO. USISDG-2026-13).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeog.2026.104230.

References

Acevedo-De-los-Rios, A., Jones-Perez, J., & Rondinel-Oviedo, D. R. (2025). Methodology for prioritizing sustainable urban regeneration interventions in informal settlements: Case study in Lima. *Cities, 157*, Article 105550.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., … Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems, 35*, 23716–23736.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., … Zhu, K. (2025). Qwen3-VL technical report. *(arXiv:2511.21631)*.

Bai, Z., Gao, F., Wang, L., Du, J., Liu, S., & Ye, Y. (2026). Vision-language models reveal spatial and qualitative inequities in urban renewal: Evidence from Shanghai, China. *Applied Geography, 195*, Article 104152.

Cai, Z., Karvonen, A., Cong, C., & Huang, W. (2026). Fine-grained urban land use simulation: Integrating spatial dynamic modeling with a pre-trained vision-language model. *Computers, Environment and Urban Systems, 126*, Article 102416.

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

Elmqvist, T., Andersson, E., McPhearson, T., Bai, X., Bettencourt, L., Brondizio, E., … Van Der Leeuw, S. (2021). Urbanization in and for the Anthropocene. *npj Urban Sustainability, 1*(1), 6.

Fan, Z., Zhang, F., Loo, B. P. Y., & Ratti, C. (2023). Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences, 120*(27), Article e2220417120.

Feng, J., Wang, S., Liu, T., Xi, Y., & Li, Y. (2025). UrbanLLaVA: A multi-modal large language model for urban intelligence. In *2025 IEEE/CVF international conference on computer vision (ICCV)* (pp. 6209–6219). IEEE.

Gao, F., Cheng, H., Li, Z., & Yu, L. (2024). Revisiting the impact of public spaces on the mental health of rural migrants in Wuhan: An integrated multi-source data analysis. *International Journal of Health Geographics, 23*(1), 7.

Gao, F., Zheng, H., & Qin, B. (2025). Using multi-source big data to identify "double-aging" neighborhoods for urban retrofitting: A case study of Beijing. *Applied Geography, 180*, Article 103658.

Gemma Team. (2025). Gemma 3 technical report. *(arXiv:2503.19786)*. *arXiv*.

GLM-V Team. (2025). GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *(arXiv:2507.01006v1)*. *arXiv*.

Guo, Z., Xu, R., Yao, Y., Cui, J., Ni, Z., Ge, C., … Huang, G. (2024). LLaVA-UHD: An LMM perceiving any aspect ratio and high-resolution images. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer vision – ECCV 2024* (pp. 390–406). Nature Switzerland: Springer.

Han, H., Kim, S., Jin, M. Y., & Pettit, C. (2021). Providing affordable housing through urban renewal projects in Australia: Expert opinions on barriers and opportunities. *International Review for Spatial Planning and Sustainable Development, 9*(2), 41–61.

He, H., Wang, H., Zheng, L., Zhao, T., Zhang, P., & Zheng, K. (2025). Factors and pathways to enhance resident satisfaction in old residential neighbourhood renovation: A configuration analysis of cases in central Nanchang, China. *Buildings, 15*(7), 1125.

He, J., Zhang, J., Yao, Y., & Li, X. (2023). Extracting human perceptions from street view images for better assessing urban renewal potential. *Cities, 134*, Article 104189.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE.

Hong, Q., Zhao, H., & Long, Y. (2026). A multi-source and multi-perspective imagery method for assessing the physical appearance of gated communities. *Environment and Planning B: Urban Analytics and City Science, 53*(7), 1617–1636.

Howell, A., Wu, N., Bagchi-Sen, S., Kim, Y., & Sun, Q. C. (2026). Multimodal large language models, street view images and urban policy-intelligence: Recovering the sustainability effects of redlining. *npj Urban Sustainability, 6*, 79.

Hu, X., & Kaplan, D. H. (2001). The emergence of affluence in Beijing: Residential social stratification in China's capital city. *Urban Geography, 22*(1), 54–77.

Huang, G., Song, X., & Jiao, H. (2026). From pixels to perception: Fine-tuning vision-language models for semantic urban perception. *International Journal of Digital Earth, 19*(1), Article 2677402.

Huang, Y., Zhang, F., Gao, Y., Tu, W., Duarte, F., Ratti, C., … Liu, Y. (2023). Comprehensive urban space representation with varying numbers of street-level images. *Computers, Environment and Urban Systems, 106*, Article 102043.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., & Zhu, X. X. (2018). Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing, 145*, 44–59.

Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening, 14*(3), 675–685.

Liu, B., Wang, X., Xia, N., & Ni, W. (2018). Critical success factors for the management of public participation in urban renewal projects: Perspectives from governments and the public in China. *Journal of Urban Planning and Development, 144*(3), Article 04018026.

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 26296–26306). IEEE.

Liu, L., Silva, E. A., Wu, C., & Wang, H. (2017). A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Computers, Environment and Urban Systems, 65*, 113–125.

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., … Ruan, C. (2024). DeepSeek-VL: Towards real-world vision-language understanding. *(arXiv:2403.05525)*. *arXiv*.

Lu, T., Zhang, F., & Wu, F. (2020). The variegated role of the state in different gated neighbourhoods in China. *Urban Studies, 57*(8), 1642–1659.

Lu, Y., & Chen, H. (2024). Using google street view to reveal environmental justice: Assessing public perceived walkability in macroscale city. *Landscape and Urban Planning, 244*, Article 104995.

Ma, X., Ma, C., Wu, C., Xi, Y., Yang, R., Peng, N., … Ren, F. (2021). Measuring human perceptions of streetscapes to better inform urban renewal: A perspective of scene semantic parsing. *Cities, 110*, Article 103086.

Nachmany, H., & Hananel, R. (2023). The urban renewal matrix. *Land Use Policy, 131*, Article 106744.

Perez, J., & Fusco, G. (2025). Streetscape analysis with generative AI (SAGAI): Vision-language assessment and mapping of urban scenes. *Geomatica, 77*(2), Article 100063.

Raghu, D., Armeni, I., & De Wolf, C. (2026). Urban-scale facade material mapping from street view images using vision-language models for circular construction planning. *Scientific Reports, 16*, Article 20691.

Ren, S., Huang, Z., Yin, G., Yan, X., Dong, Q., Qi, J., … Zhang, S. (2025). Spatially-optimized greenspace for more effective urban heat mitigation: Insights from regional cooling heterogeneity via explainable machine learning. *Landscape and Urban Planning, 256*, Article 105296.

Roberts, P. (2000). The evolution, definition and purpose of urban regeneration. In P. Roberts, & H. Sykes (Eds.), *Urban regeneration: A handbook* (pp. 9–36). SAGE.

Sun, D., Lu, Y., Qin, Y., Lu, M., Song, Z., & Ding, Z. (2024). Method for evaluating urban building renewal potential based on multimachine learning integration: A case study of Longgang and Longhua districts in Shenzhen. *Land, 14*(1), 15.

Wang, J., Xuan, W., Qi, H., Chen, Z., Chen, H., Zheng, Z., … Yokoya, N. (2026). CityVLM: Towards sustainable urban development via multi-view coordinated vision-language model. *ISPRS Journal of Photogrammetry and Remote Sensing, 232*, 62–74.

Wang, S., Cao, J., & Yu, P. S. (2020). Deep learning for spatio-temporal data mining: A survey. *IEEE Transactions on Knowledge and Data Engineering, 34*(8), 3681–3700.

Wang, Y., Ye, Y., & Weng, C. (2026). A scalable framework for street interface morphology assessment via automated multimodal large language model agents. *Land, 15*(4), 610.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Wu, Y., Galdini, R., Hui, E. C. M., & Long, H. (2020). Urban regeneration and re-use: China and Europe. *Cities, 106*, Article 102863.

Xia, C., Yeh, A. G. O., & Zhang, A. (2020). Analyzing spatial relationships between urban land use intensity and urban vitality at street block level: A case study of five Chinese megacities. *Landscape and Urban Planning, 193*, Article 103669.

Xing, J., & Sieber, R. (2023). The challenges of integrating explainable artificial intelligence into GeoAI. *Transactions in GIS, 27*(3), 626–645.

Yan, X., Huang, Z., Ren, S., Zhu, Q., Yin, G., Qi, J., & Bao, Y. (2026). Explainable urban renewal prediction at building-scale using hierarchical graph neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing, 233*, 609–622.

Yang, C., & Zhang, Y. (2026). Assessing and mapping public visual perception across urban public space typologies based on geotagged social media images. *Applied Geography, 194*, Article 104097.

Yao, Y., Liang, Z., Yuan, Z., Liu, P., Bie, Y., Zhang, J., … Guan, Q. (2019). A human-machine adversarial scoring framework for urban perception assessment using street-view images. *International Journal of Geographical Information Science, 33*(12), 2363–2384.

Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, 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.

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(8), 5625–5644.

Zhang, J., Li, Y., Fukuda, T., & Wang, B. (2025). Urban safety perception assessments via integrating multimodal large language models with street view images. *Cities, 165*, Article 106122.

Zhao, T., Wang, S., Ouyang, C., Chen, M., Liu, C., Zhang, J., … Wang, L. (2024). Artificial intelligence for geoscience: Progress, challenges, and perspectives. *The Innovation, 5*(5), Article 100691.

Zhao, X., Lu, Y., & Lin, G. (2024). An integrated deep learning approach for assessing the visual qualities of built environments utilizing street view images. *Engineering Applications of Artificial Intelligence, 130*, Article 107805.

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review, 57*(4), 99.

Zhao, X., Xia, N., & Li, M. (2025). Assessing urban renewal opportunities by combining 3D building information and geographic big data. *Geo-Spatial Information Science, 28*(3), 1108–1124.

Zheng, H. W., Shen, G. Q., & Wang, H. (2014). A review of recent studies on sustainable urban renewal. *Habitat International, 41*, 272–279.

Zheng, L., Song, Q., Kwan, M.-P., van Ameijde, J., Tian, S., & Zheng, Y. (2026). A large language model framework for urban perception mapping: Linking greenspace patterns and park attractiveness across the pandemic. *Computers, Environment and Urban Systems, 127*, Article 102433.

Zhu, H., Su, J., Chen, M., Wang, W., Deng, Y., Chen, G., & Zhang, W. (2025). PlanGPT-VL: Enhancing urban planning with domain-specific vision-language models. In S. Potdar, L. Rojas-Barahona, & S. Montella (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing: Industry track* (pp. 2461–2483). Association for Computational Linguistics.

Zhuang, T., Qian, Q. K., Visscher, H. J., Elsinga, M. G., & Wu, W. (2019). The role of stakeholders and their participation network in decision-making of urban renewal in China: The case of Chongqing. *Cities, 92*, 47–58.