buildings

MDPI

Article

Human–AI Alignment in Architectural and Streetscape Evaluation: A Multi-City Study Using Vision–Language Models and Explainable Street-View Analysis

Yuanyuan Mou [1] and Xin Li [2,*]

[1] The Bartlett School of Architecture, University College London, London WC1H 0QB, UK; yuanyuanmou2026@163.com
[2] College of Civil Engineering and Architecture, Guangxi University, Nanning 530004, China
* Correspondence: lixin1983@gxu.edu.cn

Abstract

AI systems may agree with one another about the quality of buildings and streets, even when their judgments differ from people's. We tested this possibility using 120 paired building-frontage and streetscape images selected from 39,047 street-view panoramas across six Chinese urban districts. The images were rated by 100 architecture or urban-planning students, 100 students from other disciplines, and five vision–language models (VLMs). The two student groups strongly agreed on how they ranked the images (Spearman correlations of 0.914–0.946 for building frontages and 0.923–0.943 for streetscapes), although architecture/planning students consistently gave lower scores. By contrast, the AI models agreed relatively well with one another but showed little correspondence with either student group in their rankings. We also examined whether visible elements such as vegetation, roads, sidewalks, and sky could explain the ratings. These simple image features predicted AI streetscape scores more successfully than human scores, consistent with the possibility that their evaluations reflect different visual cues, although the study could not establish how either reached its judgments. AI may support preliminary screening of large image collections, but human evaluation remains essential for architectural and urban-design decisions.

Keywords: vision–language models; Human–AI agreement; architectural evaluation; streetscape perception; street-view imagery; semantic segmentation; XGBoost; SHAP

Check for updates

Academic Editor: Nikos A. Salingaros

Received: 24 September 2026
Revised: 3 November 2026
Accepted: 13 November 2026
Published: 16 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The visual qualities of buildings and streets constitute a fundamental dimension of human experience in cities. Building façades form the immediate architectural interface between individual buildings and public space, while streetscapes integrate façades, vegetation, pedestrian facilities, roads, vehicles, open sky, and other environmental elements into a continuous perceptual field. These visual characteristics influence not only aesthetic judgments but also affective responses, environmental preference, perceived safety, comfort, and broader evaluations of urban quality. Recent research has demonstrated that architectural form, geometry, materiality, visual complexity, and spatial enclosure can elicit measurable psychological and physiological responses [1]. At the street scale, machine-learning and image-based studies have further shown that vegetation, buildings, roads, openness, and other visible elements are systematically associated with mood states, perceived safety, beauty, and overall environmental perception [2–4]. In practical terms, this

study asks whether AI systems judge the same buildings and streets in ways that agree with human observers, and whether those judgments differ between people with and without architecture-related training. The goal is to help architects and urban planners understand when AI-based visual assessments may support preliminary screening and when human judgment remains necessary.

Traditional assessments of architectural and streetscape quality rely on field observations and participant surveys, but these approaches are difficult to scale. Street-view imagery (SVI) enables systematic description of visible vegetation, buildings, sky, pedestrian facilities, and related environmental conditions [5–11]. However, the resulting indicators depend on viewpoint, field of view, and sampling procedures [5,11]. Combined with semantic segmentation and interpretable learning, SVI has been used to examine associations between scene components and perceived road safety, visual quality, restoration, and walkability [12–16].

Perceptual ratings also vary across evaluation dimensions and observer groups. Perceived safety does not necessarily reflect actual risk, and environmental preferences can differ across populations [8,12,14,17]. Reviews document substantial conceptual and methodological heterogeneity in this field [18]. Reliable human reference scores therefore require standardized stimuli, sufficient ratings per image, and transparent survey procedures [19].

Within this broader visual environment, the architectural façade deserves particular attention. A façade is not simply one semantic category within a street image; it is a designed architectural composition whose evaluation involves relationships among proportion, rhythm, material and color, articulation, scale, visual hierarchy, and architectural identity. Recent studies have therefore begun to isolate building exteriors from general streetscape perception. Liang et al. [20] developed a large-scale SVI-based framework for evaluating building exteriors across multiple cities and demonstrated that perceived characteristics of buildings contribute meaningfully to overall streetscape quality. Lee et al. [21] further examined façade complexity and visual preference using street-view images and showed that human visual processing does not necessarily correspond directly to computational measures of façade complexity. These findings suggest an important distinction between identifying visible façade properties and making higher-order architectural judgments about their quality. Although building façades form an integral part of the streetscape, evaluating an individual building frontage is not equivalent to evaluating the surrounding street environment. Façade evaluation emphasizes architectural composition, proportion, material coherence, articulation, and identity, whereas streetscape perception additionally involves vegetation, pedestrian infrastructure, roadway conditions, spatial openness, and the relationships between buildings and public space. Examining these two domains using visual stimuli derived from the same street-view locations therefore provides a common spatial basis for investigating whether Human–AI agreement differs across architectural and contextual evaluation tasks. This joint perspective allows the identification of agreement patterns that persist across both domains, as well as task-specific differences that may be obscured when façades and streetscapes are investigated independently.

Architectural education and professional experience may influence how people interpret and evaluate buildings, raising a longstanding question about the relationship between disciplinary expertise and public architectural preference. Classical environmental-psychology research identified important differences between architects and laypersons. Gifford et al. [22] reported that the aesthetic judgments of practicing architects and lay observers were almost unrelated for the modern buildings examined, partly because the groups responded to different physical characteristics of the façades. Brown and Gifford [23] further demonstrated that practicing architects had difficulty predicting lay evaluations of individual buildings, although they recognized that lay observers generally assigned

more favorable ratings. Subsequent work connected these differences to the cognitive interpretation of architectural attributes, including complexity, clarity, originality, and meaningfulness [24]. These studies established architectural expertise as an important source of heterogeneity in built-environment evaluation.

However, the relationship between architectural training and visual preference is not necessarily uniform across observer populations or stages of expertise. Šafárová et al. [25] found that young architects and non-architect peers did not significantly differ in their aesthetic evaluations of the buildings examined, suggesting that findings based on experienced professionals should not automatically be generalized to students or early-career observers. More recent virtual-environment research has also documented both commonalities and differences between architectural experts and non-experts in the perception of urban ambiance [26]. Theoretical work on architectural geometry and environmental preference further highlights the potential importance of spatial organization, symmetry, and other relational properties in subjective visual experience [27]. These findings can be understood through two conceptually distinct but potentially related pathways. First, architecture-related education may influence the relative importance assigned to visual attributes such as composition, proportion, and spatial organization, potentially affecting how environments are ranked. Second, disciplinary training may shape the evaluative criteria used to translate overall visual impressions into numerical ratings, potentially producing differences in absolute scores even when relative rankings remain similar. Accordingly, the present study distinguishes between ranking agreement and rating-level differences without assuming that the observed differences are necessarily caused by architectural education.

Against this background, recent developments in multimodal artificial intelligence introduce a new methodological opportunity. Vision–language models (VLMs) learn relationships between visual and textual representations and can perform image interpretation, comparison, description, and zero-shot evaluation without the task-specific training pipelines required by many conventional computer-vision models [28]. In architectural and urban research, generative and multimodal AI is expanding rapidly, particularly in early-stage design, image-based design exploration, planning support, and comparative evaluation [29,30]. A recent systematic review of generative AI in architectural design found that image-based applications dominate current research and that subjective comparative evaluation remains one of the principal means of validating AI-generated architectural outputs [29]. At the urban scale, large multimodal models are also increasingly discussed as potential assistants for planning analysis, scenario generation, and outcome evaluation rather than merely as text-generation systems [30].

More directly relevant to environmental perception, emerging evidence suggests that multimodal models can reproduce some dimensions of human visual judgment but struggle with others. Comparisons between human observers, GPT-based systems, and LLaVA have shown substantial alignment for several landscape-preference dimensions, while more nuanced properties such as legibility remained difficult for AI to reproduce [31]. In architectural façade research, Law et al. [32] similarly reported stronger human–AI alignment for relatively explicit objective descriptors than for affective descriptors, suggesting that machine interpretation becomes less reliable as evaluation moves from an observable form toward higher-order experiential meaning. Such findings are especially important for architectural quality assessment because judgments of composition, proportion, material coherence, identity, pleasantness, and safety operate at different levels of perceptual abstraction. At the intersection of multimodal AI and human perception, Qin et al. [33] integrated historic-street imagery and environmental audio with AI-assisted perceptual assessment and human validation. Their study highlights the importance of validating model-generated evaluations against human judgments while demonstrating that visually observable envi-

ronmental characteristics alone may not adequately represent the multisensory qualities of place. This evidence further supports the need to distinguish the reproducibility of AI-generated visual assessments from their correspondence with actual human perceptions.

Meanwhile, VLM-based built-environment analysis is progressing from experimental demonstrations toward city-scale applications. Automated street-view acquisition and transformer-based image selection have been developed to improve the efficiency of extracting usable building imagery [34]. Raghu et al. [35] recently applied multimodal AI to more than nine thousand street-level images across six cities and demonstrated that VLMs can retrieve façade materials, architectural style, condition, and related building-level characteristics with substantial accuracy. At the perceptual level, Ma and Kwan [36] combined human experience, street-view imagery, and VLMs to measure psychological restorative quality at urban scale, demonstrating the potential of vision–language approaches to extend subjective environmental assessment beyond conventional surveys. These developments suggest that VLMs are increasingly capable of interpreting built-environment imagery at scale, but they do not establish that such models can reliably reproduce architecture-specific or streetscape-specific human judgment.

Architectural-perception studies increasingly use immersive visualization, eye tracking, and biometric measures to investigate visual attention and affective responses [37–41]. These approaches indicate that numerical ratings do not fully reveal the perceptual process. Human–AI studies have likewise found task-dependent correspondence: GPT-4V and human observers sometimes agree in architectural-ambiance classifications [42], whereas multi-model comparisons with design experts reveal discrepancies in more subjective assessments [43]. Questions consequently remain about whether computational representations adequately capture architecturally meaningful qualities [44]. This study therefore evaluates image rankings alongside absolute scores, rather than treating similar average ratings as evidence of equivalent architectural judgment.

A related issue concerns interpretability. Prediction or rating consistency alone does not reveal which visible environmental characteristics are associated with human or AI assessments. This distinction is particularly important in architecture, where a numerically similar evaluation can emerge from different visual cues or evaluative criteria. Recent explainable-AI research in architectural image understanding has therefore emphasized the need to connect computational outputs with interpretable architectural evidence rather than relying solely on black-box predictions [45]. In urban perception research, SHAP and related explainable-learning techniques have likewise been used to reveal nonlinear associations between vegetation, buildings, roads, sidewalks, walls, vehicles, and subjective environmental evaluations [12,13]. Comparing the feature–rating relationships derived from AI and human evaluations can therefore provide information beyond conventional score correlation: it can indicate whether similar ratings are associated with similar observable characteristics of the street environment.

Despite this progress, several interrelated research limitations remain, centering on whether contemporary VLMs can reliably reproduce human judgments across different domains of built-environment evaluation and different human reference populations. First, most SVI-based perception studies have focused either on predicting human ratings or on extracting objective environmental indicators, whereas systematic comparisons between multiple contemporary VLMs and human evaluators remain relatively limited [13,15,18]. Second, emerging Human–AI comparisons in architectural and landscape evaluation have generally examined a limited number of models or visual scenes [31,32], while differences between observers with and without architecture-related training further complicate the definition of an appropriate human reference benchmark [46]. Importantly, such differences may involve absolute rating levels, relative preference rankings, or both. Third, building-

façade assessment and broader streetscape perception have commonly been investigated separately, even though façades constitute an integral component of the street visual environment and the two domains represent distinct but related levels of visual evaluation [20,21]. Examining paired façade and streetscape stimuli derived from the same locations provides an opportunity to determine whether Human–AI agreement patterns persist across these complementary evaluation domains or vary with the assessment task. Fourth, agreement in final scores or rankings does not necessarily indicate whether human and AI evaluations exhibit comparable associations with observable street-space characteristics. Consequently, the central research problem is not the integration of multiple analytical techniques itself, but the extent to which Human–AI alignment depends on the evaluation domain and human reference population, and whether measurable environmental features can help characterize differences between human and model-generated streetscape ratings.

Accordingly, this study investigates the extent to which contemporary VLMs reproduce human evaluations of architectural frontages and urban streetscapes, with particular attention to the influence of human reference populations and evaluation domains. Three interconnected research questions are addressed. RQ1: To what extent do architecture/urban-planning students (AUP) and students without such training (NAP) agree in their relative rankings and absolute ratings of architectural frontages and streetscapes? RQ2 (primary research question): How closely do contemporary VLMs reproduce the rankings and absolute scores of these two human groups, and how do the observed agreement patterns compare between architectural-frontage and streetscape evaluation? RQ3: To what extent can measurable street-space semantic characteristics predict human and VLM streetscape evaluations, and what do differences in predictive performance and feature attribution suggest about the limitations of semantic explanations of perceptual ratings?

To answer these questions, we establish a multi-city evaluation framework using 39,047 Baidu Street View panoramas collected from six urban districts in Guangzhou, Changsha, Beijing, Zhengzhou, Shenzhen, and Xi'an. A controlled set of 120 paired architectural-frontage and streetscape stimuli was evaluated by 100 AUP students, 100 NAP students, and five contemporary VLMs. Spearman rank correlation, Lin's concordance correlation coefficient, and mean absolute error were used to assess agreement within and between human and AI evaluation sources. For streetscape evaluation, SegFormer-derived semantic indicators and XGBoost–SHAP analyses were further employed to examine the predictive and explanatory relationships between visible environmental characteristics and evaluator scores. The empirical analysis follows the three research questions: Section 3.1 establishes the human evaluation reference, Section 3.2 examines Human–AI agreement across the two domains, Section 3.3 investigates semantic-feature associations with streetscape ratings, and Section 3.4 evaluates cross-group and cross-city heterogeneity and robustness.

2. Materials and Methods

2.1. Multi-City Panoramic Street-View Dataset

Six administrative districts were selected as study areas: Haizhu District in Guangzhou, Tianxin District in Changsha, Haidian District in Beijing, Zhongyuan District in Zhengzhou, Luohu District in Shenzhen, and Yanta District in Xi'an (Figure 1). These districts were examined as comparative urban case-study areas rather than as a statistically representative sample of Chinese cities. Using a consistent district-level spatial unit allowed Human–AI evaluation patterns to be examined across different geographical and urban contexts. Nevertheless, a single district cannot represent the full diversity of building types and street configurations within its city, and the findings should therefore be interpreted within the spatial scope of the selected study areas.

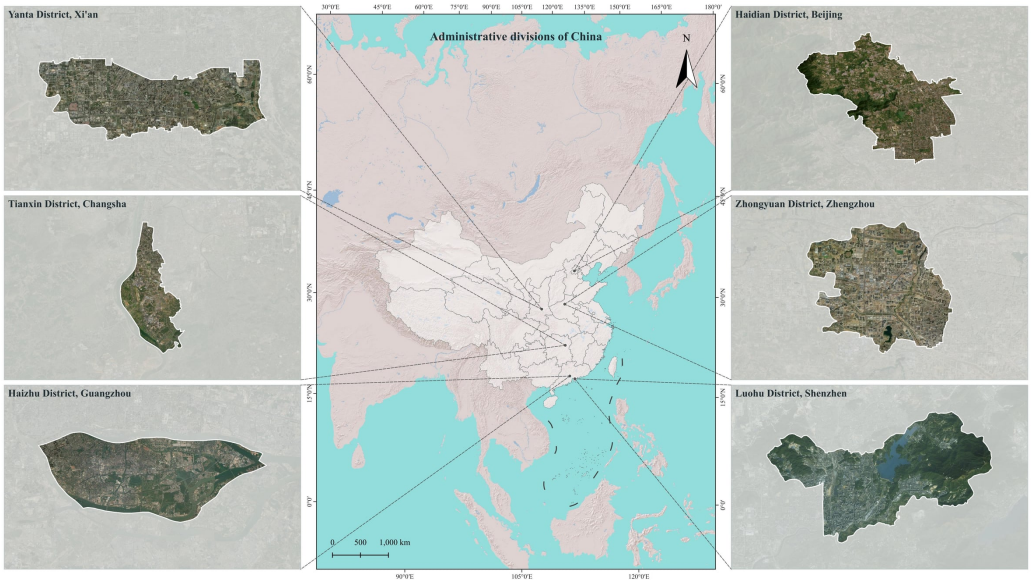

Figure 1. Research area. In the six inset maps, the darker areas indicate the administrative districts selected for this study.

Street-view sampling was conducted along urban branch roads and pedestrian-accessible streets within each district. Sampling points were generated at 100 m intervals, while expressways and road sections with limited pedestrian relevance were excluded. A georeferenced 360° panoramic image was retrieved from Baidu Street View for each valid location. Associated metadata included the sampling identifier, coordinates, Street View ID (SVID), panorama acquisition time, image filename, and acquisition status. After acquisition and integrity screening, 39,047 valid panoramic images were retained. All images had a standardized resolution of 2048 × 1024 pixels and a 2:1 aspect ratio.

All evaluation stimuli were derived from the same Baidu Street View panoramas. At each of the 120 selected locations, the original 2048 × 1024 panorama was retained for streetscape evaluation, while a corresponding building-oriented perspective image was generated for architectural-frontage evaluation. The two stimuli were linked using a unique identifier (STIM_ID). For architectural-view generation, 12 candidate rectilinear views were produced at yaw angles from 0° to 330° in 30° increments, with a fixed horizontal field of view of 90°, a pitch of 0°, and an output resolution of 768 × 768 pixels. A pretrained SegFormer-B0 model (nvidia/segformer-b0-finetuned-ade-512-512) was used to estimate the visible proportions of buildings, sky, and roads. An automated selection score favored higher overall and central building coverage while penalizing sky and roadway proportions. This procedure selected building-oriented views rather than identifying or isolating individual building instances. No fixed minimum building-coverage threshold was imposed; the median building-pixel proportion was 49.7% (interquartile range: 38.4–63.7%). Seven views with negative selection scores were flagged for visual review in the frozen stimulus manifest and retained in the analytical sample. Once the stimuli were frozen, no further cropping, recompression, or image replacement was permitted, and both human groups and all five VLMs evaluated the identical image files.

Separately, streetscape semantic indicators were derived from four 1024 × 1024 rectilinear views per panorama at yaw angles of 0°, 90°, 180°, and 270°, with a fixed pitch of 0° and a horizontal field of view of 90°, as detailed in Section 2.2. This procedure was distinct from the architectural-view selection process.

The final dataset comprised 17,144 panoramas in Haidian District, 5389 in Tianxin District, 5749 in Haizhu District, 2760 in Zhongyuan District, 3122 in Luohu District, and 4883 in

Yanta District, yielding a total of 39,047 valid panoramic sampling locations. The spatial distribution of the study areas is shown in Figure 1, and the dataset composition is summarized in Table 1.

Table 1. Overview of the study areas and image dataset.

City	Administrative District	Sampling Interval	Valid Sampling Points	Panoramic Images
Guangzhou	Haizhu District	100 m	5749	5749
Changsha	Tianxin District	100 m	5389	5389
Beijing	Haidian District	100 m	17,144	17,144
Zhengzhou	Zhongyuan District	100 m	2760	2760
Shenzhen	Luohu District	100 m	3122	3122
Xi'an	Yanta District	100 m	4883	4883
Total	—	—	39,047	39,047

2.2. Semantic Segmentation and Street-Space Indicators

Semantic segmentation was applied to the panoramic street-view imagery to quantify the visible composition of the pedestrian-scale street environment. Because the original Baidu Street View data were stored as 2048 × 1024 equirectangular panoramas, direct pixel-based segmentation could be affected by projection distortion, particularly toward the lateral boundaries of the panorama. Each 360° panorama was therefore computationally reprojected into four horizontal rectilinear views separated by 90°, thereby retaining full horizontal environmental coverage while providing conventional perspective images for semantic analysis.

A pretrained SegFormer semantic segmentation model was then used to classify image pixels into major urban elements, including buildings, vegetation, sky, sidewalks, roadways, vehicles, pedestrians, and other visible street objects. The same projection procedure, model, and inference settings were applied to all images from the six study districts to ensure methodological consistency. For each perspective view, the proportional coverage of semantic class k was calculated as:

$$P_{ijk} = \frac{N_{ijk}}{N_{ij}}, \tag{1}$$

where N_{ijk} represents the number of pixels classified as semantic class k in view i at sampling point j, and N_{ij} represents the total number of valid pixels in that view. The four directional results were subsequently averaged to obtain sampling-point-level indicators:

$$\overline{P}_{jk} = \frac{1}{4}\sum_{i=1}^{4} P_{ijk}, \tag{2}$$

thereby reducing directional dependence and representing the overall visible environment surrounding each sampling location.

Closely related semantic classes were further combined into interpretable street-space indicators. Pixels classified as cars, buses, trucks, and motorcycles were aggregated to calculate the visible motor-vehicle proportion (TRA), whereas person and bicycle pixels were combined to calculate the visible person-and-bicycle proportion (HA). The construction of the street-space indicators followed an image-based approach to characterizing observable components of the micro-scale built environment. Related research by Yan et al. [47] demonstrated the application of street-view-derived environmental indicators alongside other spatial data within an interpretable machine-learning framework for analyzing historic urban districts. In the present study, eight image-derived indicators were retained: vegetation proportion (VEG), visible sky proportion (SKY), visible building proportion (BLD), visible sidewalk proportion (SW), visible roadway proportion (RD), visible motor-vehicle proportion (TRA), visible person-and-bicycle proportion (HA), and semantic-class diversity

(SVC) (Table 2). VEG, SKY, BLD, SW, RD, TRA, and HA represent the proportions of valid image pixels assigned to their corresponding semantic categories, whereas SVC summarizes the entropy of the predicted semantic-class distribution. These indicators characterize visible scene composition rather than directly measuring spatial enclosure, walkability, traffic intensity, actual human activity, or holistic perceptual complexity. Their interpretation is therefore restricted to the image-based characteristics captured by semantic segmentation.

Table 2. Street-space indicators derived from semantic segmentation.

Indicator	Abbreviation	Calculation	Interpretation
Vegetation proportion	VEG	Vegetation pixels/total valid pixels	Proportion of the image occupied by visible vegetation
Visible sky proportion	SKY	Sky pixels/total valid pixels	Proportion of the image occupied by visible sky, rather than a direct measure of spatial openness
Visible building proportion	BLD	Building pixels/total valid pixels	Proportion of the image occupied by building surfaces, rather than a direct measure of spatial enclosure
Visible sidewalk proportion	SW	Sidewalk pixels/total valid pixels	Proportion of the image occupied by visible sidewalks, without directly measuring walkability or sidewalk quality
Visible roadway proportion	RD	Road pixels/total valid pixels	Proportion of the image occupied by roadway surfaces
Visible motor-vehicle proportion	TRA	Car + bus + truck + motorcycle pixels/ total valid pixels	Proportion of the image occupied by motor vehicles, rather than actual traffic volume or exposure
Visible person-and-bicycle proportion	HA	Person + bicycle pixels/total valid pixels	Proportion of the image occupied by persons and bicycles, rather than actual human activity intensity
Semantic-class diversity (entropy)	SVC	Shannon entropy of semantic-class proportions	Diversity and evenness of predicted semantic categories, rather than holistic visual or architectural complexity

Notes: All proportional indicators are derived from the semantic classification of image pixels and aggregated across four directional views. SVC is calculated from the predicted semantic-class proportions. The original variable abbreviations are retained for consistency with the analytical datasets, figures, and model outputs. These indicators describe observable image composition and should not be interpreted as direct measurements of environmental functions, experiential qualities, or causal mechanisms.

Semantic-class diversity (SVC) was calculated using Shannon entropy:

$$SVC_j = -\sum_{k=1}^{K} p_{jk} \ln\left(p_{jk}\right), \quad (3)$$

where p_{jk} represents the proportional coverage of semantic class k at sampling point j, and K is the total number of semantic classes included in the calculation. Higher SVC values indicate greater entropy in the predicted semantic-class distribution, rather than necessarily greater architectural or perceptual complexity.

Before statistical modeling, the derived indicators were examined for missing values, extreme observations, pairwise correlations, and variance inflation factors (VIFs) to identify potential multicollinearity. The final street-space indicators were linked to the human and VLM perception scores through the sampling-point identifiers and subsequently used as explanatory variables in the XGBoost and SHAP analyses.

2.3. Human and VLM Evaluation

A comparative human–AI visual evaluation was conducted to examine how architecture-related educational backgrounds were associated with evaluations of building frontages and urban streetscapes and to determine the extent to which contemporary VLM assessments aligned with human judgments. A total of 200 third- and fourth-year undergraduate students from the same university participated in the experiment. The architecture/urban-planning group (AUP) consisted of 100 students majoring in architecture or urban–rural planning, whereas the non-architecture/planning group (NAP) consisted of 100 students majoring in education, sociology, economics, or management, without formal architectural or planning training. Each group included 50 male and 50 female participants. All participants were aged 18 years or older and had normal or corrected-to-normal vision. Recruiting participants from the same institution and educational stages, with identical gender distributions, improved the comparability of the two groups, although differences in individual characteristics and prior experience could not be entirely excluded. The human-evaluation protocol was approved by the Institutional Review Board of School of Architecture, Changsha University of Science and Technology (Approval No. 20260902; 2 September 2026). The analyses reported here used aggregated image-level and group-level ratings without identifying individual participants.

Two complementary visual evaluation frameworks were applied. Building façade images were evaluated in terms of compositional coherence (CC), proportion and scale (PS), material and color coherence (MC), façade articulation and depth (AD), architectural identity (ID), and overall design quality (ODQ). Streetscape stimuli were evaluated in terms of pleasantness (PL), perceived environmental safety (SAF), pedestrian comfort (PC), building–street integration (BSI), and overall streetscape quality (OSQ). All eleven evaluation dimensions were rated using a common seven-point scale ranging from 1 (very poor) to 7 (excellent). These endpoints served as general evaluative anchors rather than dimension-specific verbal labels. Higher scores indicated more favorable assessments of the corresponding attribute, including greater pleasantness (PL), higher perceived environmental safety (SAF), greater perceived suitability for walking and staying (PC), stronger building–street integration (BSI), and higher overall streetscape quality (OSQ). In particular, SAF referred to perceived safety inferred from visible physical and spatial characteristics rather than actual crime risk or demographic characteristics of visible pedestrians. The operational definitions of all dimensions are provided in Table 3.

Because the complete panoramic database contained 39,047 sampling points, a controlled subset of 120 paired architectural-frontage and streetscape stimuli was used for the formal evaluation experiment. To ensure balanced geographical coverage, each of the six districts was divided into 20 spatial clusters, with one sampling location selected from each cluster, yielding 20 pairs per district. This spatially distributed sampling strategy prevented districts with larger panorama inventories from dominating the evaluation sample. Across the final 120 locations, the median within-district nearest-neighbor distance was

1.46 km, although one pair of locations was separated by approximately 85 m. The spatial clusters were used to distribute the experimental stimuli geographically and did not serve as grouping units in the subsequent predictive cross-validation. However, no formal stratification by building typology or street configuration was implemented, and the selected stimuli cannot be assumed to proportionally represent the full morphological diversity of each district. Each pair consisted of a building-oriented view and its corresponding panoramic streetscape image from the same location. The two evaluation tasks employed different visual representations and domain-specific rating dimensions and were therefore treated as complementary rather than interchangeable measures. A balanced incomplete-block design was adopted so that each participant evaluated 20 architectural-frontage images and 20 panoramic streetscape stimuli. Images were randomized within separate evaluation blocks to reduce contextual carry-over and participant fatigue. Mean ratings were subsequently calculated separately for the AUP and NAP groups to obtain image-level human consensus scores.

Table 3. Human and VLM visual evaluation framework.

Evaluation Level	Dimension	Abbreviation	Definition
Building façade	Compositional coherence	CC	Overall coherence of rhythm, balance, alignment, hierarchy, and façade organization
	Proportion and scale	PS	Quality of façade proportions, openings, solid–void relationships, and perceived scale
	Material and color coherence	MC	Visual compatibility of façade materials, textures, and colors
	Façade articulation and depth	AD	Expression of layering, recesses, projections, openings, shadows, and façade depth
	Architectural identity	ID	Clarity, distinctiveness, and consistency of the architectural design language
	Overall design quality	ODQ	Overall perceived architectural quality of the façade
Streetscape	Pleasantness	PL	Overall visual pleasantness and positive environmental impression
	Perceived environmental safety	SAF	Perceived safety derived from visible physical and spatial characteristics
	Pedestrian comfort	PC	Perceived suitability of the environment for walking and staying
	Building–street integration	BSI	Degree of visual and spatial coordination between buildings and the street environment

Table 3. *Cont.*

Evaluation Level	Dimension	Abbreviation	Definition
Streetscape	Overall streetscape quality	OSQ	Overall perceived quality of the street environment

Five VLMs were evaluated: GPT-6 (recorded model identifier: gpt-6-astra), Claude Fable 5 (claude-fable-5), DeepSeek (deepseek-flash), Kimi K3 (kimi-k3), and Qwen3-VL-Plus (qwen3-vl-plus). Each model evaluated the same 120 frozen architectural-frontage images (768 × 768 pixels) and 120 streetscape panoramas (2048 × 1024 pixels) under the recorded Prompt_v1.0 protocol. The standardized evaluation framework comprised a common system-level instruction and two task-specific prompts specifying the evaluation dimensions in Table 3, seven-point scoring criteria, and required JSON output structures. The prompts restricted evaluations to directly observable visual characteristics and prohibited unsupported inferences about building identity, geographical location, socioeconomic status, and pedestrian characteristics. The archived API implementations for Claude, DeepSeek, Kimi, and Qwen loaded the same prompt files and processed each image as an independent request. Model responses were parsed and validated to ensure that all six architectural or five streetscape scores were integers between 1 and 7. Transient API failures were handled through bounded retries, while invalid responses and unsuccessful requests were recorded separately and excluded from the final analytical datasets. All ten formal Run-1 datasets contained 120 unique valid stimulus records. Model-specific inference configurations are summarized in Supplementary Table S3, and the complete original system-level and task-specific prompts are reproduced verbatim in Supplementary Text S1.

2.4. Statistical and Explainable Modeling

All human ratings were aggregated at the image level before statistical analysis. For image i, evaluation dimension d, and participant group g, the human consensus score was calculated as:

$$HCS_{i,d}^{(g)} = \frac{1}{n_{i,g}} \sum_{r=1}^{n_{i,g}} y_{i,d,r}, \tag{4}$$

where $g \in \{A, UPNAP\}$, $n_{i,g}$ denotes the number of valid respondents from group g who evaluated image i, and $y_{i,d,r}$ is the score provided by participant r. Human consensus scores were calculated independently for the façade and streetscape evaluation dimensions.

For the explainable modeling analysis, a VLM consensus score was calculated as:

$$HVCS_{i,d} = \frac{1}{M} \sum_{m=1}^{M} s_{i,d,m}, \; M = 5, \tag{5}$$

where $s_{i,d,m}$ represents the score assigned by VLM m. Individual VLM scores were retained for model-specific human–AI comparisons, whereas *VCS* was used primarily for comparisons of street-space feature–rating relationships between AI and the two human groups.

Human–AI agreement was assessed separately for each evaluation dimension and VLM using Spearman's rank correlation coefficient (ρ), Lin's concordance correlation coefficient (CCC), and mean absolute error (MAE). Spearman's coefficient measured consistency in the ranking of visual stimuli:

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}, \tag{6}$$

where d_i denotes the difference between the paired ranks and N is the number of evaluated stimuli. Because rank correlation does not necessarily indicate agreement in absolute scores, CCC was additionally calculated:

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + \left(\mu_x - \mu_y\right)^2}, \tag{7}$$

where ρ_{xy} is the Pearson correlation between two sets of scores, μ_x and μ_y are their respective means, and σ_x^2 and σ_y^2 are their variances. Scoring deviation was quantified using:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \mid \hat{y}_i - y_i \mid, \tag{8}$$

Agreement analyses were conducted between each individual VLM and the AUP and NAP consensus scores, as well as between the two human groups themselves. This human-referenced comparison is consistent with the general validation principle illustrated by Qin et al. [33], who compared AI-assisted perceptual scores with human assessments using matched visual and auditory samples. In the present study, Spearman's rank correlation was used to evaluate whether VLMs reproduced the relative ordering of images established by human evaluators, whereas CCC and MAE were used to assess numerical agreement and scoring deviations. The three metrics therefore provide complementary evidence rather than interchangeable measures of alignment. Human consensus ratings were treated as group-specific reference judgments, rather than an objective ground truth for architectural or streetscape quality.

To quantify estimation uncertainty, pointwise 95% confidence intervals (CIs) were calculated using 2000 paired bootstrap resamples of the 120 image-level observations for each evaluation dimension. The corresponding human and VLM scores were resampled jointly, and Spearman's ρ, CCC, MAE, and paired rating differences were recalculated for each resample. The 2.5th and 97.5th percentiles of the bootstrap distributions were used as confidence limits. These intervals reflect uncertainty across the evaluated stimuli conditional on the aggregated human ratings; they do not account for participant-level resampling or alternative stimulus selection within cities.

To investigate predictive associations between image-derived semantic indicators and streetscape ratings, XGBoost regression models were developed using the semantic indicators derived in Section 2.2 as predictor variables. Separate models were constructed for PL, SAF, PC, BSI, and OSQ for the AUP, NAP, and VLM-consensus evaluations. The prediction function was expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \tag{9}$$

where x_i is the vector of street-space indicators at sampling point i, f_k denotes an individual regression tree, and K is the number of trees. Model performance was evaluated using $\boldsymbol{R}^2$, RMSE, and MAE.

XGBoost models were fitted using a squared-error objective and histogram-based tree construction. Hyperparameters were optimized through randomized search over eight candidate configurations, using three-fold inner cross-validation and negative root mean squared error (RMSE) as the selection criterion. Predictive performance was evaluated using three repetitions of five-fold outer cross-validation stratified by city, yielding 15 held-out test folds, each containing four observations from each of the six cities. Because the folds were stratified by city rather than grouped by street segment or spatial block, geographically nearby observations could occur in different training and test subsets, and residual within-city spatial dependence cannot be excluded. Two simple benchmarks, a training-set mean predictor and a city-specific training-set mean predictor, were evaluated using the same

outer folds. Cross-validation results were summarized using R^2, RMSE, and MAE. Cross-city transferability was separately assessed through leave-one-city-out (LOCO) validation, with each held-out city excluded from model training and hyperparameter tuning.

The fitted XGBoost models were interpreted using SHapley Additive exPlanations (SHAP):

$$f(x) = \phi_0 + \sum_{j=1}^{P} \phi_j, \tag{10}$$

where ϕ_0 represents the expected model prediction and ϕ_j is the contribution of predictor j. Global feature importance was calculated using:

$$I_j = \frac{1}{N} \sum_{i=1}^{N} | \phi_{ij} |, \tag{11}$$

SHAP summary and dependence plots were used to characterize the model-specific contributions and potential nonlinear associations of the eight image-derived semantic indicators defined in Table 2 with predicted streetscape ratings. Finally, the SHAP importance rankings obtained from the AUP, NAP, and VLM-consensus models were compared using Spearman correlations to determine whether AI and human groups exhibited similar street-space feature–rating relationships. The combination of image-derived built-environment indicators and XGBoost–SHAP follows a broader methodological approach also illustrated by Yan et al. [47], who used multisource spatial data and interpretable machine learning to examine micro-scale urban functional patterns. In the present analysis, SHAP values were interpreted as feature attributions within the fitted prediction models rather than estimates of independent causal effects. Their interpretation was considered in conjunction with out-of-sample predictive performance, while potential dependence among semantic pixel proportions was explicitly acknowledged. Accordingly, feature-importance comparisons involving models with negative out-of-sample predictive performance were treated as exploratory rather than as reliable evidence of differences in human perceptual mechanisms. All statistical tests were two-sided at $p < 0.05$, and the Benjamini–Hochberg false discovery rate procedure was applied where multiple comparisons were performed.

To assess the stability of SHAP feature attributions, a city-stratified bootstrap sensitivity analysis was conducted using 100 resamples for each of the 15 XGBoost models. Each resample retained 120 observations by sampling 20 locations with replacement from each of the six cities. Models were refitted using their originally selected hyperparameters, and TreeSHAP importance was recalculated on the fixed set of 120 original observations. Stability was evaluated using Spearman correlations between the bootstrap-derived and original eight-feature importance rankings, the retention frequency of the original top-ranked feature, and the overlap of the three highest-ranked features.

3. Results

3.1. Evaluation Profiles and Human Consensus Reliability

The final evaluation dataset comprised 120 paired architectural-frontage and streetscape stimuli. Each stimulus was evaluated by 16–17 participants within both the architecture/urban-planning (AUP) and non-architecture/planning (NAP) groups, resulting in 2000 valid architectural ratings and 2000 valid streetscape ratings per group. All five VLMs completed the full set of 120 architectural and 120 streetscape evaluations, with no missing stimuli in the formal Run-1 dataset.

Human evaluations showed highly stable image-level consensus across both tasks (Figure 2d). For the AUP group, Spearman–Brown-corrected split-half reliability ranged from 0.924 to 0.945 across the 11 evaluation dimensions, whereas the corresponding range for the NAP group was 0.878–0.903. Thus, although agreement was consistently higher

among architecture- and planning-trained participants, both groups produced sufficiently stable image-level consensus scores for subsequent human–AI comparisons.

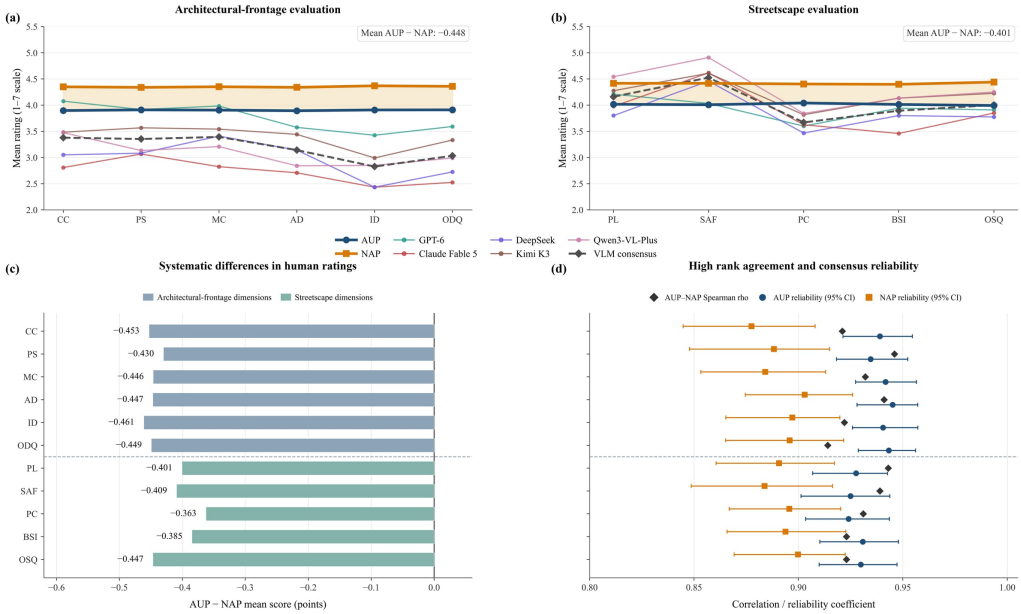

Figure 2. Human and VLM evaluation profiles, intergroup rating differences, and human consensus reliability. (**a**,**b**) Mean architectural-frontage and streetscape ratings for architecture/urban-planning students (AUP), students without architecture/planning training (NAP), five VLMs, and the VLM consensus. Shaded areas highlight the AUP–NAP rating gaps, with annotations indicating the dimension-averaged differences. (**c**) Dimension-specific AUP–NAP mean-rating differences; negative values indicate lower AUP scores. (**d**) Image-level AUP–NAP Spearman rank correlations (black diamonds) and Spearman–Brown-corrected split-half reliabilities for AUP and NAP. Horizontal error bars represent 95% intervals from 500 repeated split-half resamples for the reliability estimates. The dashed horizontal line separates architectural and streetscape dimensions. All ratings used a seven-point scale; panels (**a**,**b**) display a restricted vertical range for readability. Evaluation-dimension abbreviations are defined in Table 3.

The two human groups nevertheless differed systematically in rating level (Figure 2a,b; Table 4). For architectural-frontage evaluation, AUP mean scores were tightly clustered around the midpoint of the scale, ranging from 3.893 to 3.909, whereas NAP scores ranged from 4.339 to 4.369. The largest between-group difference occurred for architectural identity (ID; $\Delta_{AUP-NAP} = -0.461$), followed by compositional coherence (CC; −0.453) and overall design quality (ODQ; −0.449). All six architectural dimensions showed significant paired differences after Benjamini–Hochberg correction ($p_{BH} < 0.001$). Despite this systematic level difference, the ranking of the 120 architectural stimuli was highly consistent between the two groups, with Spearman correlations ranging from 0.914 to 0.946, CCC values from 0.754 to 0.770, and MAE values from 0.472 to 0.511.

A similar pattern was observed for streetscape perception. AUP mean scores ranged from 3.993 to 4.041, whereas NAP means ranged from 4.397 to 4.440. NAP participants therefore assigned consistently higher ratings across pleasantness (PL), perceived environmental safety (SAF), pedestrian comfort (PC), building–street integration (BSI), and overall streetscape quality (OSQ). The group difference ranged from −0.363 for PC to −0.447 for OSQ, and all five paired comparisons remained significant after FDR correction ($p_{BH} < 0.001$). Nevertheless, image-level rankings again remained strongly aligned, with Spearman $\rho = 0.923$–0.943, CCC = 0.771–0.807, and MAE = 0.416–0.486. These results indicate that the two human groups largely agreed on the relative ordering of environments while using systematically different locations on the seven-point rating scale.

Table 4. Human evaluation profiles, consensus reliability, and AUP–NAP agreement.

Level	Dimension	AUP Mean ± SD	NAP Mean ± SD	Reliability AUP	Reliability NAP	AUP–NAP ρ	CCC	MAE	$\Delta_{AUP-NAP}$
Architectural	CC	3.897 ± 0.943	4.350 ± 0.661	0.939	0.878	0.921	0.755	0.494	−0.453
Architectural	PS	3.909 ± 0.937	4.339 ± 0.638	0.935	0.888	0.946	0.770	0.472	−0.430
Architectural	MC	3.905 ± 0.948	4.351 ± 0.655	0.942	0.884	0.932	0.760	0.500	−0.446
Architectural	AD	3.893 ± 0.975	4.340 ± 0.675	0.945	0.903	0.941	0.768	0.511	−0.447
Architectural	ID	3.908 ± 0.934	4.369 ± 0.672	0.941	0.897	0.922	0.754	0.498	−0.461
Architectural	ODQ	3.909 ± 0.950	4.359 ± 0.666	0.943	0.896	0.914	0.754	0.506	−0.449
Streetscape	PL	4.016 ± 0.849	4.417 ± 0.662	0.928	0.891	0.943	0.807	0.429	−0.401
Streetscape	SAF	4.007 ± 0.848	4.417 ± 0.656	0.925	0.884	0.939	0.786	0.438	−0.409
Streetscape	PC	4.041 ± 0.849	4.403 ± 0.655	0.924	0.896	0.931	0.807	0.417	−0.363
Streetscape	BSI	4.013 ± 0.856	4.397 ± 0.664	0.931	0.894	0.923	0.799	0.416	−0.385
Streetscape	OSQ	3.993 ± 0.878	4.440 ± 0.682	0.930	0.900	0.923	0.771	0.486	−0.447

Notes: Reliability refers to the median Spearman–Brown-corrected split-half reliability across 500 repeated random splits. $\Delta_{AUP-NAP}$ denotes the paired mean difference. All AUP–NAP mean differences remained significant after Benjamini–Hochberg correction ($p_{BH} < 0.001$).

The five VLMs exhibited substantially greater variation in their absolute scoring profiles (Figure 2a,b). For architectural evaluation, the mean VLM-consensus score ranged from 2.827 for ID to 3.393 for MC, remaining below the corresponding AUP and NAP means for every architectural dimension. Claude Fable 5 and DeepSeek showed the lowest architectural scores, particularly for ID and ODQ, whereas GPT-6 produced the highest mean ratings for most architectural attributes. Kimi K3 occupied an intermediate position and showed the smallest overall absolute deviation from the human groups in later agreement analyses. In contrast, VLM streetscape ratings were closer in scale to human evaluations, but clear dimension-specific shifts remained. In particular, the VLM consensus was relatively high for SAF (4.525) and lower for PC (3.670), while the two human groups showed much flatter profiles across the five streetscape dimensions.

The most notable pattern was therefore a separation between preference ranking and rating level. Across all eleven evaluation dimensions, AUP participants assigned systematically lower scores than NAP participants, whereas the two groups retained very high agreement on which architectural and streetscape stimuli ranked more favorably. The results indicate a consistent group-level scoring difference without a corresponding large-scale divergence in image rankings. This distinction provides the human reference framework for interpreting the subsequent Human–AI comparisons. The corresponding pointwise 95% bootstrap confidence intervals for AUP–NAP rating differences, Spearman rank correlations, CCC, and MAE are provided in Supplementary Table S4, complementing the significance tests and point estimates reported in Table 4.

3.2. Human–AI Agreement in Architectural and Streetscape Evaluation

Human–AI agreement differed fundamentally from the strong agreement observed between the two human groups. Across architectural evaluations, none of the five individual VLMs reproduced the human image rankings to a meaningful degree (Figure 3a,b). Across the 60 model–human–dimension comparisons, Spearman correlations ranged from −0.182 to 0.081, and no individual-model correlation remained significant after BH correction. CCC values were similarly close to zero, ranging from −0.147 to 0.042, indicating that the weak correspondence was not limited to a difference in rating scale but extended to absolute agreement. Dimension-specific MAE ranged from 0.986 to 1.990 score points. The corresponding 95% bootstrap confidence intervals for the individual Human–AI agreement estimates are reported in Supplementary Table S4 to indicate their estimation precision. Importantly, the absence of statistically significant rank correlations after multiple-testing correction should not be interpreted as evidence that the underlying

associations are exactly zero; rather, the results indicate weak observed correspondence within the present evaluation sample.

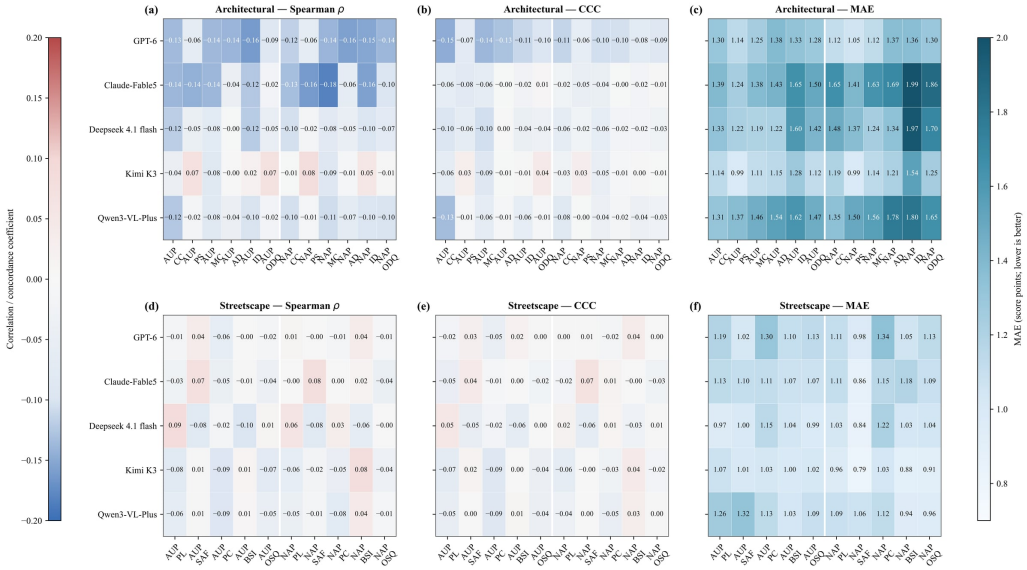

Figure 3. Human–AI agreement across architectural and streetscape evaluations. (**a**–**c**) Spearman rank correlation, Lin's concordance correlation coefficient (CCC), and mean absolute error (MAE) between each VLM and the AUP/NAP architectural-frontage consensus scores. (**d**–**f**) Corresponding agreement metrics for streetscape evaluation. Lower MAE values indicate smaller absolute scoring deviations.

Table 5 summarizes Human–AI agreement across the two evaluation domains. Among the five models, Kimi K3 showed the smallest overall architectural deviation from the human evaluations. Its mean MAE across the six dimensions was 1.131 relative to AUP and 1.221 relative to NAP. GPT-6 produced comparable errors (1.281 and 1.220, respectively) but showed systematic negative rank associations with both human groups. Claude Fable 5 showed the largest negative mean-rating bias, with mean biases of −1.176 points relative to AUP and −1.623 points relative to NAP, and its mean MAE reached 1.432 and 1.707, respectively. DeepSeek and Qwen3-VL-Plus also tended to score architectural quality below both human reference groups. At the dimension level, the lowest architectural MAE was observed for Kimi K3 versus AUP on proportion and scale (PS; MAE = 0.986), whereas the largest error occurred for Claude Fable 5 versus NAP on architectural identity (ID; MAE = 1.990).

Averaging the five VLMs did not eliminate the disagreement. The VLM-consensus architecture score retained near-zero or negative mean rank correspondence with both human groups ($\overline{\rho} = -0.099$ for AUP and −0.112 for NAP), with mean CCC values of −0.063 and −0.041. The corresponding mean MAE values were 1.225 and 1.364, respectively. This indicates that the divergence was not driven by a single anomalous model but was shared across the multimodel ensemble.

Agreement was somewhat closer in absolute-error terms for streetscape evaluation, but rank and concordance measures remained weak (Figure 3d–f). Across the 50 individual model–human comparisons, Spearman ρ ranged from −0.099 to 0.087, CCC from −0.088 to 0.067, and none of the Spearman correlations survived BH correction. MAE values ranged from 0.795 to 1.339, lower overall than those observed for architectural judgments. Kimi K3 showed the lowest mean streetscape MAE relative to NAP (0.917), while the VLM consensus yielded MAE values of 1.011 for AUP and 0.941 for NAP. DeepSeek also showed comparatively low mean absolute errors (1.031 for both human groups).

The streetscape results further revealed dimension-specific rating-level differences. Safety ratings assigned by Claude Fable 5, DeepSeek, Kimi K3, and Qwen3-VL-Plus were generally higher than their scores for pedestrian comfort, whereas AUP ratings remained close to 4.0 across all five dimensions. Qwen3-VL-Plus produced the highest mean SAF score (4.908), exceeding both AUP (4.007) and NAP (4.417), while DeepSeek assigned the lowest mean PC score (3.467). The smallest dimension-specific human–AI error was obtained for Kimi K3 versus NAP on SAF (MAE = 0.795), whereas the largest streetscape error occurred for GPT-6 versus NAP on PC (MAE = 1.339).

An important contrast emerged when agreement among the VLMs themselves was examined. Despite their weak correspondence with human consensus, the five models exhibited moderate-to-strong pairwise agreement with one another. For architectural evaluation, pairwise VLM Spearman correlations ranged from 0.516 to 0.863, with CCC values of 0.355–0.767; for streetscape evaluation, corresponding ranges were 0.533–0.832 and 0.490–0.801, respectively. All pairwise VLM rank correlations remained significant after BH correction. The strongest aggregate architectural relationship was observed between GPT-6 and Qwen3-VL-Plus, whereas the highest overall streetscape correspondence was observed among combinations involving GPT-6, Kimi K3, and Qwen3-VL-Plus. Thus, the VLMs showed a relatively coherent internal ordering of the visual stimuli, but this machine-based ordering differed markedly from that produced by both human groups.

Table 5. Overall Human–AI agreement across evaluation domains.

Domain	Model	Human Group	Mean ρ	Mean CCC	Mean MAE	Mean Bias
Architectural	GPT-6	AUP	−0.120	−0.116	1.281	−0.142
	GPT-6	NAP	−0.128	−0.090	1.220	−0.590
	Claude Fable 5	AUP	−0.101	−0.036	1.432	−1.176
	Claude Fable 5	NAP	−0.132	−0.024	1.707	−1.623
	DeepSeek	AUP	−0.069	−0.057	1.330	−0.930
	DeepSeek	NAP	−0.069	−0.034	1.517	−1.378
	Kimi K3	AUP	0.006	−0.015	1.131	−0.510
	Kimi K3	NAP	0.003	−0.010	1.221	−0.958
	Qwen3-VL-Plus	AUP	−0.065	−0.047	1.465	−0.820
	Qwen3-VL-Plus	NAP	−0.082	−0.035	1.605	−1.268
	VLM consensus	AUP	−0.099	−0.063	1.225	−0.716
	VLM consensus	NAP	−0.112	−0.041	1.364	−1.163
Streetscape	GPT-6	AUP	−0.010	−0.004	1.146	−0.076
	GPT-6	NAP	0.005	0.006	1.123	−0.476
	Claude Fable 5	AUP	−0.009	−0.008	1.098	−0.109
	Claude Fable 5	NAP	0.010	0.007	1.078	−0.510
	DeepSeek	AUP	−0.020	−0.016	1.031	−0.154
	DeepSeek	NAP	−0.009	−0.011	1.031	−0.555
	Kimi K3	AUP	−0.046	−0.036	1.025	0.198
	Kimi K3	NAP	−0.017	−0.013	0.917	−0.203
	Qwen3-VL-Plus	AUP	−0.039	−0.034	1.168	0.321
	Qwen3-VL-Plus	NAP	−0.020	−0.010	1.034	−0.080
	VLM consensus	AUP	−0.032	−0.022	1.011	0.036
	VLM consensus	NAP	−0.014	−0.004	0.941	−0.365

Mean ρ is the Fisher-z-averaged Spearman rank correlation across the dimensions within each domain. Bias is calculated as VLM minus human consensus; negative values indicate lower VLM ratings. Dimension-specific Human–AI Spearman tests for the five individual models did not remain significant after Benjamini–Hochberg correction.

3.3. Street-Space Features Associated with Perceptual Evaluation

The 120 frozen panoramic streetscape stimuli were reprojected into four horizontal rectilinear views separated by 90°, producing a total of 480 perspective images for semantic segmentation. A SegFormer-B5 model fine-tuned on Cityscapes was used to identify the principal visible street-space components, and class proportions were averaged across the four directional views to obtain sampling-point-level indicators. Figure 4 illustrates representative segmentation results from the six study cities. A qualitative visual audit covering 30 pre-specified perspective views from the six study cities is presented in Figure S2. Although this inspection provided a preliminary assessment of the segmentation outputs, no manually annotated pixel-level reference masks were available to calculate category-specific intersection over union (IoU), precision, or recall. Consequently, the visual audit should not be interpreted as quantitative evidence of segmentation accuracy, particularly for relatively small or partially obscured categories such as sidewalks, pedestrians, and bicycles. The resulting point-level dataset contained the eight pre-specified image-derived indicators defined in Table 2 (VEG, SKY, BLD, SW, RD, TRA, HA, and SVC).

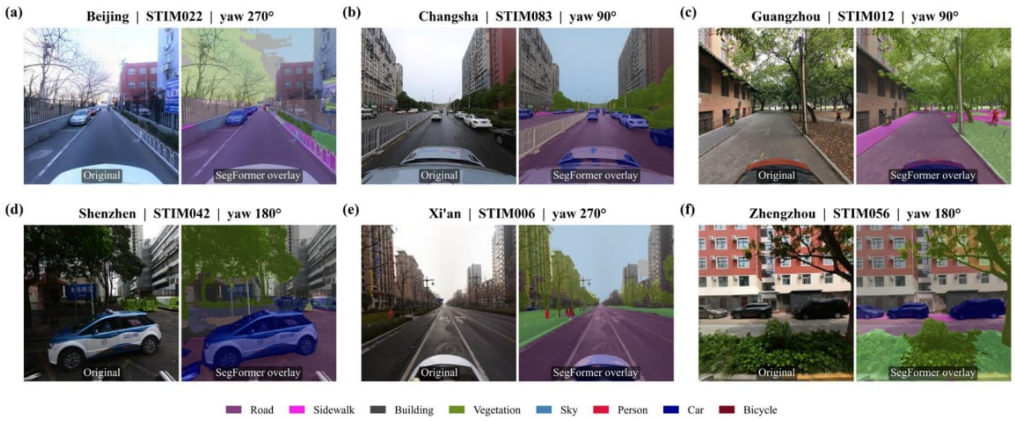

Figure 4. Representative SegFormer semantic segmentation across the six study cities. The panels show (**a**) Beijing, (**b**) Changsha, (**c**) Guangzhou, (**d**) Shenzhen, (**e**) Xi'an, and (**f**) Zhengzhou. For each city, a representative rectilinear perspective view is shown alongside its SegFormer semantic overlay. Cases were selected from the frozen analytical sample using a reproducible feature-based selection procedure. The overlays illustrate the major scene components subsequently used to derive the street-space indicators, including roads, sidewalks, buildings, vegetation, sky, pedestrians, vehicles, and bicycles.

Substantial variation was retained across the final 120 streetscapes. VEG was the largest of the focal proportional indicators ($M = 0.275$, $SD = 0.133$), followed by BLD ($M = 0.235$, $SD = 0.139$) and RD ($M = 0.226$, $SD = 0.074$). TRA averaged 0.075, whereas SW and HA were comparatively sparse, with means of 0.028 and 0.011, respectively. SKY averaged 0.060, and SVC had a mean Shannon entropy of 1.645. The sample also covered broad environmental ranges, including VEG from approximately 0 to 0.637, BLD from 0.001 to 0.626, and TRA from 0.001 to 0.311. Pairwise correlations were generally moderate (Figure S1), with the strongest association observed between VEG and BLD ($r = -0.74$); no pair exceeded the pre-specified $|r| = 0.80$ screening threshold. VIF values were nevertheless elevated for VEG (13.05) and BLD (12.95), reflecting their compositional interdependence within the finite visual field. Because the two indicators represent substantively distinct streetscape elements and the subsequent models were tree-based, both were retained, while SHAP attributions involving these correlated components were interpreted cautiously.

The predictive performance of the XGBoost models differed sharply across evaluator sources (Table 6; Figure 5a). For the AUP group, all five repeated cross-validated R^2 values

were negative. The highest value was obtained for BSI ($R^2 = -0.143$), with RMSE = 0.890 and MAE = 0.733; the remaining outcomes ranged from $R^2 = -0.165$ to -0.194. Cross-city generalizability was similarly weak, with pooled leave-one-city-out (LOCO) R^2 values remaining negative for every AUP outcome. The NAP models showed the same overall pattern: repeated-CV R^2 ranged from -0.178 for PL to -0.253 for PC, and all five LOCO R^2 values were also negative. In contrast, the VLM-consensus models produced positive out-of-sample R^2 values for all five perceptual outcomes. The strongest repeated-CV performance was observed for PL ($R^2 = 0.277$), followed by PC ($R^2 = 0.242$), OSQ ($R^2 = 0.148$), SAF ($R^2 = 0.135$), and BSI ($R^2 = 0.025$). This pattern persisted under LOCO validation, with positive cross-city R^2 values for PL (0.213), PC (0.237), SAF (0.112), OSQ (0.091), and BSI (0.006). Thus, the eight semantic street-space indicators captured a modest but transferable portion of the variation in VLM-derived perceptual scores, whereas they did not reproduce the highly reliable human consensus scores observed in Section 3.1.

Table 6. Out-of-sample predictive performance of the XGBoost models for AUP, NAP, and VLM-consensus streetscape evaluations.

Outcome	Evaluator	CV R^2	CV RMSE	CV MAE	LOCO R^2	LOCO RMSE	LOCO MAE
PL	AUP	−0.172 ± 0.280	0.886 ± 0.105	0.726 ± 0.097	−0.202	0.927	0.751
PL	NAP	−0.178 ± 0.157	0.699 ± 0.063	0.576 ± 0.059	−0.254	0.738	0.611
PL	VLM consensus	0.277 ± 0.158	0.784 ± 0.134	0.613 ± 0.104	0.213	0.834	0.643
SAF	AUP	−0.194 ± 0.193	0.902 ± 0.107	0.752 ± 0.099	−0.301	0.963	0.781
SAF	NAP	−0.212 ± 0.285	0.694 ± 0.076	0.568 ± 0.073	−0.249	0.730	0.597
SAF	VLM consensus	0.135 ± 0.195	0.678 ± 0.123	0.514 ± 0.092	0.112	0.709	0.536
PC	AUP	−0.194 ± 0.165	0.904 ± 0.102	0.745 ± 0.088	−0.265	0.951	0.774
PC	NAP	−0.253 ± 0.239	0.711 ± 0.075	0.588 ± 0.067	−0.291	0.742	0.610
PC	VLM consensus	0.242 ± 0.197	0.757 ± 0.106	0.595 ± 0.099	0.237	0.778	0.612
BSI	AUP	−0.143 ± 0.203	0.890 ± 0.081	0.733 ± 0.086	−0.242	0.950	0.791
BSI	NAP	−0.191 ± 0.141	0.706 ± 0.061	0.588 ± 0.056	−0.269	0.745	0.616
BSI	VLM consensus	0.025 ± 0.175	0.773 ± 0.105	0.616 ± 0.079	0.006	0.799	0.622
OSQ	AUP	−0.165 ± 0.133	0.925 ± 0.098	0.758 ± 0.095	−0.242	0.975	0.790
OSQ	NAP	−0.212 ± 0.145	0.731 ± 0.081	0.591 ± 0.070	−0.347	0.788	0.640
OSQ	VLM consensus	0.148 ± 0.169	0.752 ± 0.115	0.584 ± 0.085	0.091	0.796	0.605

Notes: CV values represent mean ± SD across 15 outer test folds from three repeated five-fold cross-validation runs stratified by city. LOCO denotes pooled leave-one-city-out evaluation, with all observations from the held-out city excluded from model training and tuning. Negative R^2 values indicate predictive performance below the corresponding test-set mean baseline.

The baseline comparisons further qualified these findings (Table S5). Across all five streetscape outcomes, the XGBoost models for AUP and NAP exhibited higher cross-validated RMSE than both training-mean benchmarks, indicating that the eight semantic indicators provided no predictive improvement over these simple alternatives for either human group. In contrast, the VLM-consensus models outperformed both benchmarks across all five outcomes. For example, for pleasantness, XGBoost achieved an RMSE of 0.784 compared with 0.940 for the overall training-mean predictor, whereas the corresponding AUP values were 0.886 and 0.844. These results reinforce the distinction between the out-of-sample predictability of model-generated ratings and that of human consensus scores.

The SHAP analysis was used to describe feature contributions within the fitted XGBoost models (Figure 5b–d). However, because the AUP and NAP models exhibited negative out-of-sample R^2 values and failed to outperform the mean-prediction benchmarks, their SHAP rankings should be regarded as exploratory model diagnostics rather than reliable explanations of human perceptual mechanisms. The VLM-consensus SHAP results had comparatively stronger predictive support, although their interpretation remains associative

and model-dependent. Among the AUP models, SW contributed most strongly to PL (importance share = 0.251) and SAF (0.178), while PC was most strongly associated with SVC (0.197), closely followed by HA (0.195) and SW (0.194). HA ranked first for BSI (0.180), whereas VEG had the largest contribution to OSQ (0.187). For NAP, HA ranked highest for PL (0.188), VEG was the leading feature for SAF (0.182) and PC (0.190), and SW dominated BSI (0.216) and OSQ (0.198). The VLM-consensus models exhibited a more concentrated attribution pattern: VEG was the dominant predictor for PL (0.333), PC (0.266), and OSQ (0.269), whereas RD contributed most strongly to SAF (0.229) and SKY to BSI (0.215). The supplementary SHAP summary distributions (Figure S3) reinforce this contrast by showing that high VEG values were generally associated with positive contributions to VLM-predicted PL, PC, and OSQ, whereas higher SKY values tended to contribute negatively to several outcomes. Higher RD and SW values showed more outcome-specific positive contributions, particularly for SAF and PC.

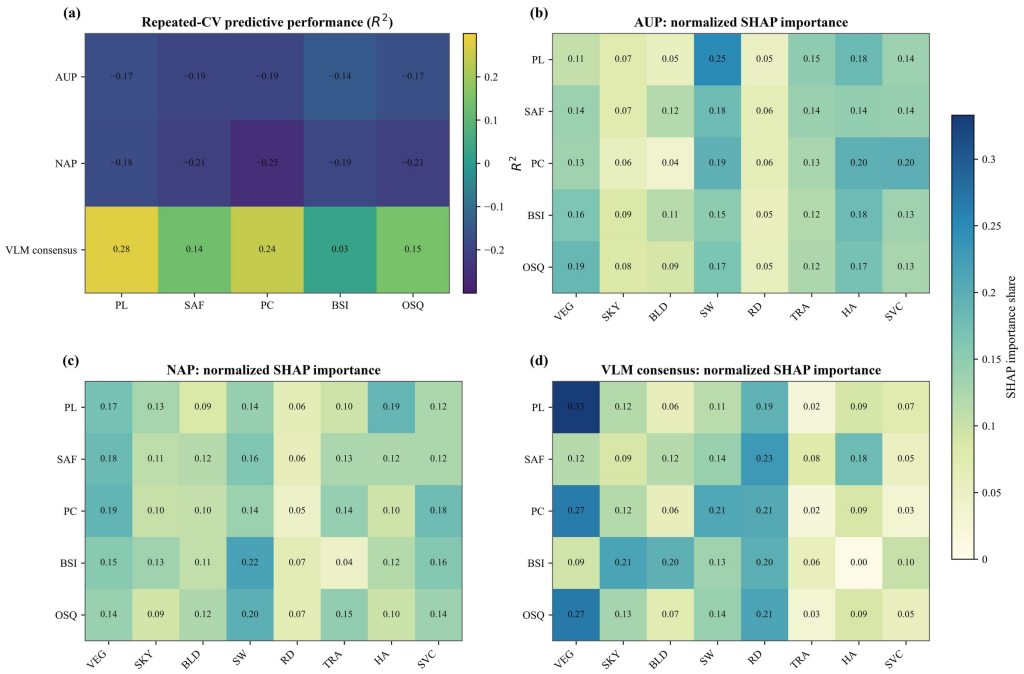

Figure 5. XGBoost predictive performance and normalized SHAP feature importance. (**a**) Mean repeated cross-validated R^2 for the five perceptual outcomes in the AUP, NAP, and VLM-consensus models. Negative values indicate performance below the test-fold mean baseline. (**b**–**d**) Normalized mean absolute SHAP importance for the eight street-space indicators across PL, SAF, PC, BSI, and OSQ for AUP, NAP, and VLM-consensus models, respectively. SHAP importance patterns for the AUP and NAP models should be interpreted as exploratory because the corresponding models exhibited negative out-of-sample predictive performance.

The bootstrap stability analysis revealed substantial differences in the reproducibility of SHAP importance rankings across evaluator sources (Table S6). For AUP and NAP, median Spearman correlations between bootstrap-derived and original feature rankings ranged from 0.345 to 0.524 and from 0.310 to 0.429, respectively. The original top-ranked feature was retained in only 17–35% of AUP resamples and 24–34% of NAP resamples, indicating considerable sensitivity to stimulus resampling. In contrast, the VLM-consensus models showed comparatively higher ranking stability (median ρ = 0.595–0.857), with top-feature retention rates of 38–94%. Nevertheless, stability varied across VLM outcomes, particularly for building–street integration. Together with the negative out-of-sample R^2 values for both human groups, these findings indicate that AUP and NAP SHAP results

should be treated as exploratory model diagnostics rather than reliable explanations of human perceptual mechanisms.

The nonlinear dependence plots in Figure 6 provide additional detail on the four feature–outcome combinations with the highest average SHAP importance across evaluator sources: VEG–PL, SW–PC, HA–PL, and SKY–BSI. For VEG–PL, the VLM-consensus model showed the clearest threshold-like relationship: vegetation below approximately 0.20 generally produced negative SHAP contributions, whereas contributions increased rapidly between roughly 0.20 and 0.45 before approaching a plateau. The AUP and NAP curves were markedly flatter and showed only limited changes over the same range. A similar evaluator-specific pattern was observed for SW–PC. In the VLM-consensus model, sidewalk visibility shifted from negative to positive contribution at approximately SW = 0.03–0.04 and became increasingly positive thereafter, whereas the fitted human-model curves declined slightly at higher sidewalk proportions. Given the negative out-of-sample R^2 of the AUP and NAP models, these human SHAP patterns should be interpreted as characteristics of the fitted models rather than as robust predictive relationships.

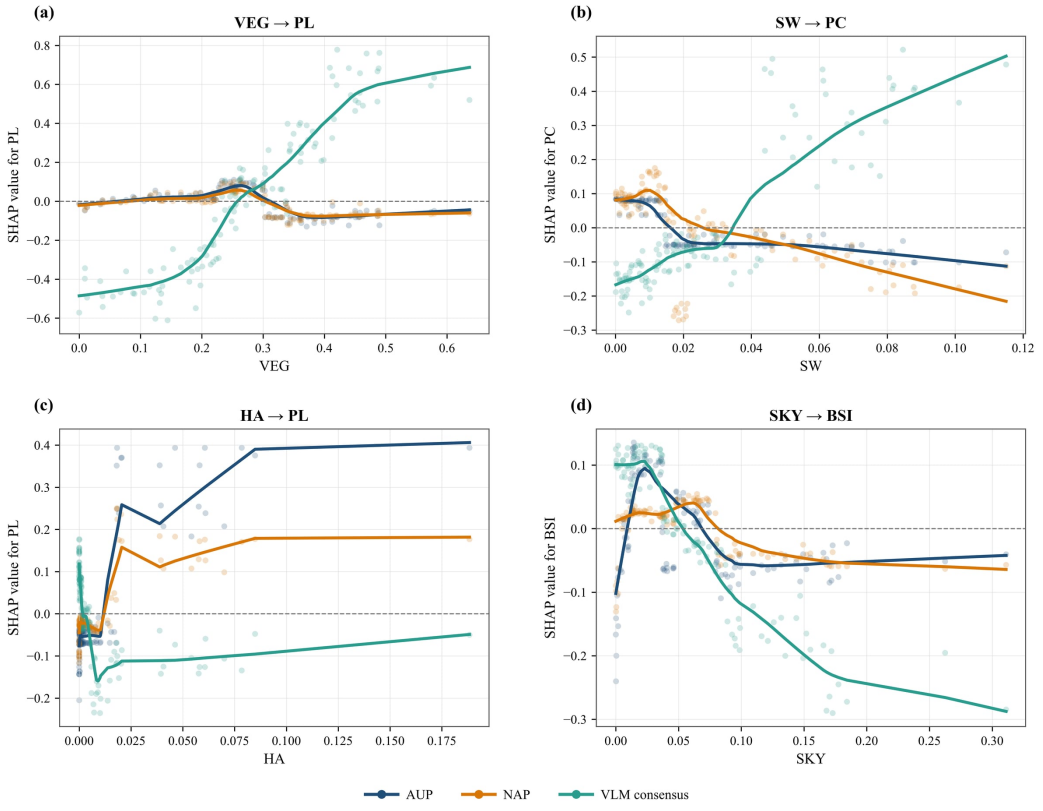

Figure 6. SHAP dependence patterns for the dominant street-space feature–perception relationships. Dependence plots show the relationships between (**a**) vegetation proportion and pleasantness, (**b**) visible sidewalk proportion and pedestrian comfort, (**c**) visible person-and-bicycle proportion and pleasantness, and (**d**) visible sky proportion and building–street integration. Points represent sampling locations and curves show LOWESS-smoothed relationships for AUP, NAP, and VLM-consensus models. Blue, orange, and teal points and curves represent the AUP, NAP, and VLM-consensus models, respectively. SHAP values above zero indicate positive contributions to model predictions. The AUP and NAP dependence curves are exploratory descriptions of fitted model behavior and should not be interpreted as validated perceptual mechanisms.

For HA–PL, the AUP and NAP fitted functions became positive once visible person-and-bicycle proportion increased beyond very low levels, while the VLM-consensus relationship remained predominantly negative over most of the observed range. In the

SKY–BSI relationship, the three evaluator sources displayed a broadly similar transition at low sky proportions but diverged substantially thereafter. The VLM-consensus SHAP contribution decreased progressively as SKY increased, becoming strongly negative at higher levels, whereas the human-model curves exhibited considerably smaller negative contributions. Taken together, these findings indicate that the VLM-consensus ratings were systematically linked to a relatively small set of directly observable semantic scene characteristics, particularly vegetation, visible roadway, sidewalk, and sky proportions. By contrast, the eight compositional indicators alone were insufficient to predict the image-level human consensus scores, despite the strong within-human agreement demonstrated in Section 3.1. Spatial configuration, architectural relationships, and contextual interpretation remain plausible sources of information not captured by these indicators, although the present models cannot establish their relative contributions to human evaluations.

3.4. Cross-Group and Cross-City Heterogeneity

Cross-group and cross-city analyses were conducted to determine whether the differences observed in Sections 3.1–3.3 were concentrated in particular urban contexts or remained stable across the six study cities. Overall, the AUP–NAP rating gap remained highly consistent across cities, whereas Human–AI agreement varied across individual model–city combinations and remained weak in the pooled analysis. City-level variation was limited for the two human groups but more apparent for the VLM-consensus streetscape ratings (Figure 7; Figure S4).

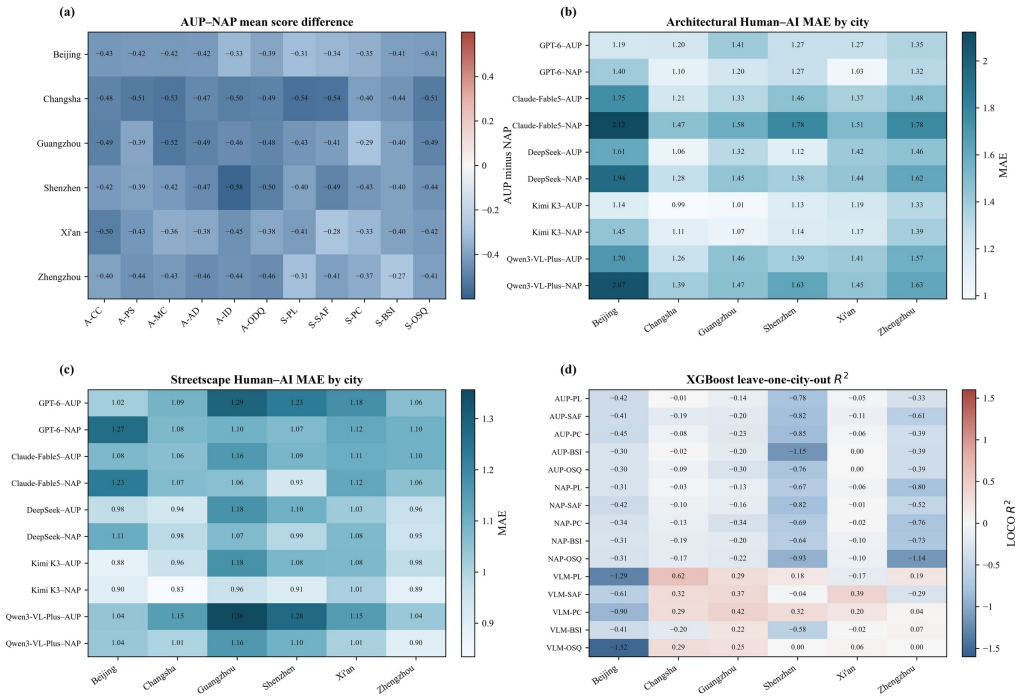

Figure 7. Cross-group and cross-city heterogeneity in Human–AI evaluation and predictive robustness. (**a**) Mean AUP–NAP score differences across the six study cities for architectural and streetscape evaluation dimensions. Negative values indicate lower AUP ratings. (**b**,**c**) City-specific mean absolute error between each VLM and the AUP/NAP consensus scores for architectural-frontage and streetscape evaluation, respectively. (**d**) Leave-one-city-out R^2 of the XGBoost models for AUP, NAP, and VLM-consensus streetscape outcomes; positive values indicate predictive performance exceeding the held-out-city mean baseline.

For the two human groups, city-level heterogeneity was weak. None of the six architectural dimensions showed significant between-city differences for either AUP or NAP after

Benjamini–Hochberg correction, and the median epsilon-squared effect size was effectively zero for both groups. Similarly, none of the five streetscape dimensions showed significant city heterogeneity for AUP or NAP. By contrast, the VLM-consensus ratings exhibited greater spatial variation. Although no architectural dimension remained significant after FDR correction, two of the five streetscape outcomes—pleasantness (PL) and overall streetscape quality (OSQ)—showed significant between-city heterogeneity, with $p_{BH} = 0.031$ for both. Their epsilon-squared effect sizes were 0.112 and 0.107, respectively.

The city-specific VLM-consensus means further illustrate this variation. For PL, the mean VLM score ranged from 3.54 in Beijing to 4.48 in Shenzhen, while OSQ ranged from 3.43 in Beijing to 4.36 in Guangzhou. In comparison, the human city means were much more stable. AUP streetscape means across the five dimensions generally remained close to 4.0, while NAP scores remained consistently higher, approximately in the mid-4 range. The complete city-level profiles are shown in Figure S4.

AUP–NAP differences were remarkably stable across the six urban contexts (Figure 7a; Table 7). For architectural evaluation, the average AUP-minus-NAP difference ranged from −0.401 in Beijing to −0.497 in Changsha. Despite this systematic scale shift, the image rankings produced by the two groups remained strongly aligned within every city. Mean Spearman correlations ranged from 0.861 in Shenzhen to 0.961 in Xi'an, while mean CCC values ranged from 0.659 to 0.824. All six architectural dimensions showed significant AUP–NAP differences after BH correction in every city.

Table 7. City-level AUP–NAP agreement and scoring differences.

Domain	City	Mean AUP	Mean NAP	AUP–NAP Difference	Mean ρ	Mean CCC	Mean MAE	BH-Significant Dimensions
Architectural	Beijing	4.081	4.482	−0.401	0.911	0.774	0.461	6/6
Architectural	Changsha	3.661	4.159	−0.497	0.924	0.702	0.518	6/6
Architectural	Guangzhou	3.810	4.282	−0.472	0.956	0.797	0.516	6/6
Architectural	Shenzhen	3.802	4.264	−0.463	0.861	0.659	0.497	6/6
Architectural	Xi'an	4.004	4.419	−0.415	0.961	0.824	0.504	6/6
Architectural	Zhengzhou	4.063	4.501	−0.438	0.910	0.718	0.486	6/6
Streetscape	Beijing	4.091	4.455	−0.364	0.929	0.827	0.393	5/5
Streetscape	Changsha	3.797	4.284	−0.487	0.936	0.730	0.511	5/5
Streetscape	Guangzhou	3.957	4.360	−0.403	0.932	0.806	0.452	5/5
Streetscape	Shenzhen	3.949	4.379	−0.430	0.872	0.730	0.445	5/5
Streetscape	Xi'an	4.093	4.462	−0.368	0.948	0.865	0.428	5/5
Streetscape	Zhengzhou	4.197	4.549	−0.352	0.921	0.748	0.393	5/5

Mean ρ, CCC, and MAE are averaged across the six architectural or five streetscape dimensions within each city. Mean difference is calculated as AUP minus NAP. BH-significant dimensions indicate paired AUP–NAP comparisons remaining significant after Benjamini–Hochberg correction.

The same pattern was observed for streetscape perception. Average AUP-minus-NAP differences ranged from −0.352 in Zhengzhou to −0.487 in Changsha, and all five streetscape dimensions remained significantly different after BH correction in all six cities. Nevertheless, within-city AUP–NAP rank agreement remained high, with mean Spearman correlations ranging from 0.872 in Shenzhen to 0.948 in Xi'an, and mean CCC values ranging from 0.730 to 0.865. Mean absolute differences remained modest, ranging from 0.393 to 0.511 score points. These results reinforce the distinction identified in Section 3.1 that AUP participants consistently assigned lower absolute scores than NAP participants while maintaining similar relative rankings of streetscapes. However, this descriptive pattern does not establish the underlying reasons for the rating-level differences.

Human–AI agreement varied across cities and models, but rank-based and concordance measures remained generally weak (Figure 7b,c; Table S1). For architectural evaluation, city-level MAE ranged from 0.986 for Kimi K3 versus AUP in Changsha to 2.123 for Claude

Fable 5 versus NAP in Beijing. Streetscape MAE ranged from 0.834 for Kimi K3 versus NAP in Changsha to 1.357 for Qwen3-VL-Plus versus AUP in Guangzhou. Thus, city-specific differences mainly concerned the magnitude of absolute scoring discrepancies rather than consistently stronger Human–AI ranking alignment.

Leave-one-city-out sensitivity analysis showed that excluding a single study city changed the model-specific mean MAE by at most 0.092 for architectural evaluation and 0.038 for streetscape evaluation (Figure S5). These findings indicate that the absolute-error results were not dominated by one city, although this analysis does not test alternative selections of stimuli within districts or the robustness of rank correlations.

XGBoost leave-one-city-out validation (Figure 7d; Table S2) yielded negative pooled R^2 values for all five AUP and NAP outcomes, but modest positive values for VLM consensus (0.006–0.237), with variation across held-out cities. This indicates that the predictive contrast was not confined to within-city cross-validation, while the limited transferability and small, nonrepresentative sample warrant caution in extending these patterns to other urban contexts.

4. Discussion

4.1. Disciplinary Background, Rating-Level Differences, and Human–AI Perception Differences

A central finding of this study is that architecture/urban-planning students (AUP) and students without such training (NAP) exhibited highly consistent visual preference rankings but systematically different absolute ratings. Across architectural dimensions, AUP–NAP Spearman correlations ranged from 0.914 to 0.946, while AUP scores were approximately 0.43–0.46 points lower. Similar differences were observed across streetscape dimensions, and the pattern persisted across all six cities. The AUP group also demonstrated higher image-level consensus reliability. However, greater within-group reliability reflects consistency rather than evaluative accuracy, and the observed rating-level differences cannot establish that either group applied more appropriate or professionally informed assessment standards.

This finding both supports and extends previous research on architectural expertise. Gifford et al. [22] identified substantial differences between architects and laypersons in architectural aesthetic judgments, while Brown and Gifford [23] found that lay observers generally provided more favorable evaluations than practicing architects. Similarly, Xu et al. [46] reported that design professionals adopted more stringent evaluations of urban public spaces. However, unlike the pronounced preference divergence reported in some classical studies, our participants largely agreed on which environments were relatively favorable. This distinction may reflect differences in participant experience and evaluation context, consistent with evidence that younger architects and non-architects do not always exhibit substantial aesthetic differences [25]. Thus, differences in rating levels should not automatically be interpreted as differences in preference rankings.

One plausible interpretation is that AUP and NAP participants formed broadly similar relative judgments but differed in how they translated these impressions into numerical ratings. Familiarity with architectural composition, proportion, identity, and spatial relationships may have contributed to the lower scores assigned by AUP participants, although differences in response-scale use or pre-existing participant characteristics could also produce a similar pattern. Previous eye-tracking research has also highlighted the relevance of domain-specific visual expertise [48], while recent reviews emphasize the role of architectural knowledge in visual attention [49]. Recent cross-national survey evidence further highlights the importance of distinguishing public architectural preferences from professional recognition. Mouratidis [50] found that respondents in Norway and Greece generally preferred traditional over contemporary building façades, including comparisons

involving contemporary buildings that had received architectural awards or nominations. Although that study did not directly compare educational backgrounds, its findings reinforce the importance of clearly defining the human reference population in architectural evaluation. Nevertheless, the present experiment did not directly measure visual attention or evaluative reasoning and did not include an independent external benchmark against which the accuracy of either group's ratings could be assessed. Accordingly, neither the lower AUP scores nor their higher consensus reliability establishes greater accuracy or more stringent evaluative standards, and the observed group differences cannot be causally attributed to architectural education.

These findings also have implications for Human–AI alignment. Although some VLMs produced absolute scores closer to AUP or NAP, their near-zero rank correlations with both groups indicate that numerical similarity does not imply comparable architectural judgment. This distinction complements recent research showing that Human–AI agreement in architectural evaluation varies across tasks and perceptual dimensions [42,43]. Accordingly, AI benchmarking should distinguish between agreement in rating levels and agreement in preference rankings, while explicitly defining the human reference population. The consistent AUP–NAP differences across cities reinforce this implication, although the undergraduate sample limits generalization to practicing architects and the broader public.

4.2. Human–AI Alignment Across Architectural and Streetscape Evaluation

The key contrast is therefore between agreement about the relative order of scenes and agreement about the numerical score assigned to each scene. Across the 120 stimuli, the two human groups showed strong rank agreement for architectural frontages (Spearman's $\rho = 0.914$–0.946) and streetscapes ($\rho = 0.923$–0.943), despite systematic differences in mean ratings. This indicates that the groups largely identified the same scenes as more or less favorable within the study sample, without establishing that they applied identical evaluation criteria. By comparison, correlations between individual VLMs and either human group were close to zero for façades ($\rho = -0.182$ to 0.081) and streetscapes ($\rho = -0.099$ to 0.087), even though pairwise inter-VLM correlations were markedly higher ($\rho = 0.516$–0.863 and 0.533–0.832, respectively). These results suggest that the models produced a relatively consistent machine-based ordering that was not shared by either human group. Because rank agreement is invariant to strictly increasing transformations of scores, correcting a model's mean rating or rescaling its output would not, by itself, recover the human ordering. Identifying why the orderings differ would require evidence about cue use or evaluative reasoning, neither of which was directly observed in the present experiment.

This finding is more cautious than some recent human–AI perception studies. Ho et al. [31], for example, reported substantial correspondence between human observers and multimodal models for several natural-landscape preference dimensions, although alignment deteriorated for more nuanced properties such as legibility. Law et al. [32] similarly found that Human–AI alignment in façade evaluation was stronger for relatively explicit descriptors than for affective attributes. More recently, Yao et al. [51] reported moderate-to-high Human–VLM consistency in several dimensions of urban street-scene perception. The weaker alignment observed in the present study therefore does not necessarily contradict those findings; rather, it suggests that alignment is strongly task dependent. The present evaluation required models to move beyond object recognition or broad visual preference and assign judgments concerning compositional coherence, proportion and scale, material coherence, articulation, architectural identity, building–street integration, and overall design quality. A particularly relevant comparison is provided by Belaroussi and Martín-Gutierrez [42], who found that GPT-4V matched the majority human response in nine of twelve categorical evaluations of architectural ambiance. Belaroussi [43] sub-

sequently extended this line of inquiry by comparing several multimodal models with architecture, engineering, and construction experts, documenting both attribute-specific agreement and discrepancies in subjective environmental assessment. The present findings should be interpreted alongside, rather than as a direct contradiction of, these studies. Their evaluations involved a small set of virtual environments and categorical judgments, whereas our experiment assessed the image-level ranking and absolute ratings of 120 real-world urban stimuli across eleven dimensions. Agreement with the dominant human category does not necessarily indicate that a model can reproduce the relative ranking of numerous scenes. Our results therefore underscore the importance of reporting multiple agreement metrics and specifying the human reference population when evaluating architectural AI.

This interpretation is also consistent with the distinction between recognition and higher-order architectural judgment identified in recent built-environment AI research. VLMs have demonstrated substantial capacity to retrieve observable façade attributes such as materials, condition, and architectural style from street imagery [35], and recent work has shown that multimodal models can operate at urban scale for environmental or restorative-quality assessment [36]. Such tasks benefit from strong object, material, and scene-recognition capabilities. Architectural evaluation, however, requires interpreting how multiple visible elements are organized into a coherent whole. Lee et al. [21] similarly showed that computational measures of façade complexity do not map straightforwardly onto human visual preference, while recent explainable architectural-image research has emphasized the need to connect machine predictions to architecturally meaningful evidence [45]. The present findings extend this distinction by demonstrating that five different VLM families could converge on a broadly similar machine ordering of architectural stimuli while that ordering remained largely unrelated to the human consensus ordering.

The strong inter-VLM agreement observed in this study is therefore itself informative. It suggests that the models may share regularities acquired from overlapping image–text corpora, broadly similar visual encoders, or comparable learned representations of visually explicit scene attributes. However, such agreement should not be interpreted as evidence of a shared human-like cognitive mechanism. Rather, the convergence indicates that several independently developed models can produce similar rating structures even when those structures differ from human judgments. This creates an important methodological caution for architectural AI studies: agreement among several AI systems does not constitute external validation unless their outputs are also compared against appropriate human reference groups.

The larger Human–AI errors observed for architectural-frontage evaluation than for broader streetscape perception become particularly informative when the two evaluation domains are considered jointly. Across both domains, AUP and NAP participants exhibited strong agreement in their relative rankings despite systematic differences in absolute rating levels, whereas the five VLMs showed consistently weak rank correspondence with both human groups. This indicates that Human–AI ranking disagreement was not confined to either architectural-frontage or streetscape evaluation. Nevertheless, the magnitude of absolute scoring discrepancies differed between the two domains, with architectural evaluations generally exhibiting larger errors. One plausible explanation concerns the different visual properties emphasized by the two assessment tasks. Streetscape dimensions such as greenery-related pleasantness, perceived environmental safety, and pedestrian comfort can be partly associated with readily detectable scene components. In contrast, architectural identity, articulation, proportional quality, and overall façade design quality involve more relational characteristics, including compositional organization, visual hierarchy, material coherence, and architectural conventions. Recent eye-tracking research in architecture similarly emphasizes that architectural visual experience involves selective attention to

form, layout, lighting, spatial organization, and other relational characteristics rather than merely counting visible elements [49]. These considerations suggest that general-purpose VLMs may face additional challenges when evaluations require judgments about how visible architectural elements are organized into a coherent whole. However, because the two domains employed different visual representations and rating dimensions, and some building-oriented views contained limited or partially obscured architectural frontages, the observed MAE differences cannot isolate intrinsic task difficulty from stimulus-framing effects or identify the underlying perceptual mechanisms. Rather, the joint analysis reveals a consistent pattern of weak Human–AI ranking agreement across evaluation domains alongside task-specific differences in absolute scoring discrepancies, highlighting the importance of examining architectural and streetscape evaluations within a common analytical framework.

4.3. Objective Street Features and Environmental Perception

A further finding from the SegFormer–XGBoost–SHAP analysis is that the same set of image-derived semantic composition indicators exhibited different predictive associations with VLM and human evaluation scores. Previous street-view research has repeatedly demonstrated that semantic indicators such as greenery, building enclosure, sky visibility, pedestrian infrastructure, traffic, and spatial openness can explain meaningful portions of human environmental perception [12–16]. Recent studies have further combined street imagery, semantic segmentation, machine learning, and SHAP to identify nonlinear relationships between street morphology and perceived urban quality [52,53]. For example, Li et al. [52] found that morphological compactness, greenery, and openness contributed differently to safety, liveliness, aesthetic, and emotional perception, while Kuang et al. [53] reported associations between street quality and factors including building density, visible sky, and greenery. At the same time, research on architectural perception has emphasized the relevance of geometric organization, visual attention, and affective responses [27,38,39]. However, because visual attention and emotional responses were not directly measured in our experiment, they remain candidate explanations for the observed limitations of semantic-composition predictors rather than empirically verified mediating mechanisms.

Against this background, the negative cross-validated R^2 values for the AUP and NAP models are notable. They do not imply that physical street conditions are irrelevant to human perception. Rather, they indicate that models based on the eight semantic composition indicators did not reliably predict the numerical human consensus scores for held-out images under the implemented cross-validation procedures. This distinction is important. Many previous studies have focused on one city or a relatively homogeneous spatial context, whereas the present analysis deliberately combined six geographically and morphologically different urban districts and assessed both repeated cross-validation and leave-one-city-out transfer. The complementary leave-one-city-out validation provides a more geographically separated assessment of transferability, although within-city spatial dependence remains a limitation of the repeated cross-validation procedure.

A second explanation concerns the representation of the environment. Semantic segmentation compresses a complex scene into proportional quantities. Two streets may contain identical amounts of vegetation, buildings, roads, and sky while differing substantially in tree placement, façade continuity, enclosure, shade, setback, maintenance, material quality, pedestrian permeability, storefront activation, and visual hierarchy. Human observers can integrate these configurational characteristics into a holistic judgment, whereas the eight predictors used in this study largely describe how much of each semantic category is visible rather than how those elements are spatially organized. In architectural and landscape design, the arrangement and combination of physical elements can be understood in

terms of design syntax. For example, Sharma [54] examined the design syntax of urban greenways through the relationships among planting, pathways, materials, and spatial assemblies. Although the present façade evaluation addresses perceived compositional qualities, the SegFormer-derived streetscape indicators primarily quantify the visible proportions of individual elements rather than their spatial organization. Thus, these indicators capture aspects of the visual design vocabulary but do not directly measure the relationships constituting design syntax. The contrast between high human consensus reliability and poor out-of-sample XGBoost performance suggests that simple semantic proportions were insufficient to capture the variation in human ratings. However, the available results do not establish which omitted visual or contextual characteristics account for this limitation, and SHAP patterns from the poorly performing human models should not be interpreted as validated perceptual mechanisms.

The VLM-consensus results followed a different pattern. All five perception outcomes produced positive repeated-CV R^2 values, and most retained positive performance under cross-city validation. Pleasantness and pedestrian comfort showed the strongest transferable predictability. SHAP further revealed relatively concentrated feature structures: vegetation dominated VLM predictions of PL, PC, and OSQ; roadway proportion was most influential for SAF; and visible sky proportion was particularly important for BSI. These findings partly resemble earlier image-based perception studies in which greenery and openness were repeatedly associated with positive environmental evaluation [2,14,52,53]. Read alongside the weak Human–AI rank agreement, the greater out-of-sample predictability of VLM ratings from these eight semantic indicators is consistent with the possibility that model scores reflect a more restricted set of visible scene cues. This remains a hypothesis rather than evidence of specific human or model reasoning processes, which were not directly measured.

The VEG–PL and SW–PC patterns in Figure 6 illustrate model-dependent nonlinear associations that are broadly consistent with the relevance of greenery in urban perception [2,14,18]. However, the flatter AUP and NAP curves should not be interpreted as evidence that human observers disregard these features, given the negative out-of-sample R^2 of their underlying models.

The SKY–BSI pattern further demonstrates why SHAP relationships should be interpreted as model associations rather than causal design effects. In the VLM-consensus model, a higher visible sky proportion (SKY) was associated with increasingly negative SHAP contributions to building–street integration. One plausible architectural interpretation is that higher sky exposure may occur in scenes with weaker enclosure, larger setbacks, or discontinuous building frontages. However, these configurational characteristics were not directly measured, and sky proportion itself cannot be regarded as an independent causal factor. Similar interpretive limitations arise from the compositional dependence among the semantic indicators. In particular, the elevated VIF values for vegetation proportion (VEG = 13.05) and visible building proportion (BLD = 12.95) indicate substantial linear dependence, potentially affecting the allocation of predictive importance between correlated features. Related research by Yan et al. [47] demonstrates the application of XGBoost–SHAP to nonlinear spatial associations involving multisource GIS and street-view information; nevertheless, model-derived feature contributions should not be equated with causal mechanisms. This distinction is especially important in the present study because the AUP and NAP models exhibited negative out-of-sample R^2 values across all five streetscape-perception outcomes (Table 6). The bootstrap analysis further demonstrated that the AUP and NAP feature-importance rankings were sensitive to resampling. Combined with their negative out-of-sample R^2 values, this instability reinforces the need to interpret the human SHAP results as exploratory characteristics of fitted models rather than validated

perceptual mechanisms. Although the VLM-consensus models exhibited comparatively more reproducible importance rankings and modest positive predictive performance, their feature attributions remain model-dependent and cannot establish causal environmental effects. The resulting feature contributions therefore indicate model-dependent statistical associations, not independent causal effects of individual street-space elements.

This distinction also explains why the present results differ from studies that achieve stronger human-perception prediction using richer feature sets. Recent work has incorporated not only semantic proportions but morphological compactness, street-network structure, accessibility, image-derived spatial indicators, and nonlinear interaction terms [13,52]. Future human-perception models should therefore move beyond class proportions toward descriptors of spatial arrangement, continuity, depth, enclosure, façade organization, vegetation configuration, and object relationships. Such enrichment may narrow the gap between the relatively simple visual heuristics captured by the VLMs and the more holistic human assessments observed here.

4.4. Design Implications, Limitations, and Future Research

The findings have several implications for architectural and urban-design practice. First, current VLMs appear more suitable for large-scale preliminary screening and decision support than for autonomous design-quality assessment. Their ability to process large image collections rapidly and generate internally consistent evaluations may help identify atypical environments, prioritize locations for further review, and support exploratory urban audits, consistent with recent VLM-based street-scene applications [51]. However, the weak Human–AI correspondence observed here indicates that VLM scores should not be treated as substitutes for professional or public judgment. For architects and planners, this could mean using AI-generated evaluations to flag potentially unusual building frontages or street environments for closer inspection, while leaving judgments about design quality and appropriate interventions to human assessors.

Second, the stronger dependence of VLM ratings on explicit semantic features such as vegetation, visible roadway, sidewalk, and sky proportions suggests that multimodal models may be particularly useful for diagnosing clearly observable environmental conditions. Nevertheless, these features should not be translated mechanically into design prescriptions. Human architectural and streetscape evaluation involves relational and contextual qualities—including spatial configuration, façade continuity, enclosure, hierarchy, material expression, and building–street relationships—that cannot be represented by isolated pixel proportions. The stable difference between AUP and NAP ratings further shows that "human preference" is not a single universal benchmark. AI systems intended for participatory planning should therefore be calibrated against appropriate public groups, whereas tools for architectural review should incorporate discipline-specific reference populations.

The framework also offers potential applications in architecture and urban-planning education. In design studios, students could compare their own evaluations of building frontages and streetscapes with peer and VLM-generated scores, using disagreements to discuss architectural composition, spatial relationships, and contextual judgment. This proposed educational application was not directly evaluated as a learning outcome in the present study.

Several limitations should be acknowledged. The formal Human–AI experiment used 120 balanced stimuli from six Chinese urban districts, although these were derived from a much larger database of 39,047 panoramas. This controlled design supported repeated evaluation and cross-city comparison but limits generalization to other cultural contexts, architectural traditions, and urban morphologies. The evaluation sample included one location from each of 20 spatial clusters per district, but was not stratified by building typology or

street morphology. Furthermore, some automatically selected architectural views contained limited or partially obscured building frontages, and the resulting differences in visual framing may have influenced the evaluation scores. Although the leave-one-city-out MAE analysis indicated limited sensitivity to city exclusion, it did not directly assess the effects of alternative within-district stimulus selection or the flagged low-visibility views. Furthermore, because repeated XGBoost cross-validation was stratified by city rather than spatially blocked, geographically nearby observations could have appeared in different training and test folds, potentially affecting predictive performance estimates through residual spatial autocorrelation. Although the separate LOCO validation eliminated same-city overlap between training and test observations, it did not directly quantify within-city spatial dependence. The participant sample was also relatively narrow, comprising third- and fourth-year undergraduates from a single university, although both groups had identical gender distributions. Potential differences in exact age distributions, academic-year composition, and prior visual experience may nevertheless have contributed to the observed evaluation patterns. Future work should extend the benchmark to practicing architects, urban designers, residents, and more diverse demographic groups. Eye-tracking or verbal-protocol methods could also clarify whether groups that produce similar ratings actually attend to the same visual evidence [48,49]. An additional limitation is that the experiment measured explicit ratings but did not directly observe visual attention, emotional responses, or physiological reactions during evaluation. Furthermore, the use of generic seven-point verbal anchors across different perceptual dimensions may have introduced variation in how participants interpreted the response scale, particularly for pleasantness, perceived safety, and pedestrian comfort; future studies should consider piloting dimension-specific anchors. Previous studies have demonstrated the feasibility of measuring gaze allocation in architectural and urban scenes [39], investigating computational attention estimates [40], and examining facial-expression-based correlates of architectural responses [41]. Future work could combine these methods with architecture-related expertise comparisons to investigate whether systematic rating-level differences are associated with visual attention, interpretive criteria, affective experience, or response-scale use. Longitudinal studies following students through architectural education, together with comparisons involving experienced practitioners and more diverse public samples, would also help distinguish the effects of training from selection and cohort differences.

Methodological limitations also arise from the visual representation and modeling pipeline. This limitation is also highlighted by Qin et al. [33], who demonstrated that the preservation of visible historic streetscape characteristics does not necessarily correspond to the preservation of auditory locality within the same urban environment. Their findings suggest that visual and auditory dimensions may capture different aspects of place-based perception. Accordingly, the present Human–AI comparisons should be interpreted as assessments of image-based architectural and streetscape qualities rather than comprehensive evaluations of multisensory environmental experience. Future research could integrate synchronized street-view imagery, environmental audio, and human perceptual assessments to investigate whether multimodal inputs improve the correspondence between AI-generated evaluations and human judgments. Static street-view imagery cannot reproduce movement, sound, thermal conditions, depth perception, or sequential pedestrian experience, and image acquisition conditions may introduce additional variation. The SegFormer-B5 model was fine-tuned on Cityscapes rather than locally annotated Baidu Street View imagery, and its segmentation outputs were assessed through a qualitative audit of 30 perspective views without a quantitative pixel-level validation benchmark. Differences in image acquisition conditions, projection characteristics, and urban scene composition may introduce domain shift and class-specific classification errors. This limitation is partic-

ularly relevant to relatively small categories, such as sidewalks, pedestrians, and bicycles, for which misclassification may affect the derived visible sidewalk proportion (SW) and visible person-and-bicycle proportion (HA) indicators. Because these indicators were subsequently used in the XGBoost–SHAP analysis, segmentation errors could propagate into model predictions and feature attributions. Moreover, positive downstream predictive performance does not independently establish segmentation accuracy. Accordingly, the reported SHAP relationships should be interpreted as conditional on the unverified accuracy of the semantic indicators rather than as validated effects of individual street-space elements. Future research should evaluate segmentation performance using geographically distributed, manually annotated street-view samples and report category-specific IoU, precision, and recall. In addition, the eight semantic predictors described scene composition but not spatial topology. Future studies should therefore combine semantic segmentation with depth estimation, façade continuity, enclosure, setback, object adjacency, scene graphs, or learned spatial embeddings. Such richer representations may better capture the relational characteristics emphasized in architectural interpretation [28,45].

Finally, the explainable-learning and VLM results should be interpreted conservatively. Negative cross-validated R^2 values for AUP and NAP indicate that their SHAP patterns describe fitted model behavior rather than robust predictors of human perception, whereas the VLM-consensus SHAP results were better supported by positive cross-validation and LOCO performance. In addition, rapidly evolving model versions, prompts, and inference settings may affect VLM outputs. Although the present study froze stimuli, prompts, scoring criteria, and formal model runs, future research should examine prompt sensitivity, repeated-inference stability, updated model versions, and pairwise or ranking-based evaluation schemes.

5. Conclusions

This study investigated Human–AI alignment across two complementary domains of built-environment evaluation—architectural frontages and urban streetscapes—using paired visual stimuli from six Chinese urban districts. Across both evaluation domains, architecture/urban-planning students and students without such training exhibited strong agreement in the relative ranking of visual stimuli despite systematic differences in their absolute ratings. In contrast, the five VLMs demonstrated weak correspondence with both human reference groups, even though the models showed moderate-to-strong agreement with one another. The joint analysis revealed that Human–AI ranking discrepancies persisted across architectural and streetscape evaluations, while the magnitude of absolute scoring errors varied between the two tasks. These findings establish the importance of distinguishing evaluation domains, human reference populations, and agreement metrics when assessing AI-generated judgments of built environments. The principal contribution of this study therefore lies in providing a human-referenced, two-domain framework for evaluating Human–AI alignment, rather than simply integrating multiple computational techniques. Such a framework highlights the need to assess ranking consistency and absolute scoring agreement separately and cautions against interpreting agreement among VLMs as evidence of correspondence with human architectural judgment.

The explainable street-space analysis further revealed a fundamental difference between human and VLM evaluation. Eight SegFormer-derived indicators describing the visible proportions of vegetation, sky, buildings, sidewalks, roads, motor vehicles, persons and bicycles, together with semantic-class diversity did not achieve positive out-of-sample predictive performance for either human group. By contrast, VLM-consensus ratings showed modest but transferable predictability, particularly for pleasantness and pedestrian comfort, and were more strongly structured by explicit semantic cues such as vegetation,

visible roadway, sidewalk, and sky proportions. Cross-city and leave-one-city-out analyses confirmed that the principal Human–AI differences were not attributable to a single urban context.

These findings suggest that current VLMs should be regarded as complementary analytical tools rather than substitutes for human architectural judgment. For everyday architectural and planning practice, the central message is simple: AI can help review large collections of building and street images efficiently, but it cannot yet reliably determine which environments people will judge as well designed. Human reference groups remain essential when AI systems are used in architectural review, urban-design assessment, or participatory planning. In architecture and planning education, the same comparative framework may also support critical reflection on the differences between human and AI-generated design judgments. Future research should expand human benchmarks to professional and culturally diverse populations, incorporate richer spatial and relational visual descriptors, and examine whether improved multimodal models can move beyond explicit semantic cues toward the contextual and configurational reasoning that characterizes human evaluation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/buildings16224568/s1, Figure S1: Pairwise correlations among the eight SegFormer-derived street-space indicators. Figure S2: Visual quality audit of SegFormer semantic segmentation across 30 pre-specified perspective views sampled from the six study cities. Figure S3: SHAP summary distributions for the VLM-consensus XGBoost models across five streetscape-perception outcomes. Figure S4: City-level evaluation profiles for the AUP, NAP, and VLM-consensus groups. Figure S5: Leave-one-city-out sensitivity of Human–AI agreement. Table S1: City-level Human–AI agreement across architectural-frontage and streetscape evaluation. Table S2: Pooled leave-one-city-out predictive performance of XGBoost models for streetscape perception. Table S3: Documented VLM configurations and inference settings. Table S4: Pointwise bootstrap 95% confidence intervals for AUP–NAP and Human–AI agreement metrics. Table S5: Cross-validated predictive performance of XGBoost and mean-prediction benchmarks. Table S6: SHAP feature-importance ranking stability under city-stratified bootstrap resampling. Text S1: Verbatim prompts used in the formal VLM evaluation. Text S2: XGBoost configuration and hyperparameter optimization. Supplementary Data S1: Model-specific optimized XGBoost hyperparameters and fold-level cross-validation results.

Author Contributions: Conceptualization, Y.M.; Methodology, Y.M. and X.L.; Software, Y.M.; Writing—original draft, Y.M.; Writing—review & editing, X.L.; Supervision, X.L.; Project administration, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (Grant No. 52568001).

Institutional Review Board Statement: The study was conducted in accordance with the applicable institutional research ethics requirements and was approved by the Institutional Review Board of School of Architecture, Changsha University of Science and Technology (Approval No. 20260902; approval date: 2 September 2026).

Informed Consent Statement: Informed consent for participation was obtained from all participants prior to the visual evaluation experiment.

Data Availability Statement: The summary statistics, model configurations, evaluation prompts, and supplementary robustness analyses supporting the findings of this study are provided in the article and its Supplementary Materials. The derived image-level human consensus ratings, VLM evaluation scores, semantic street-space indicators, and analytical scripts are available from the corresponding author upon reasonable request. The original Baidu Street View imagery was obtained from a third-party platform and is not redistributed by the authors; access to and reuse of these images are subject to the data provider's applicable terms. Individual-level participant responses are

not publicly shared to protect participant privacy, and requests for de-identified research data will be considered in accordance with applicable ethical requirements and participant consent.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AD	Façade articulation and depth
AUP	Architecture or urban–rural planning
BLD	Visible building proportion
BSI	Building–street integration
CC	Compositional coherence
CCC	Lin's concordance correlation coefficient
HA	Visible person-and-bicycle proportion
ID	Architectural identity
LOCO	Leave-one-city-out
MAE	Mean absolute error
MC	Material and color coherence
NAP	Non-architecture/planning
ODQ	Overall design quality
OSQ	Overall streetscape quality
PC	Pedestrian comfort
PL	Pleasantness
PS	Proportion and scale
RD	Visible roadway proportion
SAF	Perceived environmental safety
SHAP	SHapley Additive exPlanations
SKY	Visible sky proportion
SVC	Semantic-class diversity (entropy)
SVI	Street-view imagery
SVID	Street View ID
SW	Visible sidewalk proportion
TRA	Visible motor-vehicle proportion
VEG	Vegetation proportion
VIF	Variance inflation factor
VLMs	Vision–language models

References

1. St-Jean, P.; Clark, O.G.; Jemtrud, M. A review of the effects of architectural stimuli on human psychology and physiology. *Build. Environ.* **2022**, *219*, 109182. [CrossRef]
2. Chen, C.; Li, H.; Luo, W.; Xie, J.; Yao, J.; Wu, L.; Xia, Y. Predicting the effect of street environment on residents' mood states in large urban areas using machine learning and street view images. *Sci. Total Environ.* **2022**, *816*, 151605. [CrossRef] [PubMed]
3. Wei, J.; Yue, W.; Li, M.; Gao, J. Mapping human perception of urban landscape from street-view images: A deep-learning approach. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102886. [CrossRef]
4. Zhou, H.; Wang, J.; Wilson, K. Impacts of perceived safety and beauty of park environments on time spent in parks: Examining the potential of street view imagery and phone-based GPS data. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103078. [CrossRef]
5. Biljecki, F.; Zhao, T.; Liang, X.; Hou, Y. Sensitivity of measuring the urban form and greenery using street-level imagery: A comparative study of approaches and visual perspectives. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *122*, 103385. [CrossRef]
6. Chen, S.; Biljecki, F. Automatic assessment of public open spaces using street view imagery. *Cities* **2023**, *137*, 104329. [CrossRef]
7. Jeon, J.; Woo, A. Deep learning analysis of street panorama images to evaluate the streetscape walkability of neighborhoods for subsidized families in Seoul, Korea. *Landsc. Urban Plan.* **2023**, *230*, 104631. [CrossRef]
8. Cui, Q.; Zhang, Y.; Yang, G.; Huang, Y.; Chen, Y. Analysing gender differences in the perceived safety from street view imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103537. [CrossRef]
9. Liang, X.; Zhao, T.; Biljecki, F. Revealing spatio-temporal evolution of urban visual environments with street view imagery. *Landsc. Urban Plan.* **2023**, *237*, 104802. [CrossRef]

10. Koo, B.W.; Guhathakurta, S.; Botchwey, N.; Hipp, A. Can good microscale pedestrian streetscapes enhance the benefits of macroscale accessible urban form? An automated audit approach using Google street view images. *Landsc. Urban Plan.* **2023**, *237*, 104816. [CrossRef]
11. Ki, D.; Park, K.; Chen, Z. Bridging the gap between pedestrian and street views for human-centric environment measurement: A GIS-based 3D virtual environment. *Landsc. Urban Plan.* **2023**, *240*, 104873. [CrossRef]
12. Hamim, O.F.; Ukkusuri, S.V. Towards safer streets: A framework for unveiling pedestrians' perceived road safety using street view imagery. *Accid. Anal. Prev.* **2023**, *195*, 107400. [CrossRef] [PubMed]
13. Zhao, X.; Lu, Y.; Lin, G. An integrated deep learning approach for assessing the visual qualities of built environments utilizing street view images. *Eng. Appl. Artif. Intell.* **2024**, *130*, 107805. [CrossRef]
14. Wu, Y.; Liu, Q.; Hang, T.; Yang, Y.; Wang, Y.; Cao, L. Integrating restorative perception into urban street planning: A framework using street view images, deep learning, and space syntax. *Cities* **2024**, *147*, 104791. [CrossRef]
15. Ogawa, Y.; Oki, T.; Zhao, C.; Sekimoto, Y. Evaluating the subjective perceptions of streetscapes using street-view images. *Landsc. Urban Plan.* **2024**, *247*, 105073. [CrossRef]
16. Li, J.; Zhang, X.; Li, L.; Wang, X.; Cheng, J.; Gao, C.; Ling, J. An estimation method for multidimensional urban street walkability based on panoramic semantic segmentation and domain adaptation. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108905. [CrossRef]
17. Yu, X.; Ma, J.; Tang, Y.; Yang, T.; Jiang, F. Can we trust our eyes? Interpreting the misperception of road safety from street view images and deep learning. *Accid. Anal. Prev.* **2024**, *197*, 107455. [CrossRef] [PubMed]
18. Ito, K.; Kang, Y.; Zhang, Y.; Zhang, F.; Biljecki, F. Understanding urban perception with visual data: A systematic review. *Cities* **2024**, *152*, 105169. [CrossRef]
19. Gu, Y.; Quintana, M.; Liang, X.; Ito, K.; Yap, W.; Biljecki, F. Designing effective image-based surveys for urban visual perception. *Landsc. Urban Plan.* **2025**, *260*, 105368. [CrossRef]
20. Liang, X.; Chang, J.H.; Gao, S.; Zhao, T.; Biljecki, F. Evaluating human perception of building exteriors using street view imagery. *Build. Environ.* **2024**, *263*, 111875. [CrossRef]
21. Lee, C.-C.; Wang, Y.; Tang, C.; Li, X.; Yin, J. A visual preference analysis of building façades in street view images: A case study of Nanshan District, Shenzhen. *Buildings* **2025**, *15*, 611. [CrossRef]
22. Gifford, R.; Hine, D.W.; Muller-Clemm, W.; Reynolds, D.J., Jr.; Shaw, K.T. Decoding Modern Architecture: A Lens Model Approach for Understanding the Aesthetic Differences of Architects and Laypersons. *Environ. Behav.* **2000**, *32*, 163–187.
23. Brown, G.; Gifford, R. Architects Predict Lay Evaluations of Large Contemporary Buildings: Whose Conceptual Properties? *J. Environ. Psychol.* **2001**, *21*, 93–99. [CrossRef]
24. Gifford, R.; Hine, D.W.; Muller-Clemm, W.; Shaw, K.T. Why Architects and Laypersons Judge Buildings Differently: Cognitive Properties and Physical Bases. *J. Archit. Plan. Res.* **2002**, *19*, 131–148.
25. Šafárová, K.; Pírko, M.; Juřík, V.; Pavlica, T.; Németh, O. Differences between Young Architects' and Non-Architects' Aesthetic Evaluation of Buildings. *Front. Archit. Res.* **2019**, *8*, 229–237. [CrossRef]
26. Belaroussi, R.; Díaz González, E.; Dupin, F.; Martin-Gutierrez, J. Appraisal of Architectural Ambiances in a Future District. *Sustainability* **2023**, *15*, 13295. [CrossRef]
27. Mehaffy, M.W. The Impacts of Symmetry in Architecture and Urbanism: Toward a New Research Agenda. *Buildings* **2020**, *10*, 249. [CrossRef]
28. Zhang, J.; Huang, J.; Jin, S.; Lu, S. Vision-Language Models for Vision Tasks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5625–5644. [CrossRef] [PubMed]
29. Jang, S.; Roh, H.; Lee, G. Generative AI in architectural design: Application, data, and evaluation methods. *Autom. Constr.* **2025**, *174*, 106174. [CrossRef]
30. Zheng, Y.; Xu, F.; Lin, Y.; Santi, P.; Ratti, C.; Wang, Q.R.; Li, Y. Urban planning in the era of large language models. *Nat. Comput. Sci.* **2025**, *5*, 727–736. Correction in *Nat. Comput. Sci.* **2025**, *5*, 1091. [CrossRef] [PubMed]
31. Ho, L.; Chen, C.-C.; Chang, C.-Y. Research note: Assessing human preferences for natural landscapes—An analysis of ChatGPT-4 and LLaVA models. *Landsc. Urban Plan.* **2025**, *259*, 105371. [CrossRef]
32. Law, S.; Valentine, C.; Kahlon, Y.; Seresinhe, C.I.; Tang, J.; Morad, M.G.; Fujii, H. Generative AI for Architectural Façade Design: Measuring Perceptual Alignment Across Geographical, Objective, and Affective Descriptors. *Buildings* **2025**, *15*, 3212. [CrossRef]
33. Qin, Y.; Yang, D.; Huang, Y. When Visual Conservation Meets Auditory Homogenization: Multimodal Evidence from Ximen Street Historic District, Qujing, China. *Buildings* **2026**, *16*, 2871. [CrossRef]
34. Wang, S. Development of approach to an automated acquisition of static street view images using transformer architecture for analysis of building characteristics. *Sci. Rep.* **2025**, *15*, 29062. [CrossRef] [PubMed]
35. Raghu, D.; Armeni, I.; De Wolf, C. Urban-scale facade material mapping from street view images using vision–language models for circular construction planning. *Sci. Rep.* **2026**, *16*, 20691. [CrossRef] [PubMed]
36. Ma, H.; Kwan, M.-P. Measuring the psychological restorative quality of urban spaces: A vision language model-based method. *Sci. Rep.* **2026**, *16*, 16534. [CrossRef] [PubMed]

37. Mouratidis, K.; Hassan, R. Contemporary versus Traditional Styles in Architecture and Public Space: A Virtual Reality Study with 360-Degree Videos. *Cities* **2020**, *97*, 102499. [CrossRef]
38. Sussman, A.; Hollander, J.B. *Cognitive Architecture: Designing for How We Respond to the Built Environment*, 2nd ed.; Routledge: London, UK, 2021.
39. Hollander, J.B.; Sussman, A.; Purdy Levering, A.; Foster-Karim, C. Using Eye-Tracking to Understand Human Responses to Traditional Neighborhood Designs. *Plan. Pract. Res.* **2020**, *35*, 485–509. [CrossRef]
40. Hollander, J.B.; Sussman, A.; Lowitt, P.; Angus, N.; Situ, M. Eye-Tracking Emulation Software: A Promising Urban Design Tool. *Archit. Sci. Rev.* **2021**, *64*, 383–393. [CrossRef]
41. Lavdas, A.A.; Sussman, A. Quantifying Emotional Responses to Traditional and Modern Architecture: The Case of US Federal Buildings. *Appl. Sci.* **2026**, *16*, 4406. [CrossRef]
42. Belaroussi, R.; Martín-Gutierrez, J. Architectural Ambiance: ChatGPT Versus Human Perception. *Electronics* **2025**, *14*, 2184. [CrossRef]
43. Belaroussi, R. Subjective Assessment of a Built Environment by ChatGPT, Gemini and Grok: Comparison with Architecture, Engineering and Construction Expert Perception. *Big Data Cogn. Comput.* **2025**, *9*, 100. [CrossRef]
44. Lavdas, A.A.; Mehaffy, M.W.; Salingaros, N.A. AI, the Beauty of Places, and the Metaverse: Beyond "Geometrical Fundamentalism". *Archit. Intell.* **2023**, *2*, 8. [CrossRef]
45. Yin, J.; Li, P.; Li, T.; Zhong, J.; Jin, Z.; Feng, T.; Russell, P. ArchiExplain: Multi-Level Evidence Chains for Precedent-Based Interpretability in Architectural Image Understanding. *Buildings* **2026**, *16*, 2394. [CrossRef]
46. Xu, L.; Zhang, Y.; Li, F.; Yin, J. Perceptual difference of urban public spaces between design professionals and 'laypersons': Evidence, health implications and ready-made urban design templates. *Indoor Built Environ.* **2023**, *32*, 494–520. [CrossRef]
47. Yan, Y.; Xu, X.; Huang, Y.; Zhang, Q. Micro-scale Land-Use Functional Patterns and Driving Mechanisms in Historic Districts: A Multi-Source GIS and Interpretable Machine Learning Approach. *Front. Environ. Sci.* **2026**, *14*, 1784087. [CrossRef]
48. Ivy, S.; Rohovit, T.; Stefanucci, J.; Stokes, D.; Mills, M.; Drew, T. Visual Expertise Is More Than Meets the Eye: An Examination of Holistic Visual Processing in Radiologists and Architects. *J. Med. Imaging* **2023**, *10*, 015501. [CrossRef] [PubMed]
49. Cruz, M.B.; Rebelo, F.; Cruz Pinto, J. Eye-Tracking Advancements in Architecture: A Review of Recent Studies. *Buildings* **2025**, *15*, 3496. [CrossRef]
50. Mouratidis, K. Aesthetic preferences for traditional versus contemporary architecture: Cross-national evidence from real-world buildings. *Cities* **2026**, *179*, 107410. [CrossRef]
51. Yao, Y.; Dall'Ò, G.; Lu, F. Urban Street-Scene Perception and Renewal Strategies Powered by Vision–Language Models. *Land* **2026**, *15*, 244. [CrossRef]
52. Li, Y.; Lu, J.; Meng, Y.; Luo, Y.; Ren, J. Exploring Urban Spatial Quality Through Street View Imagery and Human Perception Analysis. *Buildings* **2025**, *15*, 3116. [CrossRef]
53. Kuang, B.; Yang, H.; Jung, T. The Impact of Visual Elements in Street View on Street Quality: A Quantitative Study Based on Deep Learning, Elastic Net Regression, and SHapley Additive exPlanations (SHAP). *Sustainability* **2025**, *17*, 3454. [CrossRef]
54. Sharma, A. Urban greenways: Operationalizing design syntax and integrating mathematics and science in design. *Front. Archit. Res.* **2015**, *4*, 24–34. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.