land

MDPI

Article

A Scalable Framework for Street Interface Morphology Assessment via Automated Multimodal Large Language Model Agents

Yuchen Wang, Yu Ye * and Chao Weng

College of Architecture and Urban Planning, Tongji University, Siping Road, Shanghai 200092, China
* Correspondence: yye@tongji.edu.cn

Abstract

Evaluating street interface morphology is essential for urban design, yet existing approaches often struggle to combine large-scale applicability with higher-level morphological interpretation. This study proposes a scalable framework for assessing street interface morphology using an automated multimodal large language model (MLLM) agent. Using street view imagery (SVI), the framework evaluates four core morphological dimensions—enclosure, continuity, transparency, and roughness–through two complementary analytical streams: objective geometric measurement and subjective morphological assessment. To support reliable evaluation, the framework incorporates a dual-benchmark strategy consisting of manually derived geometric measurements and expert-consensus ratings for calibration and validation. Applied in Shanghai, the framework demonstrated reliable performance across the evaluated dimensions. The optimized agent was further extended to continuous street-segment analysis, demonstrating its applicability to large-scale urban assessment. By integrating objective and subjective evaluation within a scalable and interpretable workflow, the proposed methodology provides a practical tool for street interface morphology analysis and urban design assessment.

Keywords: street interface morphology; multimodal large language models (MLLMs); agent-based workflow; street view imagery (SVI); morphological evaluation; streetscape analysis

Check for updates

Academic Editor: Thomas Panagopoulos

Received: 10 February 2026
Revised: 19 March 2026
Accepted: 27 March 2026
Published: 8 April 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

1.1. Street Interface Morphological Evaluation

Streets serve as primary spaces for daily life and social interaction, shaping the structural framework of urban environments and directly influencing the quality of urban living [1,2]. As the spatial boundaries between buildings and public streets, street interfaces represent one of the most perceptually significant elements of urban form. They influence how pedestrians perceive enclosure, continuity, transparency, and visual complexity within urban environments. These morphological characteristics affect not only aesthetic qualities but also broader urban design outcomes related to walkability, spatial legibility, and the visual experience of urban space [3,4]. Empirical studies have shown that interfaces with higher transparency and continuity tend to encourage walking activities and enhance perceived safety [5,6]. Early urban design scholars such as Jan Gehl also emphasized the importance of active and permeable ground-floor edges in supporting social activities

 https://doi.org/10.3390/land15040610

along streets [7]. Similarly, Norberg-Schulz highlighted how spatial boundaries and architectural form shape human perception and the experiential meaning of urban space [8]. Consequently, the systematic assessment of street interface morphology has long been a central topic in urban design and planning research.

Traditional approaches to evaluating street interfaces have largely relied on field surveys and expert assessments [4]. Urban designers typically conduct visual inspections to interpret morphological qualities such as enclosure using indicators like the width-to- height metric (*D/H*) and façade continuity [9]. Although these approaches provide nuanced interpretations grounded in professional expertise, they are time-consuming, difficult to replicate at large spatial scales [10], and often subject to variability between evaluators. With the deepening of urban research, the need for scalable and reproducible methods for large-scale street morphology assessment has become increasingly important.

At the same time, urban design research has increasingly emphasized the importance of human perception in evaluating street environments [11–14]. A representative example is the framework proposed by Ewing and Handy, which links perceptual qualities such as imageability, enclosure, and complexity with measurable spatial indicators, demonstrating how objective urban form can be related to perceptual experience [15]. Beyond static visual appraisal, Cullen's concept of serial vision suggests that urban space is also perceived sequentially through movement, as spatial qualities unfold progressively along a path [16]. However, many existing methods struggle to support perceptual evaluation in a scalable and systematic manner. Experimental approaches, such as physiological sensing or eye-tracking, can capture detailed perceptual responses, yet they remain difficult to implement at large spatial scales, limiting their applicability for systematic urban analysis [14,17].

1.2. Street-View Imagery and Computer Vision Approaches

Recent advances in geospatial technologies and computer vision have created new opportunities for automated urban morphological analysis. Geographic information systems (GISs), remote sensing, and street-view imagery (SVI) now allow researchers to derive measurable indicators of urban form at increasingly large spatial scales. In particular, widely available platforms such as Google Street View and Baidu Street View provide extensive visual records of urban streetscapes from a pedestrian perspective [13,18–21]. Using these data sources, computer vision models can detect urban elements such as buildings, vegetation, sidewalks, and signage, enabling large-scale analyses of streetscape characteristics.

Despite these advances, existing approaches often face a fundamental limitation. Most SVI-based methods rely on extracting visual elements through image segmentation, pixel-level classification or object detection. Deep convolutional neural network architectures (DCNNs), such as DeepLab, PSPNet, SegNet, and FCN, are commonly used to perform pixel-level semantic segmentation of streetscape elements [22], including façades, vegetation, sky, and ground [23–28]. While these techniques effectively quantify the presence of urban elements, they often struggle to translate such information into higher-level morphological interpretations that reflect how urban designers understand street interfaces [12,29–32]. They also have difficulty interpreting nuanced interface types or complex spatial configurations, such as façades partially occluded by vegetation or overlapping architectural elements [31]. Consequently, conventional element-based visual models often struggle both to capture the higher-order spatial relationships underlying street interface morphology and to approximate human perception-based evaluation of the same morphological attributes.

Another research direction predicts human perceptions of urban environments using machine learning models trained on survey data. These studies estimate perceptual attributes such as safety, beauty, or liveliness from street images [33–43]. While these studies have expanded perception-oriented urban analysis, they often rely on black-box prediction and provide limited interpretability in relation to specific morphological attributes [12,30,44].

1.3. Multimodal Large Language Models and New Possibilities

Recent advances in artificial intelligence, particularly the emergence of multimodal large language models (MLLMs), offer new possibilities for addressing these challenges [45,46]. By combining the reasoning of large language models with the visual perception of vision-language models (VLMs), MLLMs interpret complex visual scenes through natural-language-guided reasoning. Furthermore, they demonstrate strong zero-shot abilities, aligning spatial visual features with professional descriptions without task-specific retraining [47]. Beyond pixel-level recognition, MLLMs can leverage contextual cues in street-view imagery to infer occluded elements and interpret higher-level spatial relationships. Recent studies further suggest that such models can approximate aspects of expert judgment by drawing on extensive pre-trained knowledge [48–51].

When embedded within automated agent-based workflows, MLLMs can further decompose complex analytical tasks, invoke external tools when required, and perform multi-step reasoning through "reasoning and acting" (ReAct) processes, generate structured data outputs, and enable iterative refinement of functions and workflows. This makes them particularly promising for street interface analysis, where scalable application depends not only on visual interpretation but also on workflow integration and controllable reasoning. Also, MLLM-based agents provide a potential bridge between low-level visual features and higher-level morphological interpretation of street interface.

Despite the potential of multimodal AI, several limitations remain in existing research for street interface analysis.

First, most existing large-scale SVI studies still rely on conventional computer vision pipelines rather than directly leveraging the native visual reasoning capabilities of new multimodal large language models such as Gemini. In many cases, street-view images are first processed through segmentation models such as DeepLab, and the resulting structured outputs are then passed to large language models for further analysis. This indirect workflow may discard important semantic and spatial information that is embedded in the original visual scene.

Second, many existing analytical workflows remain fragmented across multiple platforms, which reduces their operational coherence and limits their broader applicability. Researchers often need to coordinate among multiple platforms, including SVI APIs, GIS databases, and specialized computer vision libraries, while also managing repeated data conversion and transfer across these systems. Such fragmentation increases technical complexity, lowers analytical efficiency, and makes large-scale streetscape analysis more difficult to deploy in a streamlined and widely usable manner.

Third, many existing approaches remain heavily dependent on conventionally trained deep-learning models and labeled datasets, which constrains their transferability across different urban contexts. Models calibrated for specific cities, streetscape types, or annotation schemes often generalize poorly to environments with different spatial textures, interface configurations, or architectural forms. As a result, cross-scenario application usually requires additional data preparation, retraining, or model adaptation, rather than taking advantage of the broader zero-shot generalizability offered by multimodal large language models [44].

In addition, many automated evaluation pathways still operate as "black-box" systems. Some end-to-end models can produce efficient predictions through statistical fitting but lack interpretability, making it difficult to trace how specific spatial features contribute to evaluation outcomes. Although recent multimodal AI studies demonstrate strong scene understanding capabilities, most applications still focus on broad social indicators rather than the refined morphology of street interfaces [46,52]. Consequently, a structured and interpretable framework for the systematic evaluation of street interface morphology using MLLMs remains largely absent.

Finally, the use of such multimodal agent-based workflows in urban morphological analysis remains limited. As a result, the potential of MLLM-based agents for scalable, interpretable, and workflow-integrated street interface analysis remains underexplored.

1.4. Research Scope and Questions

In this study, street interface morphology refers to the observable form of the street edge at the pedestrian level. It is treated as the boundary of the street canyon, formed by building fronts, boundary structures, or street-edge vegetation. Elements within the street space itself, such as median trees, are not included. Four represented indicators—Enclosure, Continuity, Transparency, and Roughness—were selected as the fundamental features of street interface, which will be explained in detail later. In this study, the term "automated MLLM agent" refers to a task-oriented tool deployed on an AI platform that can automatically execute a predefined multi-step analytical workflow once user parameters are specified, including street-view image acquisition, rule-guided perception, structured output generation, and iterative refinement through calibration and validation against reference values.

To address the challenges identified above, this study proposes an automated MLLM-agent workflow for assessing street interface morphology using street-view imagery. The framework integrates two complementary assessment dimensions—the objective method and the subjective method. These two analytical streams are integrated through a dual-benchmark validation strategy, which compares objective outputs with field or satellite measurements and subjective results with expert consensus.

This study addresses three main research questions:

Q1: Can the proposed MLLM-agent workflow effectively assess street interface morphology from street-view imagery, through both objective geometric measurement and subjective perception-based morphological evaluation?

Q2: How can the evaluation capability of the MLLM agent be optimized within the proposed workflow, for example through prompt design or calibration strategies?

Q3: Can the proposed workflow support scalable and continuous street-level analysis while maintaining interpretability for urban design applications?

2. Materials and Methods

2.1. Analytical Framework

To address the limitations of previous workflows, this study develops a multimodal agent- based framework for large-scale assessment of street interface morphology from street-view imagery. The proposed framework is designed as a transferable analytical system supported by a dual-benchmark calibration mechanism. It incorporates two complementary evaluation streams—objective geometric measurement and subjective perception-based morphological assessment—and uses the dual-benchmark mechanism to evaluate and refine the agent's capability in both streams (Figure 1).

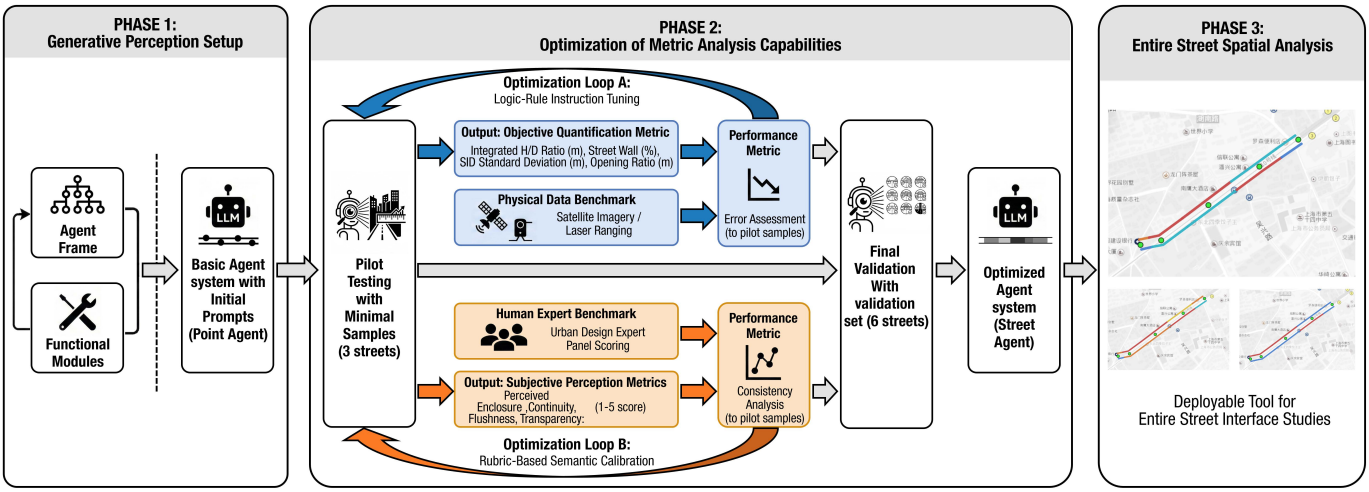


Figure 1. Methodological framework of the proposed street interface assessment workflow. Gray arrows indicate the overall progression of the workflow, while the blue and orange curved arrows represent the iterative optimization loops for the objective and subjective evaluation streams, respectively. Chinese labels appearing in the embedded map are place names from the original basemap. Source: Authors' elaboration.

In this study, objective metrics refer to the actual geometric values derived from spatial measurements, whereas subjective metrics represent perceived morphological characteristics rather than broader psychological states such as comfort, safety, or liveliness. Rather than linking physical form to affective or experiential responses, this study treats geometric measurement and perception-based judgment as two complementary ways of characterizing the same morphological attributes of street interfaces. This distinction is important because the framework aims to assess street interface morphology itself, rather than its secondary psychological or emotional effects.

Conventional SVI-based approaches typically rely on pixel-level feature extraction and visual element classification. While such techniques can identify urban components, they often capture only surface-level relationships between elements. In contrast, the proposed framework introduces a multimodal large language model (MLLM) as a cognitive reasoning engine capable of interpreting the semantic structure of street interfaces. This capability allows the system to infer higher-level morphological dimensions, such as enclosure and continuity, that are commonly used in urban design analysis.

Building on the generalized reasoning ability of multimodal LLMs, this study further integrates an agent-based architecture to construct an automated urban analysis agent. The agent is designed to replicate key steps of a professional street interface survey conducted by urban planners. It automatically samples street-view imagery and performs a dual assessment of street morphology.

First, the agent estimates objective geometric indicators by interpreting spatial relationships within the street scene. Second, it generates subjective perception-based morphological scores by approximating expert-like visual judgment of the same morphological attributes. These two analytical processes complement each other: objective measurements capture measurable geometric properties, while subjective evaluations represent how the same morphological attributes are visually assessed. Together, they provide a more comprehensive description of street interface morphology.

The overall workflow establishes a progressive research path consisting of three stages: "single-point setup—dual-benchmark calibration—continuous street analysis." This strategy first validates and optimizes the agent's analytical capability through controlled experiments at discrete sampling points. It then extends the workflow to continuous street

segments for large-scale analysis. Finally, the optimized model is applied to representative areas in Shanghai to test its applicability and to compare street interface characteristics across different urban contexts.

In detail, the overall analytical framework comprises the following three phases:

Phase 1: Automated visual sampling and generative perception setup.

This phase establishes the baseline infrastructure of the MLLM agent. It integrates a standardized street-view data acquisition module with a foundational morphological analysis engine powered by the multimodal large language model (MLLM).

The objective of this phase is to initialize the perceptual capability of the agent, enabling the automated transformation of unstructured street-view imagery into structured morphological indicator data. These indicators correspond to the four core street interface dimensions: Enclosure, Continuity, Roughness, and Transparency, which are commonly adopted in street morphology studies as important indicators [15,32].

Phase 2: Optimization of metric analysis capabilities.

To bridge the gap between generative AI-based interpretation and professional urban measurement, a series of calibration experiments were conducted. Because objective geometric measurements and subjective perceptual evaluations may produce divergent interpretations of the same street interface, both dimensions were integrated into the analytical framework to better capture the complexity of street morphological characteristics [53].

A ground-truth dataset covering 9 street segments across three representative morphological zones was constructed to support a dual-feedback optimization process:

1. Loop A (Objective metrics): This loop iteratively refines the agent's geometric reasoning capability, for example by correcting *H/D* ratio estimation through comparison with field-measured spatial data.
2. Loop B (Subjective metrics): This loop aligns the agent's semantic scoring logic with the evaluations provided by a panel of expert urban designers.
3. Validation and Scalable Extension: After verifying the stability and professional reliability of the workflow, the optimized analytical core was extended into a comprehensive street-analysis agent equipped with spatial visualization capabilities.

Phase 3: Continuous Street Spatial Analysis.

In the final phase, the optimized street analysis agent was deployed across the three representative urban zones. The agent conducted quantitative assessments of continuous street segments using sampling points generated at user-defined spatial intervals.

By analyzing the resulting visualized metric polylines, the system produced quantifiable evidence of spatial variations in street interface morphology. This analysis reveals the distinctive interface characteristics of different urban contexts, while also highlighting the spatial differences between objective geometric measurements and subjective perceptual evaluations of street interface indicators.

2.2. Study Area and Data Source

This study selects Shanghai, China, as the empirical case study area. As a high-density metropolis with a complex urban fabric, Shanghai exhibits a wide diversity of street interface morphologies. This diversity provides an appropriate testing environment for evaluating the adaptability of the proposed agent across multiple urban scenarios.

To ensure methodological applicability across different urban contexts, three typical zones representing distinct street morphology types were selected as the empirical sampling framework. Within each zone, a Zone–Street–Point hierarchical sampling strategy was established. Specifically, three representative street segments were selected

within each zone, resulting in a total of nine streets. Along each street segment, four key sampling points were identified, leading to a total of 36 sampling points. This sampling strategy balances analytical efficiency with analytical coverage, while ensuring that different street interface conditions are adequately represented (Figure 2). The geographic coordinates of the experimental sampling points are provided in Appendix A, Table A1.

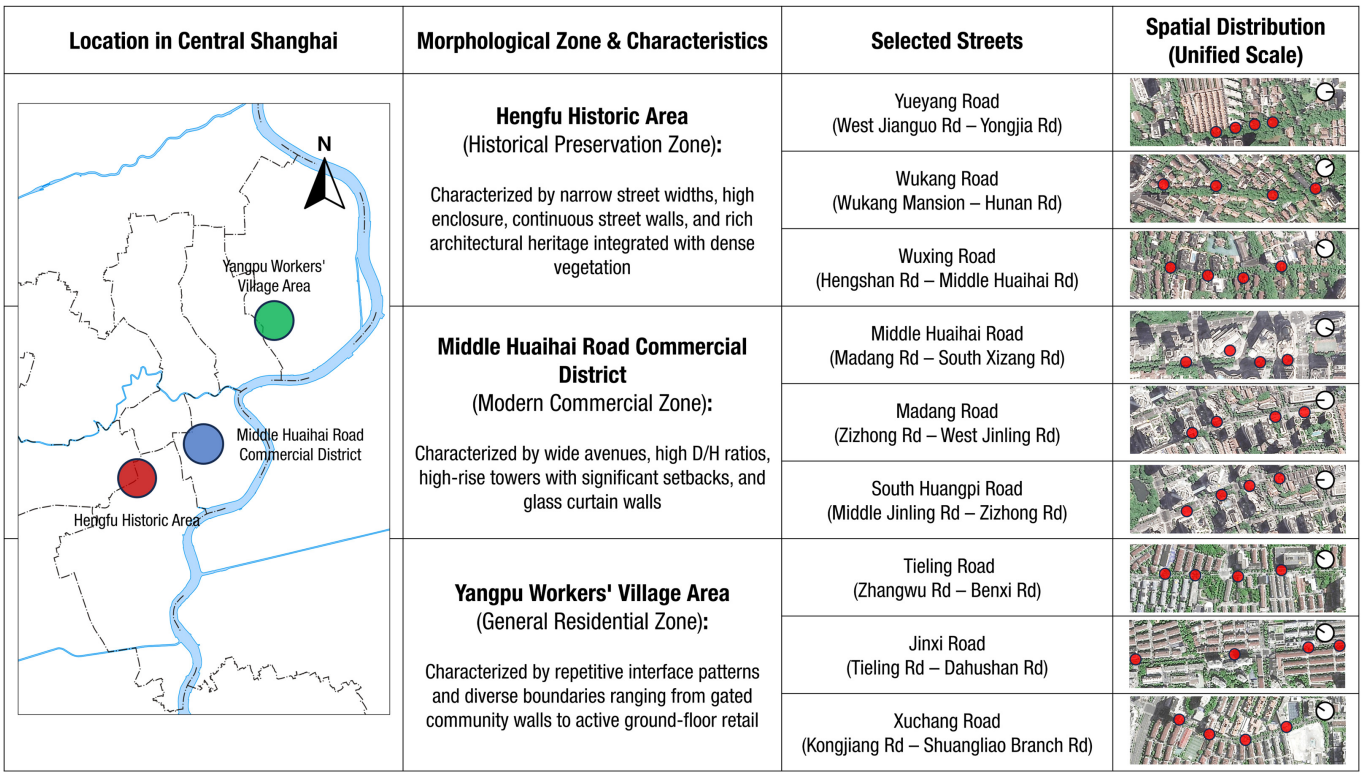


Figure 2. Overview of the study area and the Zone–Street–Point hierarchical sampling strategy. Source: Authors' elaboration.

For the subsequent calibration and validation experiments, the 9 sampled streets were divided into two subsets: one street segment from each zone (3 streets with 12 sampling points in total) was assigned to the pilot subset, and the remaining 6 street segments (24 sampling points) were reserved as an independent hold-out dataset.

The three zones and their corresponding sampling streets are summarized as follows:

1. Historical Preservation Zone (Hengfu Historic Area). This zone includes Yueyang Road, Wukang Road, and Wuxing Road. These streets are characterized by narrow widths, strong spatial enclosure, and complex interactions between historic buildings and dense vegetation. Such characteristics make the identification of street interfaces difficult when relying solely on satellite imagery, highlighting the importance of street-level observations.
2. Modern Commercial Zone (Middle Huaihai Road Commercial District). This zone includes Middle Huaihai Road, Madang Road, and South Huangpi Road. Streets in this district are characterized by wide avenues, high-rise commercial towers with significant building setbacks, and extensive glass curtain-wall façades, creating a spatial scale markedly different from that of the historical area.
3. General Residential Zone (Yangpu Workers' Village Area). This zone includes Tieling Road, Jinxi Road, and Xuchang Road. The street interfaces here exhibit relatively repetitive morphological patterns, with diverse boundary conditions ranging from gated residential walls to active ground-floor commercial frontage.

The selection of sampling points prioritized morphological representativeness rather than strict equidistant spacing. Redundant points located within highly repetitive interface segments were intentionally avoided, while points representing distinct spatial characteristics were retained for analysis.

To acquire street-level imagery, we used Baidu Map as the street-view source and matched each sampling point to the corresponding available street-view scene based on its geographic coordinates. The resulting standardized image sets served as the primary visual input for the perception module of the proposed agent.

2.3. Multi-Dimensional Morphological Evaluation Model

To implement the analytical framework described above, a multi-dimensional evaluation system was constructed. This system evaluates street interface morphology through two complementary dimensions: objective geometric metrics and subjective perceptual assessments. The following sections describe the construction of the generic street interface model and the corresponding measurement rules.

2.3.1. The Generic Street Interface Model

To provide a unified morphological basis for measurements, a generic street interface model was defined (Figure 3). This model simplifies the complex street environment into computable spatial elements, identifying key parameters such as Effective Street Width (D), Interface Height (H), and Street Interface Depth (*SID*). It also evaluates a continuous three-dimensional spatial segment rather than a single two-dimensional cross-section or plan layout. Within each sampling segment, the spatial arrangement and combination of diverse elements potentially influence the measurement of different metrics.

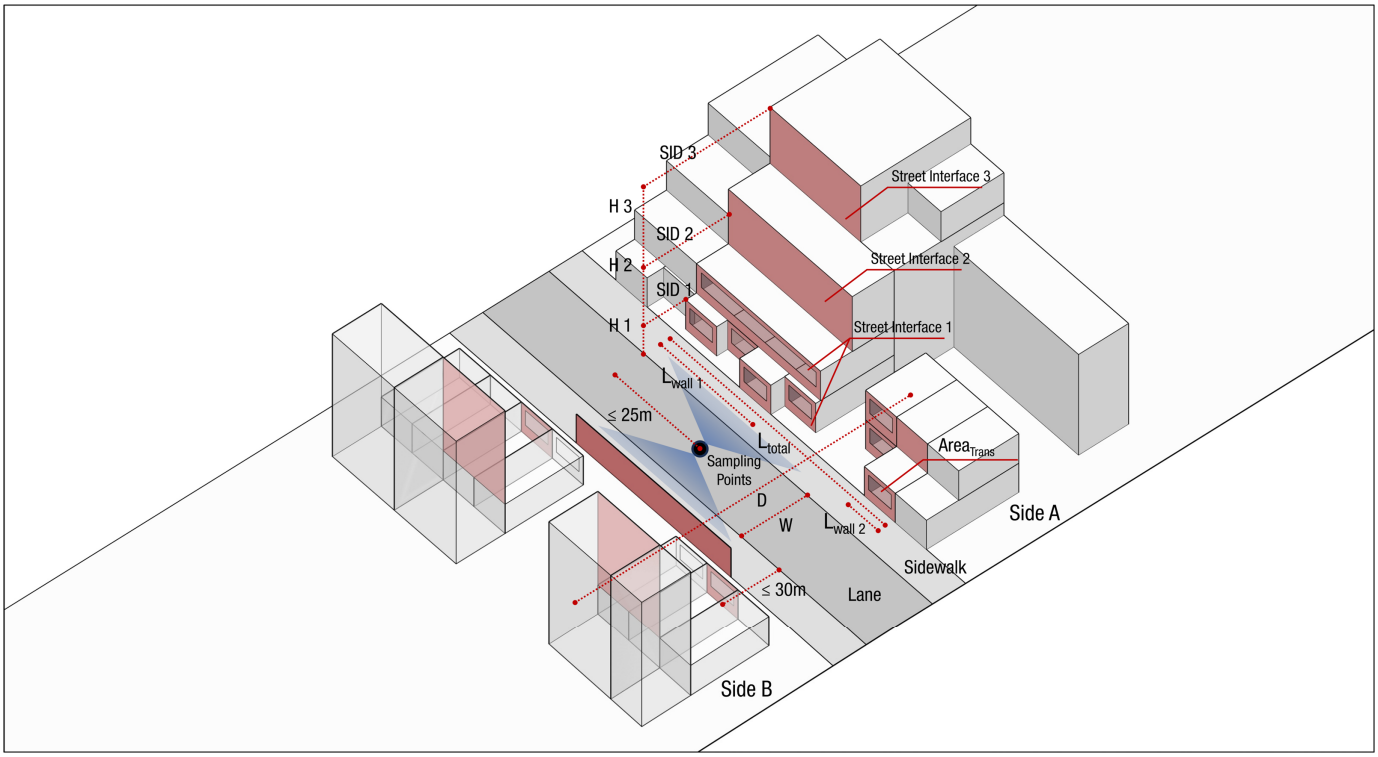


Figure 3. Schematic diagram of the generic street interface model. The effective calculation segment corresponds to the visual range of the sampling point, defining the spatial scope for metric calculation. Valid interfaces inside the effective calculation segment are marked in red. Source: Authors' elaboration.

Considering the typical characteristics of street interfaces in Chinese cities, a set of operational rules was established to standardize how these parameters are measured. The following rules define spatial scope, visual validity, and metric calculation procedures.

1. Effective calculation segment: To balance physical environment and human perception belonging to individual sampling points, calculations are not based on the entire street length but are restricted to the visual interface segment. This rule also helps reconcile different measurement logics. Traditional D/H is based on a cross-sectional view of the street, whereas human perception is shaped by a continuous street façade. This segment covers the interface extending approximately 25 m to the front and back of the sampling point. Only interfaces within this visually effective range are included in the calculation; distant or visually obscured elements are excluded. L_{total} is defined as the total length of this sampling segment along the street, normally at 50 m (double 25 m).
2. Definition of valid interface: A vertical element is counted as a valid interface if it provides a clear and tangible visual boundary that can be seen from street view. In addition, this validity is constrained by spatial proximity. Elements positioned beyond the effective visual threshold, where their capacity to enclose the street diminishes due to excessive setback from the curb line (over 30 m), are excluded from the calculation. Considering the Chinese urban context, this includes not only building facades but also boundary structures (including both solid masonry and permeable walls) and Structural Vegetation (specifically dense, continuous hedges that act as visual screens) exceeding 1.5 m in height. Individual street trees and sparse landscaping are excluded. The effective wall length (L_{walls}) is defined as the sum of the lengths of these valid continuous elements within the segment.
3. Handling spatial recesses: We distinguish between valid interfaces and spatial gaps based on scale. If a recess has a depth proportionate to the street scale and maintains a sense of enclosure, it is treated as a valid interface. Conversely, if the recess is too deep relative to the street scale, creating a perceived void, it is classified as a gap (discontinuity).
4. Multi-layered interfaces: This identifies up to three potential depth layers vertically per side. Unlike previous studies that measure setbacks from the street centerline, this study defines Street Interface Depth (*SID*) as the perpendicular distance from the curb line (the edge of the vehicle lane) to the dominant vertical face. This definition is easier for the agent to apply and better reflects the pedestrian perspective. This ensures the metric reflects the pedestrian's experience on the sidewalk. If no valid interface is detected on a specific side, the corresponding *SID* value is calculated as the sidewalk width. For the Maximum Interface Height (H_{max}) on a single side used in enclosure calculations, the maximum height among all identified layers (H_1, H_2 and H_3) is adopted. Notably, effective street width (D) used for enclosure calculation is defined as the sum of the road width (W) and the setback distances (*SID*) of the highest interface layers detected on both sides:

$$D = W_{ROAD} + SID_{A,Hmax} + SID_{B,Hmax} \quad (1)$$

5. Layer merging: To address minor architectural articulations, adjacent layers are merged when their depth difference is less than 2.0 m (for example, between the ground floor and the second floor). In such cases, the layers are treated as a single unified interface layer for morphological analysis.

2.3.2. Evaluation Metrics and Calculation Methods

The use of both objective and subjective metrics within the same morphological dimension is theoretically justified. Although the two metrics share identical conceptual definitions and spatial scopes, they capture different aspects of street morphology. Objective measurements rely on geometric rules and quantitative formulas, whereas subjective evaluations capture perceptual effects that may not be fully represented by purely mathematical calculations.

Based on the generic street interface model, the four core dimensions are evaluated as follows. The detailed calculation methods and scoring criteria are summarized in Table 1. Enclosure considers all sides of the street to capture the overall spatial atmosphere, while the other metrics only consider a single side of the street.

Table 1. Summary of objective and subjective metrics for street interface evaluation. Source: Authors' elaboration.

Core Dimension	Interface Spatial Scope	Objective Metric	Calculation Formula	Subjective Metric	Scoring Standard (Likert 1–5)
Enclosure (Overall)	Maximum interface height	Integrated *H/D* Ratio	$R_{enclo} = \frac{H_{max,A}+H_{max,B}}{2\times D}$	Perceived spatial envelope	1 (Open)–5 (Enclosed)
Continuity (Single-sided)	All interface above 1.5 m	Street Wall Ratio (%)	$R_{cont} = \frac{L_{wall}}{L_{total}} \times 100\%$	Perceived spatial continuity of street interface	1 (Fragmented)–5 (Continuous)
Transparency (Single-sided)	0–5 m in first interface layer	Opening Ratio (%)	$R_{trans} = \frac{Area_{trans}}{Area_{total}} \times 100\%$	Perceived visual permeability of street interface	1 (Opaque)–5 (Transparent)
Roughness (Single-sided)	First interface layer	SID Standard Deviation	$V_{Rough} = \sqrt{\frac{\sum(SID_i-\overline{SID})^2}{N}}$	Perceived roughness of street interface	1 (Aligned)–5 (Staggered)

1. Enclosure (Overall, or double sides):

Spatial scope: This dimension evaluates the degree of spatial enclosure of a street section, defining the envelope and sky view limits. The evaluation scope adopts the Maximum Interface Height (H_{max}) on each side.

Objective metric: Measured by the Sectional *H/D* Ratio (Height-to-Width Ratio), which inverts the traditional *D/H* metric to maintain a positive correlation with the sense of enclosure and to avoid mathematical singularities when the interface height approaches zero. If the height exceeds the visual field and cannot be estimated, indicating an excessive vertical dimension, the *H/D* ratio is capped at 4.0.

Subjective metric: Perceived spatial envelope (1–5). It evaluates the psychological sense of spatial containment.

2. Continuity (Single-sided):

Spatial scope: This dimension focuses on the horizontal spatial integrity of the street boundary on both sides. The scope includes all constructed vertical elements that act as visual barriers above 1.5 m, regardless of their total height.

Objective metric: Measured by Street Wall Continuity (%). It is calculated as the ratio of the length of continuous valid vertical elements (L_{walls}) to the total sampling length (L_{total}) within the visual field. Note that spatial gaps and vacant lots are included in the denominator L_{total} as negative factors, effectively reducing the continuity ratio.

Subjective metric: Perceived spatial continuity of the street interface (Likert 1–5). It assesses the coherence of the street walls, penalizing fragmentation caused by vacant lots, abrupt gaps, or incoherent interface transitions that disrupt the visual flow.

3. Transparency (Single-sided):

Spatial scope: This is restricted to the 0–5 m Pedestrian Zone. It measures the visual permeability of the street interface separating street space and plot. It is not limited to seeing into buildings, but also seeing into deep spaces, such as courtyards, gardens, or lobbies.

Objective metric: Measured by the Interface Opening Ratio (%). This is measured by the area ratio of transparent elements (e.g., glass windows, glazed doors, open fences) attached to the street walls. Note that structural voids, such as open passageways (tunnels), recessed entrances without doors, or gaps between buildings, are excluded from this metric.

Subjective metric: Perceived visual permeability of the street interface (Likert 1–5). This reflects the perceived degree of visual openness and spatial penetration across the street-edge boundary between the public realm and adjacent private or semi-public spaces, with higher scores indicating a stronger sense of permeability.

4. Roughness (Single-sided):

Spatial scope: Measured by the spatial stagger of the street wall. To capture the alignment of the street wall, this dimension strictly targets the first interface layer as the primary street-defining surface. This evaluates the degree of stagger, depth variation, and unevenness of several interfaces from different buildings.

Objective metric: Measured by the *SID* Standard Deviation, where N represent the number of facades counted. A higher standard V_{Rough} indicates a staggered interface, while a lower value indicates high alignment.

Subjective metric: Perceived roughness of the street interface (Likert 1–5). It evaluates the degree of visual irregularity and façade variation along the street-side interface, with higher scores indicating more fragmented, uneven, and morphologically complex frontage conditions.

To validate the proposed agent's performance, a dual-benchmark validation framework was established. Table 2 details the distinct data acquisition methods used to construct both the ground truth baseline and the agent-based estimation.

Table 2. Data acquisition methods and measurement methodologies for the dual-benchmark framework. Source: Authors' elaboration.

Measurement Dimension	Ground Truth	Agent-Based Estimation
Objective metrics	Satellite measurement and manual field survey	Direct estimation guided by geometric definitions and formulas
Subjective metrics	Expert scoring panel	Evaluation based on indicator prompts

For objective metrics, the ground truth serves as the physical reference, derived from satellite measurement and manual field surveys to ensure accuracy. In contrast, the agent was instructed to estimate the final quantitative values under formula-guided geometric reasoning based on the metric definitions in Table 1, rather than through a fully externalized parameter-to-formula computation pipeline.

For subjective metrics, the ground truth was established through an expert scoring panel, representing the consensus of professional urban designers. The agent-based estimation approximates this expert subjective judgment process through evidence-based semantic reasoning. Rather than generating generic descriptions, the agent identifies specific visual cues defined in the evaluation rubric (such as identifying "glass curtain walls" versus "solid masonry") and uses them to infer the corresponding morphology-based scores. In this way, the

agent simulates the rubric-based evaluation process used by human experts while producing structured scores with one-decimal precision.

2.4. Agent Construction and Implementation Mechanism

2.4.1. Implementation Framework

To operationalize the proposed measurement methods, we constructed a customized urban analysis agent using the new-generation Google Gemini model (gemini-3-flash-preview). The implementation framework consists of two core components: an Interactive Visualization Module for configuration, visualization, and result inspection, and an Automated Execution Engine for scalable and controlled automated processing.

1. Interactive Visualization Module
   Serving as the primary user interface for researchers (Figure 4), this module integrates two functional panels that support interactive analysis and validation:
   Analysis Center (left panel): This panel facilities task configuration, such as setting sampling interval (e.g., 50 m), setting start point to end point of sampling path. Centrally, the system integrates a dynamic GIS mapping engine powered by the Baidu Map API. This module precisely aligns sampling coordinates with the urban road network and visualizes the morphological analysis results as vector graphics on the map. Crucially, the lower section of this panel integrates a visualization parameter control module. This allows users to dynamically toggle between different display modes of the sampling path and the result graphics, ensuring that the GIS mapping output aligns with specific analytical needs. Additionally, a Heads-Up Display (HUD) overlay provides real-time monitoring, showing processes including sampling planning, image capturing and image analysis.
   Result Archive (right panel): Dedicated to quality control, this panel displays aggregated analyzed statistical summaries, including all the result values of objective and subjective metrics of measurement dimensions. Crucially, the intermediate outputs, ranging from the captured images to the extracted fundamental parameters and natural-language descriptions, can be explicitly rendered and logged, making the entire reasoning chain fully traceable. Each captured image can be downloaded. Also, this provides direct access from result of each point to the source sampling points on the map on the left panel, enabling researchers to cross-reference AI-analyzed results with actual street views, ensuring the reliability of the automated assessment.
2. Automated Execution Engine
   The system's core logic is governed by an automated engine that manages the ReAct (Reasoning + Acting) process, enabling complex logical reasoning and controllable execution. To maintain stability during street analyzing on a larger scale (e.g., processing hundreds of points), the engine implements a rigorous Finite State Machine (FSM). It enforces a sequential lifecycle for each sampling point—Initializing, Capturing, Analyzing, Cooling—while managing memory allocation and API rate limits. Simultaneously, the engine serves as the interface between the system and the Gemini model, packaging visual data into structured prompts and parsing the output into standardized JSON formats. After optimization, the processing time for each sampling point is significantly reduced, enabling efficient analysis of large-scale street networks.

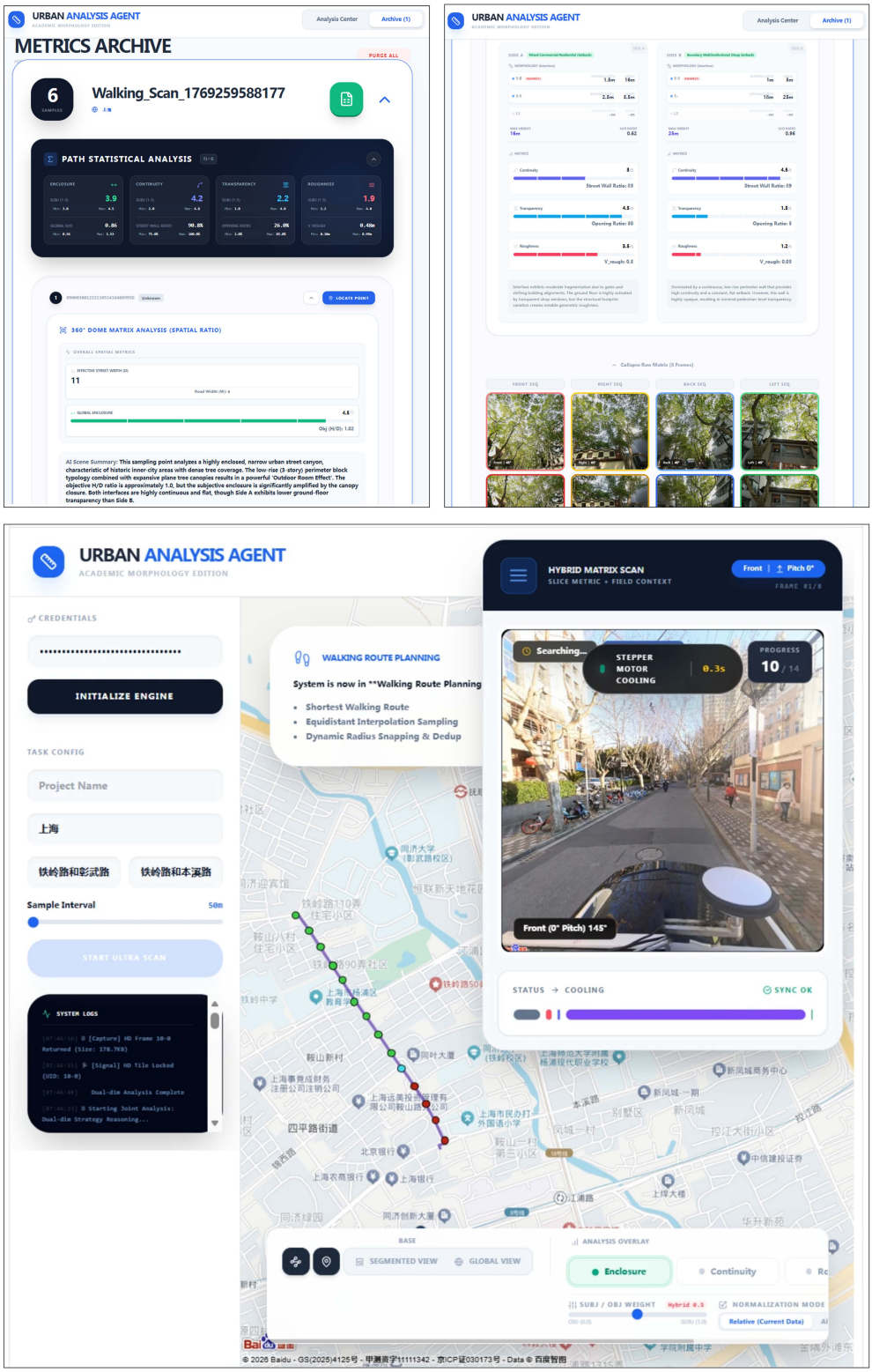


Figure 4. The interactive interface of the urban analysis agent. Chinese labels appearing in the embedded map are place names from the original basemap, while the Chinese text on the left in TASK CONFIG part denotes the input address entered by the user. Source: Authors' elaboration.

2.4.2. Operational Workflow

As illustrated in Figure 5, the agent operates through an end-to-end workflow. The overall detailed operational process is described below:

1. Step 1 (sampling planning): The agent first generates the shortest walking path based on the user-defined configurations, such as start and end points, and places sampling points at fixed intervals (e.g., 50 m). For each point, it then matches the nearest valid street-view scene available on Baidu Map within a predefined search radius, ensuring that only locations with available street-view imagery are processed.
2. Step 2 (image preparation): For each valid sampling point, the corresponding street-view scene from Baidu Map was loaded, and a standardized image set was generated from the scene for downstream analysis. These images are then aggregated into a unified visual matrix to serve as the input for the subsequent reasoning phase.
3. Step 3 (visual analyzing): The visual image matrix is packaged and transmitted to the MLLM (Gemini). Guided by the metric definitions, evaluation rules, formulas, and output requirements embedded in the prompt, the agent jointly analyzes and reasons the images and directly generates structured outputs, including objective indicator estimates, subjective perceptual scores, and supporting natural-language descriptions. For objective metrics, the MLLM estimates the quantitative indicators under formula-guided geometric reasoning. For subjective metrics, the MLLM uses prompt-defined visual cues to infer the corresponding subjective scores.
4. Step 4 (result parsing and structuring): The execution engine parses the model responses and converts them into a standardized JSON structure. The resulting records are then organized according to the four morphological dimensions, and prepared for downstream storage, and export.
5. Step 5 (spatial visualization): Finally, the structured results are transmitted to the GIS engine for real-time spatial visualization. The outputs are simultaneously logged in the results panel and exported as CSV files. Before processing the next sampling node, the system performs a brief memory reset and resource cleanup to maintain stable execution during large-scale analysis.

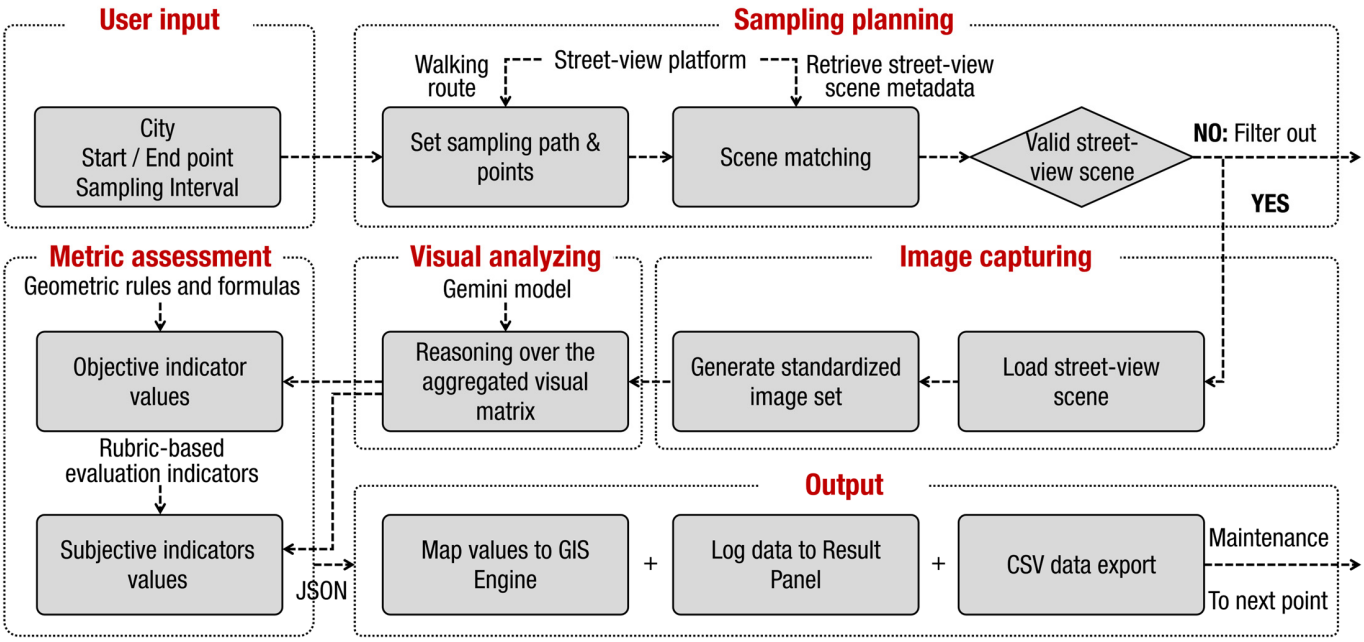


Figure 5. Operational workflow of the urban analysis agent. Dashed arrows indicate the direction of workflow progression and data transfer throughout the analytical process. Source: Authors' elaboration.

2.4.3. Visual Sampling Strategy and Joint Reasoning Mechanism

To enable the MLLM to perform joint reasoning on the street interface morphology, we adopted a visual matrix approach, which is sending several images of one sampling point to the MLLM as a batch. We evaluated various configurations, including different aspect ratios (e.g., 16:9, 1:1, 9:16), frame counts (e.g., 8, 12) and resolution, to balance information completeness, accuracy, and generation speed. Ultimately, we implemented an 8-frame Matrix configuration to maximize the Field of View (FOV) (1:1 aspect ratio, 1024 × 1024 resolution) covering two elevation tiers, ensuring that the full 360° panoramic scope of the surroundings is captured (Figure 6). This dual-tier approach ensures that necessary vertical elements, such as high-rise setbacks and tree crowns, and view are almost fully contained within the agent's viewport:

1. Tier 1 (0° Elevation): 4 images (Front, Back, Left, Right) to capture street-level to mid-level spatial information, including pedestrian-scale interfaces, pavement details, lower and mid-level façades.
2. Tier 2 (45° Elevation): 4 images (Front, Back, Left, Right) to capture mid-level to upper-level spatial information, including upper façades, upper setbacks, and the building skyline.

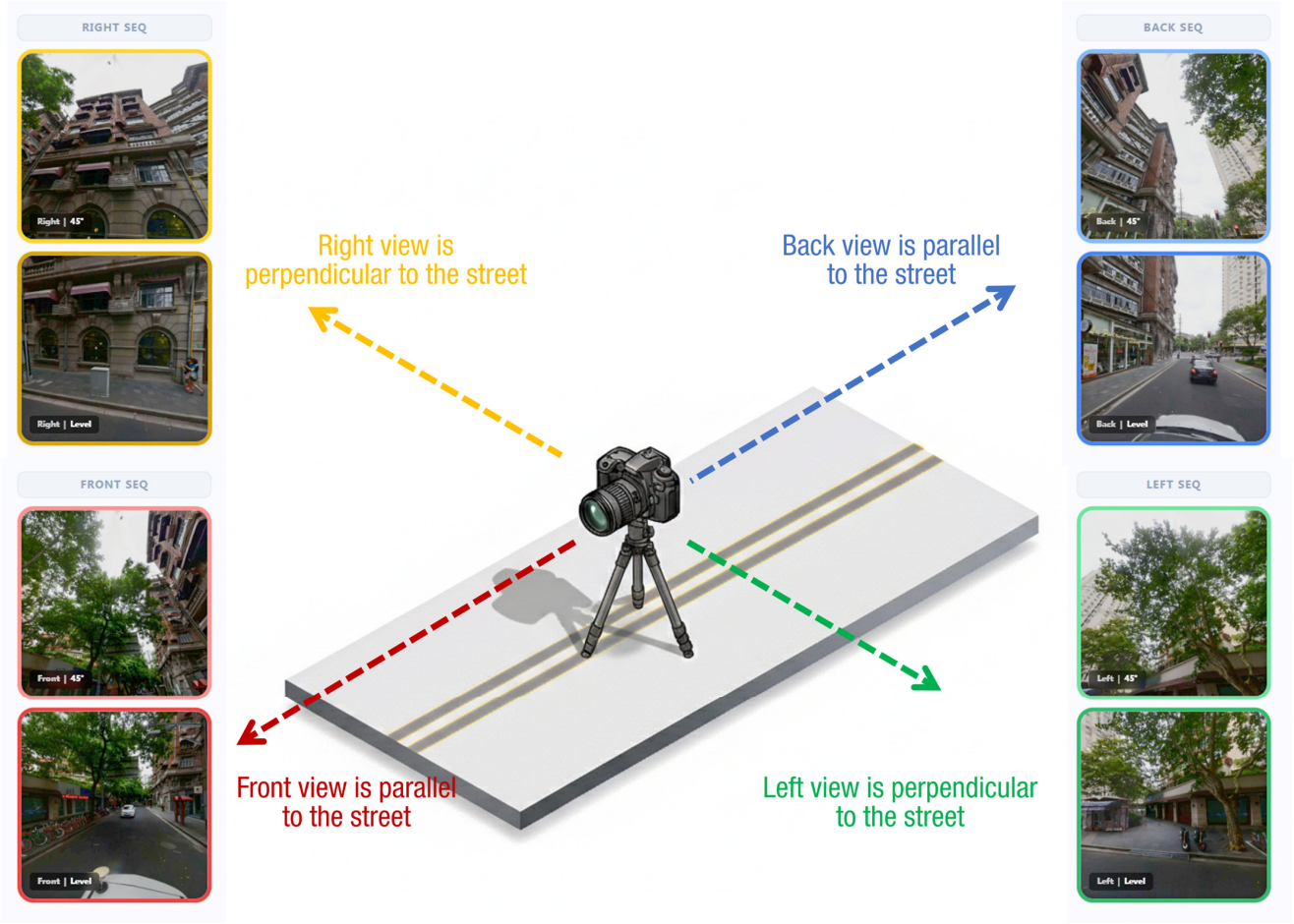


Figure 6. Visual sampling configuration for the 8-frame street-view image matrix. Source: Authors' elaboration.

Leveraging the MLLM's capability for joint reasoning and semantic alignment across multiple images, the agent processes the 8-frame matrix as a single, coherent visual context rather than analyzing individual images in isolation. This enables integrated reasoning over street interface morphology.

2.4.4. Agent Validation Experiment Design

To improve the reliability and consistency of the proposed agent, we designed a structured set of calibration and validation experiments. The strategy included internal diagnostic checks, dual-benchmark construction, pilot-based prompt refinement, independent hold-out testing, and a temporal robustness assessment.

1. Validation dataset and experimental split

The overall validation dataset comprised 36 sampling points from 9 street segments across 3 representative urban zones, following the Zone–Street–Point sampling framework described in Section 2.2. To support calibration and independent validation, the dataset was divided into two subsets. One street segment from each zone, corresponding to 12 sampling points in total, was assigned to the pilot subset for iterative prompt calibration and workflow optimization. The remaining 6 street segments, corresponding to 24 sampling points, were reserved as an independent hold-out dataset for final validation. This split was designed to separate iterative prompt calibration from final validation and to provide a more rigorous test of the agent's performance on unseen street samples.

2. Internal alignment and stability test

Before external benchmarking, a set of internal diagnostic checks was conducted to verify whether the workflow operated coherently as an analytical system. Specifically, six aspects were examined. Items 1–4 and 6 were examined during the pilot calibration stage, whereas Item 5 was calculated on the full sampled dataset to evaluate the overall internal relationship between the two analytical streams:

(1) Spatial consistency of image sampling, by checking whether the sampled image sequences followed the intended street route and viewing directions.
(2) Side-specific directional consistency, by checking whether features identified on each street side were mapped to the correct single-sided score outputs.
(3) Rule compliance of the objective metric stream, by examining whether the model-generated objective estimates were consistent with the geometric definitions, spatial scope, and formula-guided reasoning rules specified for each metric.
(4) Semantic compliance of the subjective metric stream, by examining whether prompt-defined evaluation indicators were reflected in the corresponding subjective scores.
(5) Internal correlation between the agent-generated subjective scores and the corresponding objective metrics, assessed using Spearman's rank correlation coefficient on the full sampled dataset (N = 36), as shown below:

$$\rho = \frac{\sum_{i=1}^{N} (R_i - \overline{R})(Q_i - \overline{Q})}{\sqrt{\sum_{i=1}^{N} (R_i - \overline{R})^2} \sqrt{\sum_{i=1}^{N} (Q_i - \overline{Q})^2}} \quad (2)$$

where R_i and Q_i represent the ranks of the agent-generated subjective score and the corresponding objective metric for the *i-th* sample, respectively, and $\overline{R}$ and $\overline{Q}$ denote their mean ranks.

(6) Repeated-run stability, assessed through five repeated executions on the pilot subset.

3. Construction of the objective validation benchmark

For the objective evaluation stream, the ground-truth benchmark was established as a physical reference dataset derived from satellite measurement and manual field survey. All fundamental geometric parameters required by the objective indicators were manually recorded for the sampled street points, and the corresponding benchmark values were calculated using the same spatial scope, interface definitions, and metric formulations described in Table 1. This benchmark served as the external reference for calibrating the

agent's geometric reasoning process and for evaluating the final objective measurement accuracy of the workflow.

4. Construction of the subjective validation benchmark and agreement screening

For the subjective evaluation stream, an expert-consensus benchmark was established through a professional online rating panel. Ten experts with backgrounds in urban design, including doctoral students, postdoctoral researchers, university faculty members, and practicing urban designers, independently rated the same 8-frame street-view image matrices used by the agent. Before the evaluation, all experts were provided with detailed scoring guidelines, including illustrated examples explaining the spatial scope and scoring criteria for each morphological dimension with text and figures. For each sampling point, the experts assessed seven subjective items derived from the four core dimensions, including overall enclosure and the single-sided continuity, transparency, and roughness. All ratings were completed using a five-point Likert scale.

During the rating process, the experts did not have access to the agent-generated outputs, the ratings of other experts, or the final consensus scores. No post-rating discussion or score revision was conducted before the agreement analysis. However, formal full blinding was not implemented, because the raters were aware of the study context and the evaluation task.

To assess the reliability of the expert benchmark, we calculated the within-group agreement of ten experts across 36 samples using the r_{WG} index from James, Demaree, and Wolf [54], due to the frequent occurrence of identical ratings among experts. The formula is as follows:

$$r_{WG} = 1 - \frac{S_x^2}{\sigma_E^2} \tag{3}$$

where S_x^2 represents the observed variance of expert scores, and σ_E^2 refers to the expected variance under a random distribution. The r_{WG} value was calculated separately for each sampling point and each subjective item. Agreement screening was conducted at the sample–item level rather than at the whole-sample level, allowing for the same sampling point to be retained for one metric but excluded for another. We adopted a threshold of $r_{WG} > 0.70$ as a conventional cutoff for acceptable within-group agreement in rating-based studies. To further summarize benchmark reliability at the sample–item level, an Agreement Rate was also calculated as the proportion of sample–item pairs with $r_{WG} > 0.70$ among all samples for each subjective item. A sample–item pair was retained only when the corresponding r_{WG} value exceeded 0.70. For each retained sample–item pair, the expert-consensus ground truth was defined as the mean score across the expert panel. Only these retained consensus scores were used in the subsequent prompt calibration and benchmark-based validation analyses.

5. Pilot calibration and prompt refinement

Using the pilot subset of 12 sampling points as learning anchors, the agent's outputs were compared with the corresponding objective and subjective benchmarks derived from the same image sequences.

For objective metrics, calibration focused on improving the extraction and interpretation of geometric parameters, with performance evaluated against the physical benchmark using error-based measures. To evaluate the precision of the agent's objective estimations against ground truth data, the Root Mean Square Error (*RMSE*) was employed as the primary metric, as shown in the formula

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4}$$

where $\hat{y}_i$ represents the *i-th* precision value from the agent and y_i represents the real value measured on real site and from satellite map. In the *H/D* ratio, the maximum value is set at 4.0, since the agent struggles to recognize the exact height of a skyscraper from the 45° view in a narrow street, and also when reaching 4.0, it can already be regarded as being totally enclosed. To further assess the practical reliability of the data beyond error magnitude, we introduced the Qualified Rate (*Acc*) to measure the proportion of samples that fall within acceptable tolerance thresholds. The formula

$$Acc = \frac{Count(|\hat{y}_i - y_i| \leq T_i)}{N} \times 100\% \quad (5)$$

was used to calculate the percentage of samples where the agent's estimation error falls within the acceptable corresponding threshold (T_i). For Enclosure, Continuity, and Transparency, fixed thresholds were applied, with T_i set to 0.4, 10%, and 10%, respectively. For Roughness, a dual-condition criterion was adopted: accuracy was achieved either if both the agent and ground truth identified the interface as flush (values < 0.5) or if the absolute estimation error fell within a dynamic threshold T_i, set to 0.25 times the ground truth *SID*. The goal is to raise the Acc of each metric to above 70%.

For subjective metrics, calibration focused on improving semantic alignment between the agent's prompt-guided reasoning and the expert-consensus benchmark. Because the aim was not to reproduce the exact expert scores but to approximate the relative distribution of higher and lower ratings, Spearman's rank correlation rather than *RMSE* was used to assess the correspondence between the agent scores and expert scores for each subjective item. Using the same Spearman formulation defined above, R_i and Q_i in this analysis represent the ranks of the agent generated score and the expert's score for the *i-th* image, respectively. Cases showing relatively large discrepancies between the agent outputs and the benchmark values were then fed back to the agent, together with the associated street-view images and reference scores. Based on this feedback, the agent was further prompted to identify likely sources of discrepancies, reflect on weaknesses in its current reasoning logic, and propose targeted revisions to its prompt configuration, indicator cues, and reasoning sequence. This feedback-driven process enabled iterative refinement of the analytical workflow, allowing the agent to progressively improve its alignment with the benchmark during calibration.

6. Independent hold-out validation with baseline and ablation-style comparisons

After the calibration stage, the optimized analytical engine was evaluated on an independent hold-out dataset of remaining 24 unseen sampling points. To ensure rigorous scientific control, this validation stage integrated three distinct forms of evaluation within a single analytical framework. The performance of every agent variant in this stage was quantified by benchmarking its generated metrics directly against the ground truth data, i.e., field-measured objective values and expert-annotated subjective scores. Crucially, all comparative tests were executed within the exact same automated API environment. By isolating the prompt architecture as the sole independent variable, we systematically evaluated the system through the following three approaches:

First, an independent hold-out test was conducted to assess the generalization performance and stability of the final optimized agent on unseen street samples.

Second, a zero-shot baseline comparison was introduced to distinguish the structural contribution of the proposed structured workflow from the more generic capabilities of the same Gemini-based agent environment. In this baseline setting, the same basic agent framework was retained, but the model was given only a minimal direct instruction to output the four-dimensional metrics, without explicit spatial definitions, filtering rules, or mathematical formulas.

Third, an ablation-style comparison was conducted to isolate the specific impact of the prompt-based mathematical grounding and semantic calibration. In this comparison, we evaluated a "pre-optimized" agent variant. While this variant retained the detailed semantic scoring rubrics of the final agent, it lacked the explicit cognitive constraints of the final version.

For objective metrics, the comparison was based on the discrepancy between the AI-generated values and the physical benchmark using the same error-based measures as in the pilot stage. For subjective metrics, the comparison was based on the correspondence between the AI-generated scores and the expert-consensus benchmark using the same rank-based criterion as in the pilot stage. Because the subjective benchmark was constructed after r_{WG}-based agreement screening, the effective sample size for subjective validation varied across items and was not necessarily equal to the full 24-point hold-out set.

7. Temporal robustness test

In addition, a temporal robustness test was designed to examine whether the agent produced consistent evaluations under time-related visual variation. Specifically, the test compared metrics derived from images captured in different years, including variation in vegetation density, foliage coverage, and lighting conditions. Because historical street-view imagery was not available for all sampling locations, only a subset of sites was included in this analysis. These selected sites had multi-temporal imagery, minimal spatial displacement, and no substantial changes in the street interface itself. Since the agent cannot directly retrieve historical street-view imagery through the street-view platform, an additional module was implemented in the agent to allow users to manually upload a set of eight images as input. For the selected locations, valid historical street-view images were manually captured and uploaded into the agent using this interface, after which the same analytical pipeline and prompt configuration were applied to ensure that the images were processed under identical reasoning logic.

Through these validation procedures, the stability and professional applicability of the workflow can be assessed. Once the system demonstrates stable and consistent analytical performance, the calibrated reasoning core can be extended into a street-scale visualization agent capable of continuous street interface analysis and spatial result mapping.

2.4.5. Mixed Geometry-Perception Evaluation Experiment Design

To transition from discrete single-point measurements to street-scale interpretation, the system integrates a post-processing module for spatial aggregation, weighted composite evaluation and vectorized visualization.

1. Vectorized mapping and linear segment generation

Within the Analysis Center panel, the system maps discrete sampling points onto the entire street, converting point-based results into continuous vector segments. The algorithm constructs three parallel polylines with color encoding (representing the left interface, right interface, and the integrated interface of the corresponding metrics) to provide an intuitive visual representation.

Users can switch between segmented view and global view to examine either detailed variations or overall street-level results directly on the Analysis Center panel. In segmented view, color-coded segments are used to represent metric values at each sampling interval. For Continuity, Transparency, and Roughness, the visualization simultaneously displays the values of the left and right interfaces, enabling researchers to identify potential bilateral asymmetries along the street. In contrast, for Enclosure, only the central polyline is displayed, reflecting the integrated value derived from both sides of the street.

In global view, the system displays the overall average value of the entire street using a single monochromatic polyline, allowing users to quickly grasp the general performance of the street interface.

2. Weighted composite evaluation

Recognizing that objective geometric measurements and subjective perceptual assessments contribute differently to urban spatial quality, the system supports dynamic weighted aggregation of the two evaluation components.

To ensure comparability across diverse dimensions, both objective and subjective indicators are first standardized through a normalization procedure. This mechanism supports two normalization strategies: relative normalization, which scales values within a specific case study for intra-case comparison, and absolute normalization, which applies manually defined thresholds to enable comparisons across different study areas. For the latter, reference value ranges are provided within the interface as user-defined benchmarks.

Users can subsequently assign customized weights to the normalized objective and subjective indicators, enabling the generation of composite street evaluation scores tailored to different planning priorities. The composite metric is calculated using the following formula:

$$Q_{composite} = w_{obj} \cdot Q_{obj} + w_{subj} \cdot Q_{subj}, \quad w_{obj} + w_{subj} = 1 \tag{6}$$

3. Results

3.1. Reliability Verification of Experimental Framework

3.1.1. Verification of Internal Workflow Consistency and Stability

Before benchmarking the agent against external ground-truth references, we first examined whether the workflow operated consistently and coherently as an internal analytical system. The results confirmed that the proposed agent maintained reliable alignment across image sampling, visual interpretation, rule-based calculation, semantic scoring, and repeated execution:

1. Spatial consistency of image sampling

The sampled image sequences were verified to be spatially consistent with the intended street route and viewing directions. The static image sets generated for each sampling point correctly followed the predefined origin-to-destination path and preserved the intended orientation relationships across the street-view sequence. This indicates that the visual input used for subsequent analysis was geometrically aligned with the actual street segment being evaluated.

2. Side-specific directional consistency of interpretation and scoring

Cross-checking the intermediate textual descriptions against corresponding street-view images confirmed that the agent correctly assigned each single-sided interface analysis to the corresponding street side, without reversing the left and right interfaces during analysis.

3. Rule compliance of the objective metric stream

The objective metric stream showed satisfactory rule compliance. The model-generated objective outputs were broadly consistent with the geometric logic and spatial scope defined for each metric, and no obvious contradiction was observed between the visual evidence and the final quantitative estimates in the inspected cases.

4. Semantic compliance of the subjective metric stream

The subjective metric stream also demonstrated satisfactory semantic compliance. Visual indicators specified in the prompts, such as continuous street walls, interface gaps, transparent openings, and staggered façade conditions, were appropriately identified and reflected in the corresponding subjective scores. This suggests that the agent's perceptual evaluation did not function as arbitrary text generation, but instead followed the intended indicator-based semantic reasoning framework.

5. Internal correlation logic

The analysis revealed varying degrees of correlation between the agent's subjective scores and objective metrics. For example, a moderate correlation was observed between subjective and objective Enclosure ($\rho = 0.46$, $p = 0.005$, $N = 36$), while a strong correlation was found between subjective and objective Continuity ($\rho = 0.80$, $p < 0.001$, $N = 36$). These results indicate that the agent's subjective evaluation and objective metric estimation were generated through different prompt-based reasoning processes rather than a single unified calculation. The varying levels of correlation across dimensions further suggest that the two evaluation pathways captured related but not identical aspects of street interface morphology.

6. Stability checking

Results from five repeated generation rounds on the pilot subset showed minimal variance in the output values, confirming that the workflow maintained a reliable level of stability with minimal randomness across different runs.

3.1.2. Verification of Subjective Ground Truth Reliability

The proportion of samples meeting this criterion ($r_{WG} > 0.7$) across the seven subjective items is presented in Table 3, together with the corresponding retained and excluded sample counts. Results show a generally high level of within-group agreement across all dimensions, with the Agreement Rate for all items exceeding 70%. Across the full dataset of 36 samples, 29 were retained for enclosure, 28 and 26 for continuity on Sides A and B, 26 and 26 for transparency on Sides A and B, and 27 and 27 for roughness on Sides A and B; the corresponding exclusion counts were 7, 8, 10, 10, 10, 9, and 9, respectively. Among the 12 pilot samples used for prompt calibration, the corresponding excluded counts were 3 for enclosure, 2 and 3 for continuity on Sides A and B, 4 and 7 for transparency on Sides A and B, and 4 and 6 for roughness on Sides A and B, as indicated by the values in parentheses in Table 3.

Table 3. Within-group agreement and retained/excluded sample counts for expert subjective assessments across the seven subjective items. Source: Authors' elaboration.

Dimension	Scope	Retained Samples (n)	Excluded Samples (n)	Agreement Rate ($r_{WG} > 0.7$) (%)
Enclosure	Overall	29 (9)	7 (3)	80.6
Continuity	Side A	28 (10)	8 (2)	77.8
	Side B	26 (9)	10 (3)	72.2
Transparency	Side A	26 (8)	10 (4)	72.2
	Side B	26 (5)	10 (7)	72.2
Roughness	Side A	27 (8)	9 (4)	75.0
	Side B	27 (6)	9 (6)	75.0

Note: Values outside parentheses indicate counts based on the full dataset of 36 sampling points, whereas values in parentheses indicate the corresponding counts within the 12-sample pilot subset. Agreement Rate was calculated based on the full dataset. Only sample—item pairs with $r_{WG} > 0.7$ were retained for subsequent calibration and validation analyses.

However, some low-agreement cases were still observed, with the minimum r_{WG} value dropping to 0.22. These cases mainly corresponded to street interfaces with more complex, irregular, or visually ambiguous spatial layouts, where the street boundary was less clearly defined. This pattern suggests that expert judgments tend to diverge more in morphologically heterogeneous environments. Finally, only sample—item pairs with $r_{WG} > 0.7$ were retained for subsequent calibration and validation analyses. For each retained sample–item pair, the expert-consensus ground truth was defined as the mean score across the ten experts.

3.2. Pilot Study and Prompt Calibration

In the pilot study phase, we took the median of the values generated by the agent five times to compare with the ground truth benchmark of 3 streets among all 9 streets.

3.2.1. Calibration of Objective Measurement Mechanism

The preliminary pilot-sample analysis indicated that while some metrics were satisfactory, others required improvement. After refinement, accuracies of all metrics significantly improved and successfully met the acceptance criteria. For instance, the *RMSE* for Enclosure dropped from 1.63 to 0.43, with its Acc rising from 50.0% to 75.0%; similarly, for Transparency (Side A), the *RMSE* decreased from 19.63% to 9.13%, while the Acc surged from 58.3% to 83.3%.

3.2.2. Calibration of Subjective Measurement Mechanism

Before refine the agent's prompt, most correlations were low, with Spearman's ρ ranking from 0.33 to 0.73. The optimization aimed to align the agent's scoring with expert evaluations in the pilot study, targeting value of ρ of approximately 0.7. To adjust the prompts, the image sequences of pilot samples were analyzed by the agent. Over several iterations, the common characteristics of high and low-value images were identified and integrated. These insights were used to refine the description prompts in the corresponding sections. The updated prompts specify more detailed strategies for scoring high and low values, utilizing conditional descriptions to adapt to various scenarios.

Table 4 shows the most representative prompt items excluding their definitions before and after the optimization.

Results after the optimization demonstrate a significant improvement in consistency. For instance, Continuity (Side A) increased from 0.35 to 0.66, while Roughness (Side B) surged from 0.41 to 0.93.

Table 4. Comparison of representative prompts before and after optimization. Source: Authors' elaboration.

Dimension	Prompt (Before)	Prompt (After)
Enclosure	Low: Vast sky view, low buildings, wide setbacks, empty lots High: Narrow sky strip, tall street walls, canyon effect	Low Score Indicators: - Horizontal span dominance: The effective street width visually overwhelms the maximum interface height; - Indistinct spatial edges: Lack of continuous boundaries, causing visual leakage into voids; - Asymmetric Street Profile: One side is significantly lower or open, even if the other side has high interface. High Score Indicators: - Obscured Height Boost: If buildings are >5 stories and their upper parts are obscured by canopy, enclosure is reinforced; - Vertical Field Dominance (Vertical boundaries visually overpower the horizontal street width.

Table 4. *Cont.*

Dimension	Prompt (Before)	Prompt (After)
Continuity	Low: Large gaps, vacant lots, inconsistent fencing High: Continuous wall, hedge (>1.5 m), retail frontage	Low Score Indicators: - Physical Breaks: Significant interruptions (e.g., surface parking, empty lots) that physically sever the linear street boundary; - Perpendicular Massing: Narrow gable ends face the street, disrupting the upper silhouette and breaking the continuous cornice line. High Score Indicators: - Complete Linear Coverage: Vertical interfaces extend along the full street segment, forming a physically unbroken boundary; - Rhythmic Cohesion: Unified architectural patterns (e.g., aligned cornices) create a visual rhythm that integrates distinct buildings into a seamless ribbon.
Transparency	Low: Solid masonry, shuttered doors, dense privacy hedges High: Glass curtain walls, open metal fencing, large shop windows	Low Score Indicators: - Solid Surface Dominance: The area of opaque materials (e.g., concrete walls) overwhelmingly exceeds that of openings, creating a visual barrier that blocks interior views; - Blocked Green Barrier: Dense vegetation completely obstructs views into the property, acting as an opaque "green wall" rather than a permeable boundary. High Score Indicators: - Layered Transparency: Background transparency (e.g., lit interiors or glass) takes precedence over permeable foreground layers like fences or railings; - Visual Permeability Dominance: Transparent elements (e.g., glass lobbies, arcades) exceed solid walls, allowing for deep visual penetration that effectively merges the street with the interior.
Roughness	Low: Straight wall alignment High: Complex setbacks, jagged building line	Low Score Indicators: - Monolithic Curtain Walls: Large-scale surfaces functioning as a single continuous 2D plane, prioritizing planar alignment over surface articulation; - Coplanar Detailing: Architectural elements (e.g., windows, spandrels) sit flush with the facade plane, creating a seamless, taut surface lacking depth. - Horizontal Linear Elements: Pronounced horizontal features (e.g., ribbon windows, decorative bands) emphasize lateral continuity, creating a planar reading that reduces vertical volumetric depth. High Score Indicators: - Deep Geometric Stagger: Buildings arranged in staggered or stepped patterns where upper levels or adjacent masses recede progressively, breaking vertical continuity; - Curvilinear or Organic Forms: Fluid, non-planar geometries (e.g., curved or twisted forms) that generate macro-level depth.

3.3. Final Validation of Optimized Agent Against Ground Truth

3.3.1. Validation of Objective and Subjective Measurements

As shown in Tables 5–7, the optimized agent generally achieved the best objective performance on the independent hold-out dataset (N = 24), indicating that prompt refinement substantially improved alignment with the physical benchmark. Compared with the pre-optimized agent, the optimized agent showed markedly lower *RMSE* values for Enclosure (from 1.38 to 0.35), Continuity on Side B (from 28.16% to 16.35%), Transparency on both sides (from 32.03% and 23.05% to 13.88% and 13.65%, respectively), and Roughness on both sides (from 1.91 m and 1.83 m to 1.24 m and 1.41 m, respectively). Accuracy also improved across most dimensions, reaching 79.2% for Enclosure, 83.3% for Continuity on Side B, 66.7% and 70.8% for Transparency on Sides A and B, and 79.2% and 75.0% for Roughness on Sides A and B, respectively. These results confirm that the optimized prompt configuration substantially strengthened the agent's geometric measurement performance on unseen street samples.

Table 5. Performance of the direct-prompt baseline within the same Gemini-based agent environment on objective metrics (n = 24). Source: Authors' elaboration.

Dimension	Scope	*RMSE*	*Acc* (%)
Enclosure	Overall	0.71	45.8
Continuity	Side A	10.24%	75.0
	Side B	22.39%	66.7
Transparency	Side A	36.23%	20.8
	Side B	23.79%	33.3
Roughness	Side A	2.00 m	70.8
	Side B	1.71 m	70.8

Note: The evaluation dataset contains 24 sampling points. For a small number of cases (1–2 samples), certain metrics could not be calculated due to the absence of valid street interface boundaries; these samples were excluded from the corresponding metric calculations.

Table 6. Performance of the pre-optimized agent on objective metrics (n = 24). Source: Authors' elaboration.

Dimension	Scope	*RMSE*	*Acc* (%)
Enclosure	Overall	1.38	33.3
Continuity	Side A	15.01%	62.5
	Side B	28.16%	47.8
Transparency	Side A	32.03%	29.2
	Side B	23.05%	50.0
Roughness	Side A	1.91 m	66.7
	Side B	1.83 m	70.8

Note: The evaluation dataset contains 24 sampling points. For a small number of cases (1–2 samples), certain metrics could not be calculated due to the absence of valid street interface boundaries; these samples were excluded from the corresponding metric calculations.

At the same time, the direct-prompt baseline provides a more informative reference for interpreting these gains. In some objective dimensions, especially Enclosure and Continuity, the direct-prompt baseline already outperformed the pre-optimized agent. For example, the baseline produced lower *RMSE* values for Enclosure (0.71 versus 1.38), Continuity on Side A (10.24% versus 15.01%), and Continuity on Side B (22.39% versus 28.16%), with correspondingly higher accuracy values in these cases. This suggests

that simply adding more detailed prompt instructions does not necessarily improve performance unless the prompt logic is properly calibrated; in some cases, an insufficiently optimized structured prompt may introduce additional reasoning noise and perform worse than a simpler direct-prompt setting. Nevertheless, after optimization, the proposed agent generally outperformed both the direct-prompt baseline and the pre-optimized agent across the objective metrics, although a few local exceptions remained, such as Continuity on Side A, where the baseline still showed a slightly lower *RMSE* than the optimized version.

Table 7. Performance of the optimized agent on objective metrics (n = 24). Source: Authors' elaboration.

Dimension	Scope	*RMSE*	*Acc* (%)
Enclosure	Overall	0.35	79.2
Continuity	Side A	12.70%	75.0
	Side B	16.35%	83.3
Transparency	Side A	13.88%	66.7
	Side B	13.65%	70.8
Roughness	Side A	1.24 m	79.2
	Side B	1.41 m	75.0

Note: The evaluation dataset contains 24 sampling points. For a small number of cases (1–2 samples), certain metrics could not be calculated due to the absence of valid street interface boundaries; these samples were excluded from the corresponding metric calculations.

Roughness remained the most challenging objective dimension. Its calculation depends on a more complex combination of spatial depth judgment, setback recognition, and interface classification, which makes it more difficult for the model to estimate consistently from street-view images. Although the optimized agent reduced the *RMSE* relative to the pre-optimized agent and maintained acceptable accuracy, some discrepancies still occurred in cases involving deep façade recesses, irregular setbacks, or visually layered street interfaces. These remaining discrepancies suggest that the current prompt logic for Roughness could be further improved by incorporating more refined rules for judging spatial depth and interface discontinuity.

For subjective evaluation, we calculated the Spearman correlation between the agent predictions and the expert-consensus scores on the independent hold-out dataset, using the retained high-agreement sample–item pairs for each metric (24 samples before item-specific agreement screening), as shown in Tables 8–10.

Table 8. Performance of the direct-prompt baseline on subjective metrics. Source: Authors' elaboration.

Dimension	Scope	n	Spearman's ρ	p-Value
Enclosure	Overall	20	0.37	0.108
Continuity	Side A	18	0.54	0.021
	Side B	17	0.48	0.051
Transparency	Side A	18	0.21	0.403
	Side B	21	0.48	0.028
Roughness	Side A	19	0.34	0.154
	Side B	21	0.07	0.763

Note: The original dataset contains 24 samples. Samples with low within-group agreement among experts were excluded from the analysis, resulting in effective sample sizes ranging from n = 17 to 21 across metrics.

Table 9. Performance of the pre-optimized agent on subjective metrics. Source: Authors' elaboration.

Dimension	Scope	*n*	Spearman's ρ	*p*-Value
Enclosure	Overall	20	0.44	0.052
Continuity	Side A	18	0.68	0.002
	Side B	17	−0.13	0.619
Transparency	Side A	18	0.16	0.526
	Side B	21	0.29	0.202
Roughness	Side A	19	0.34	0.154
	Side B	21	0.21	0.361

Note: The original dataset contains 24 samples. Samples with low within-group agreement among experts were excluded from the analysis, resulting in effective sample sizes ranging from n = 17 to 21 across metrics.

Table 10. Performance of the optimized agent on subjective metrics. Source: Authors' elaboration.

Dimension	Scope	*n*	Spearman's ρ	*p*-Value
Enclosure	Overall	20	0.65	0.002
Continuity	Side A	18	0.71	0.001
	Side B	17	0.79	<0.001
Transparency	Side A	18	0.65	0.004
	Side B	21	0.78	<0.001
Roughness	Side A	19	0.72	<0.001
	Side B	21	0.71	<0.001

Note: The original dataset contains 24 samples. Samples with low within-group agreement among experts were excluded from the analysis, resulting in effective sample sizes ranging from n = 17 to 21 across metrics.

The direct-prompt baseline showed limited and uneven alignment with expert ratings. Only Continuity on Side A (ρ = 0.54, p = 0.021) and Transparency on Side B (ρ = 0.48, p = 0.028) reached moderate and statistically significant correlations, whereas several other metrics remained weak or non-significant, such as Transparency on Side A (ρ = 0.21, p = 0.403) and Roughness on Side B (ρ = 0.07, p = 0.763). These results suggest that direct prompting with the underlying Gemini model alone was insufficient to achieve stable agreement with expert perceptual judgments across all subjective dimensions.

The pre-optimized agent showed mixed performance. While Enclosure (ρ = 0.44, p = 0.052) and Continuity on Side A (ρ = 0.68, p = 0.002) showed moderate to strong alignment, several other metrics remained weak, including Continuity on Side B (ρ = −0.13, p = 0.619), Transparency on Side A (ρ = 0.16, p = 0.526), and Roughness on Side B (ρ = 0.21, p = 0.361). This indicates that the initial workflow structure alone did not yet provide consistently reliable semantic alignment with expert ratings.

After optimization, the agent's performance improved substantially across all seven subjective items. All correlations increased to $\rho \geq 0.65$, and all reached statistical significance ($p \leq 0.004$). In particular, Continuity on Side B increased from ρ = −0.13 in the pre-optimized agent to ρ = 0.79 in the optimized agent ($p < 0.001$), Transparency on Side A increased from ρ = 0.16 to ρ = 0.65 (p = 0.004), and Roughness on Side B increased from ρ = 0.21 to ρ = 0.71 ($p < 0.001$). Compared with the direct-prompt baseline and the pre-optimized agent, the optimized agent also showed consistently stronger correspondence with expert judgments across all metrics.

These findings reaffirm the robustness of the optimized prompts and improved agent logic, laying a solid foundation for expanding the model to larger street-scale datasets. In addition, the baseline comparison with baseline indicates that the observed gains cannot

be attributed to the underlying Gemini model alone, but are associated with the proposed workflow design and its prompt-based semantic calibration process. At the same time, these results demonstrate improved alignment within the tested Gemini-based setting, but do not imply universal superiority across alternative models or workflows.

3.3.2. Validation of Temporal Robustness

To test temporal robustness, ten sampling locations with multi-temporal street-view imagery were selected. Two image sets captured in different years were manually uploaded into the agent's image-input module and analyzed using the same workflow as the automatic method (Figure 7). The results across the four core dimensions were averaged and compared between the two temporal groups (Table 11).

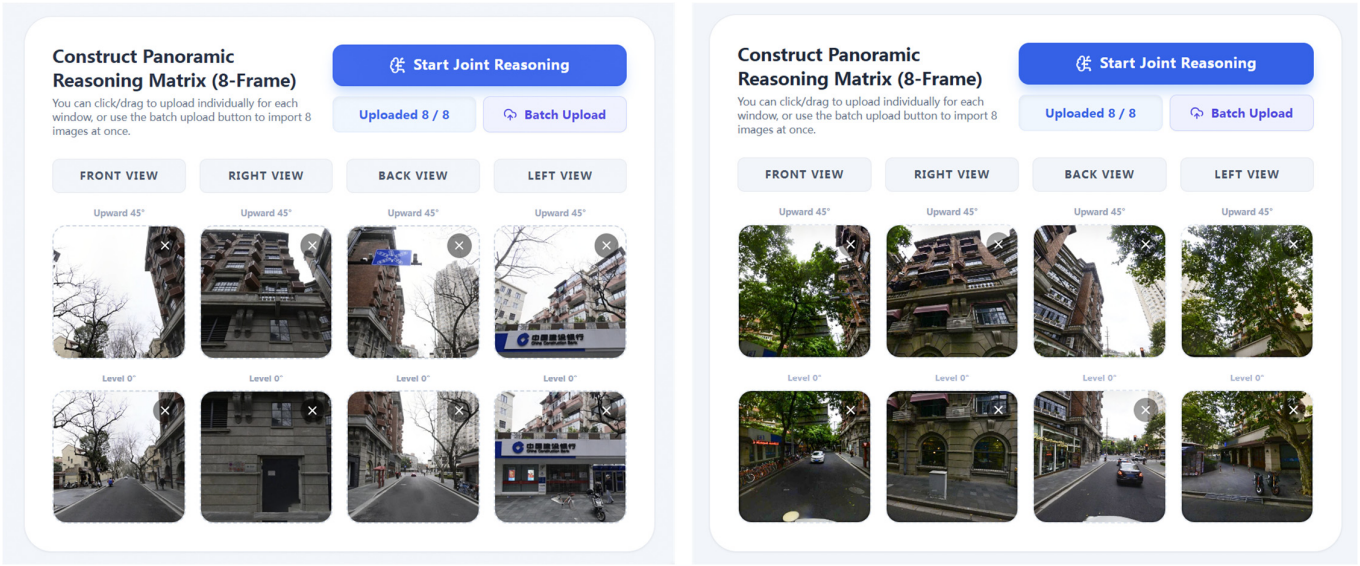


Figure 7. Example of multi-temporal street-view image matrices used in the temporal robustness test. Any Chinese text visible in the street-view images comes from the original urban scene and does not affect interpretation of the figure. Source: Authors' elaboration.

Table 11. Temporal robustness of the optimized agent across objective and subjective metrics. Source: Authors' elaboration.

Dimension	Average Difference (Objective)	Average Difference (Subjective)
Enclosure	0.09	0.23
Continuity	2.93%	0.20
Transparency	8.50%	0.57
Roughness	0.15 m	0.14

The overall results demonstrated a high degree of consistency, with average differences regarded as small parts within the range of each metric. This is attributed to the detailed prompt-based evaluation rules that were used. These rules effectively minimized the influence of seasonal factors, such as changes in lighting and tree coverage, which otherwise might have affected the evaluation.

However, Transparency showed noticeable variation. This can be explained by the presence of "opaque glass backed by solid panels" in some locations, which may appear to be windows but are actually non-transparent storefronts or signage. The agent was able to accurately identify these features based on these prompt-defined criteria, but in different

temporal images, the storefront renovations could lead to varying appearances of these glass features.

Overall, the agent adhered closely to the rules defined in the prompts. However, when objective elements of the street boundary (such as physical objects in the environment) change substantially according to the rules, differences in the evaluation results arise. Therefore, how to scientifically design prompts depends on the researcher's desired focus on measurement orientation.

3.4. Scalable Deployment: Continuous Street Interface Morphological Analysis

To demonstrate the agent's capability in scalable deployment, the optimized model was applied to three new representative streets segments, Hengshan Road (from Huashan Road to Baoqing Road), Middel Huaihai Road (from Changshu Road to South Chongqing Road) and Funshun Road (from An'shan Road to Huangxing Road), selected from the three morphological zones defined previously. The following analysis is organized into two parts: intra-case spatial research, which investigates the detailed internal characteristics of a single case, and cross-case spatial research, which compares the distinctive features across these different urban contexts.

3.4.1. Intra-Case Spatial Research

We took Hengshan Road as an example to analyze its morphological characteristics in different inside areas using the agent's capability. To maximize the visibility of variations in morphological metric values among different segments along the street, we applied Relative Normalization Mode, scaling metrics based on the street's own min–max range (except for the Roughness objective metric, which utilized a manually defined scale from 0 to 0.6 m to mitigate the masking effect of isolated outliers). Furthermore, to investigate the relationship between physical geometry and human perception, we adopted the strategy that adjusts the weighting system to isolate pure objective measurement (100% Objective weight) and isolate pure subjective measurement (100% Subjective weight). Sampling Spacing is set at 50 m.

Figures 8 and 9 show the comparative analysis of subjective and objective measurement of the four metrics. Note that the colors in each diagram just reflect the relative value variance within that diagram only, but do not reflect absolute value comparisons across different diagrams.

Regarding Enclosure, the distribution trends of objective and subjective metrics align similarly across the streetscape. Both accurately identify low values in the southwest section (near Xujiahui Park), where low-rise structures and the open park interface create asymmetry that significantly diminishes enclosure. Conversely, the central section registers high values due to the combined effect of tall buildings and dense tree canopies hiding the upper layers. However, a notable divergence between the two evaluations appears in the northeast section.

Regarding Continuity, both metrics maintain a high baseline with synchronized fluctuation locations, yet the subjective metric exhibits significantly higher volatility. A detailed comparison reveals that this discrepancy may results from the subjective measurement's higher sensitivity to specific interface conditions, such as stylistically inconsistent low-rise boundaries, where it perceives greater discontinuity than the objective measurement.

Regarding Transparency, the spatial distribution of subjective and objective values aligns similarly, despite minor differences in magnitude. Both metrics similarly identify low transparency in areas dominated by construction hoardings or vegetation-obscured walls, while assigning high values to street-level retail and restaurant interfaces featuring extensive glazing.

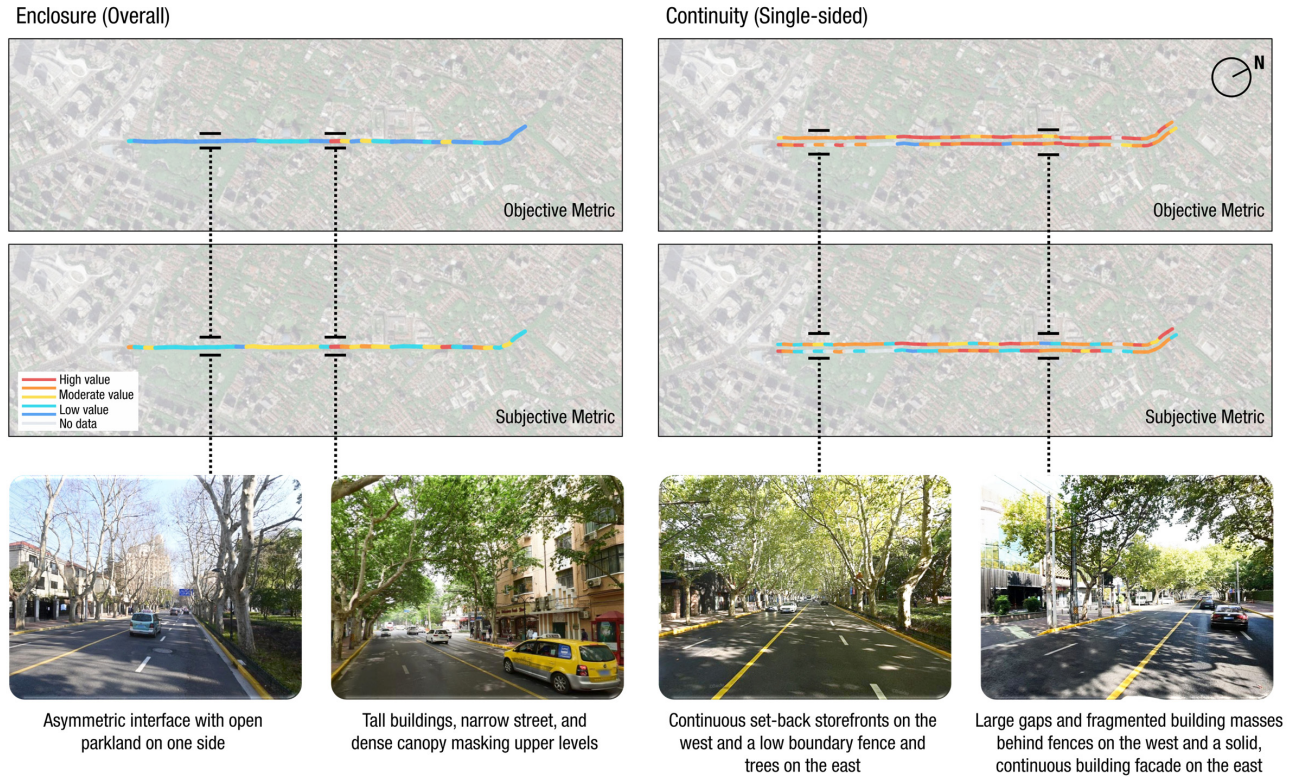


Figure 8. Intra-case spatial comparison of objective and subjective evaluations: Enclosure and Continuity. Source: Authors' elaboration.

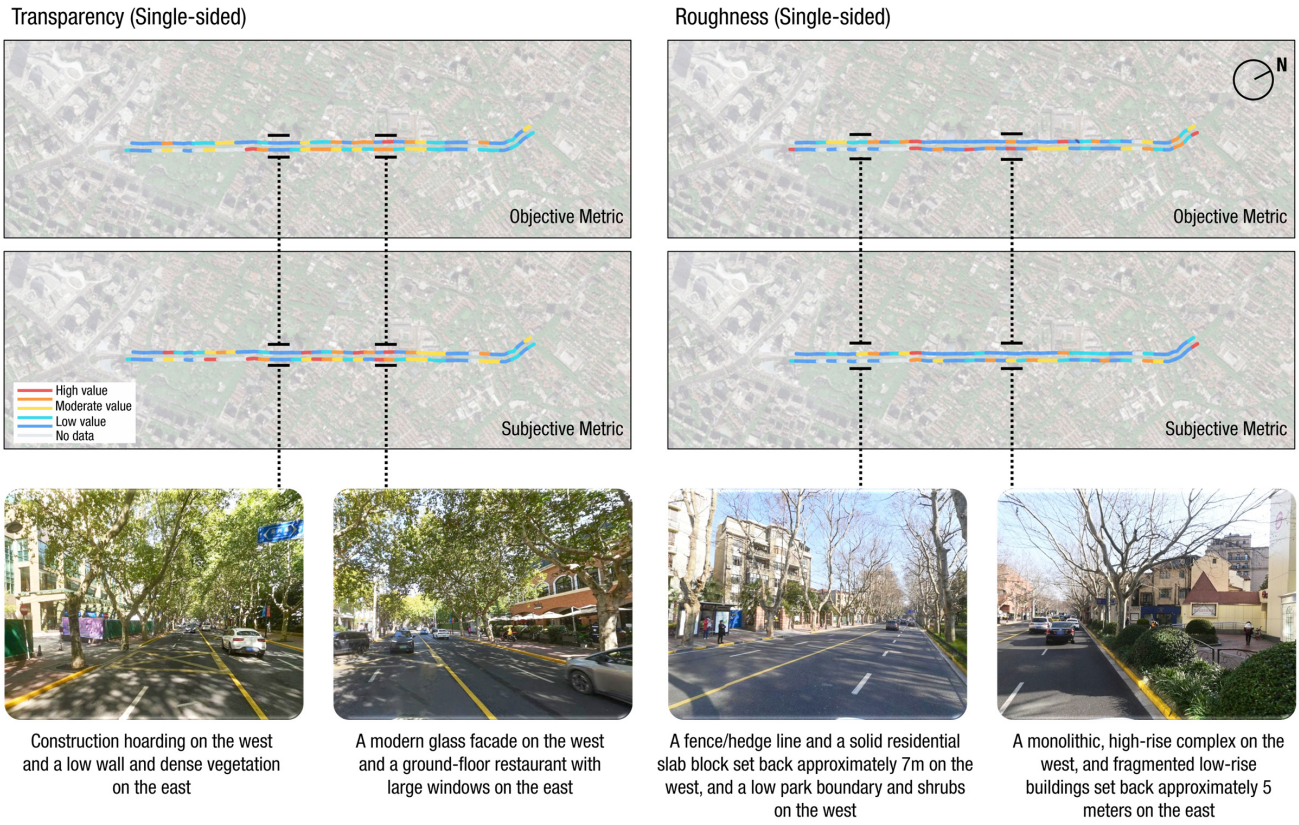


Figure 9. Intra-case spatial comparison of objective and subjective evaluations: Transparency and Roughness. Source: Authors' elaboration.

Regarding Roughness, the overall trends of high and low values are largely similar between the two metrics, though local divergences exist. For instance, while planar permeable fences are objectively rated as smooth, the subjective agent occasionally perceives higher roughness due to visible depth variations behind the boundary. Conversely, both metrics similarly attribute high values to areas where significant architectural setbacks actually exist.

3.4.2. Cross-Case Spatial Research

We compared the distinctive morphological characters across the three typological cases. To ensure rigorous comparability across different urban contexts, we switched to Absolute Normalization Mode (scaling metrics against a unified threshold). To generate a synthesized assessment of the street environment, we applied the weighted evaluation framework under a balanced scenario (w_{obj} = 0.5, w_{subj} = 0.5). This method first normalizes each subjective and objective metric into a unified range across different cases, and then combines the two values for comparison (Figure 10).

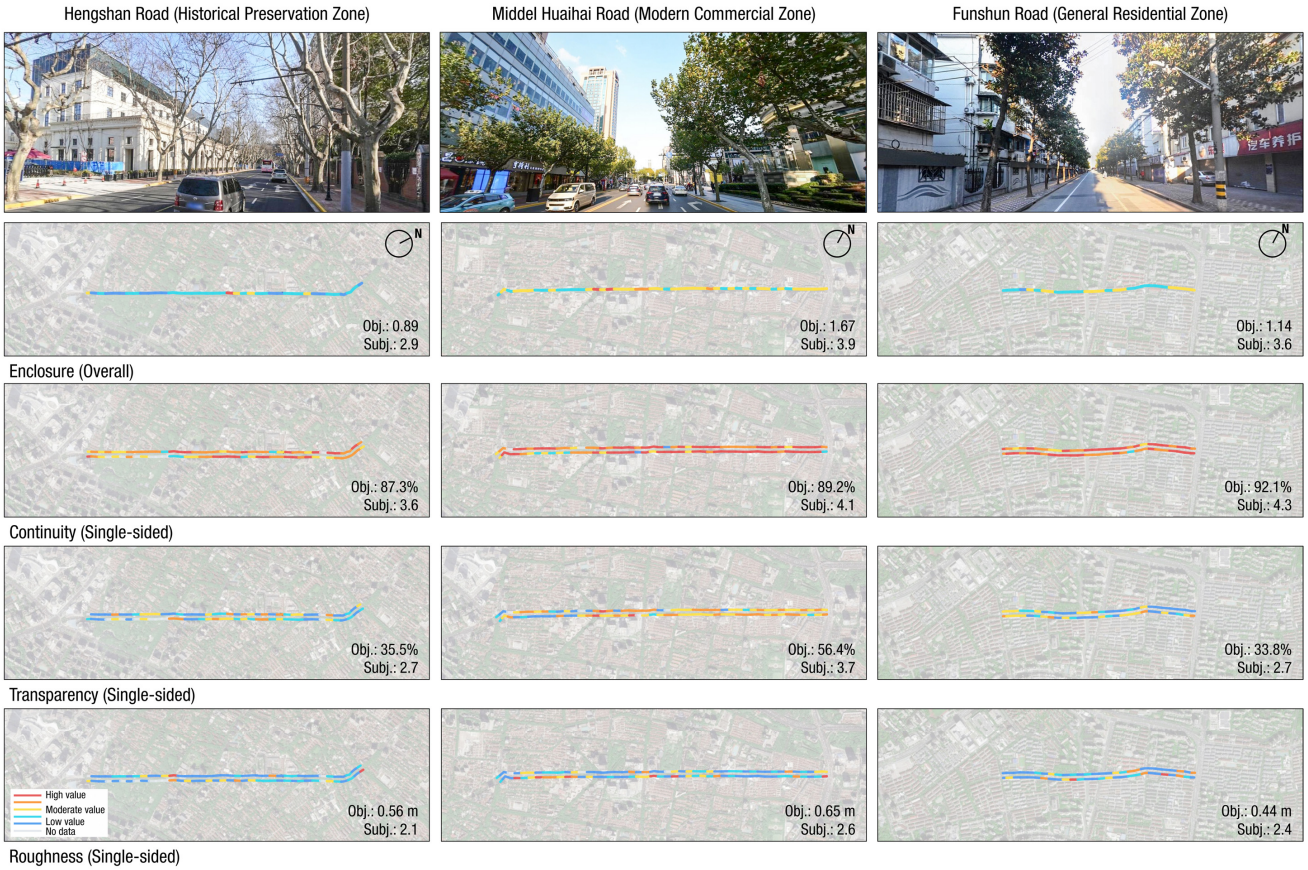


Figure 10. Cross-case comparison of integrated spatial metrics across three street typologies (mapped at the same scale), annotated with segment-wide average values in the bottom-right corner. Any Chinese text visible in the street-view images comes from the original urban scene and does not affect interpretation of the figure. Source: Authors' elaboration.

The cross-case comparison shows that across distinct urban fabrics, individual street metrics exhibit specific characteristics while also sharing certain similarities. Notably, with the exception of Enclosure, the other three metrics show a relatively high degree of bilateral proximity, with similar spatial fluctuation patterns for high and low values on both sides of the street. This suggests that in most urban environments, the morphological features of opposing street interfaces remain relatively symmetrical.

Turning to the specific differences across the cases, Hengshan Road generally has lower Enclosure than the other cases, despite its historic context. This is likely attributed to lower building heights and the intermittent lack of building interfaces on one side. Conversely, Middle Huaihai Road and Fushun Road maintain higher enclosure due to taller buildings and narrower street widths, respectively. In terms of Continuity, all three cases are similar, featuring intact interfaces without significant gaps. Regarding Transparency, Middle Huaihai Road scores the highest, driven by its continuous modern commercial windows. Finally, Roughness is similarly across all cases, which are dominated by planar interfaces with only minor localized fluctuations.

In general, these quantitative results align with established urban typologies and common sense. The distinct metric profiles successfully capture the unique spatial signatures of each case, validating that the integrated evaluation model accurately reflects the actual physical environment.

4. Discussion and Conclusions

4.1. Principal Findings and Contribution

Evaluating street interface morphology has long faced the challenge of combining scalable urban analysis with the higher-level interpretive depth required for urban design. This study addresses this gap by developing an automated MLLM-agent framework. Through the systematic implementation of the proposed methodology, this research contributes to addressing several key limitations in existing street interface morphology research:

1. Complementary assessment of street interface morphology: Traditionally, urban design research has often been divided between rigid geometric measurement and nuanced perceptual assessment. Addressing the first research objective, this study shows that the proposed MLLM-agent workflow can effectively conduct joint assessments of both objective geometric and subjective perceptual indicators directly from street-view imagery. By leveraging the joint reasoning capabilities of a natively multimodal model within an agent-based workflow, the system uses structured prompts to estimate objective geometric indicators and infer subjective morphological scores in a unified analytical process, thereby establishing an automated framework for street interface morphology assessment.
2. Codifying expert logic into automated workflows: A significant barrier in large-scale urban analysis is the difficulty of replicating the interpretive and rule-based reasoning used by professional planners. For our second research objective, we demonstrated that through structured dual-benchmark calibration and prompt engineering, the agent's reasoning can be aligned with ground-truth geometric measurement and expert consensus. This contributes a viable pathway for translating refined expert evaluation into scalable digital workflows. By mitigating the randomness of AI outputs, the framework helps ensure that automated assessments are not merely data-driven, but are anchored in established urban design principles and expert evaluative logic. Notably, this framework enables a rapid optimization process that is independent of large-scale manual annotation, thereby ensuring robust scenario generalizability across diverse and complex urban environments.
3. Capturing the continuous rhythm of streetscapes: Urban design practice emphasizes the street as a continuous experience rather than a collection of isolated points. Responding to the third research objective, our workflow successfully transitioned from static point-based sampling to continuous vectorized mapping. This advancement addresses the gap in capturing the spatial rhythm of streets, allowing for a more accurate reflection of how street interface qualities fluctuate along a journey.

4.2. Advantages and Inherent Limitations

The theoretical and practical significance of the proposed MLLM-agent framework is further clarified when situated within the broader landscape of existing urban evaluation methodologies. The advantages of our framework can be highlighted across four methodological aspects:

1. Beyond pixel-level segmentation: Traditional computer vision approaches, primarily represented by Deep Convolutional Neural Networks (DCNNs), focus on the statistical aggregation of visual elements via pixel-level semantic segmentation. While effective at quantifying the physical presence of elements, these models often encounter a "semantic gap" when attempting to synthesize fragmented visual extractions into abstract spatial relationships or complex morphological configurations. The proposed MLLM-agent overcomes this by executing contextual semantic reasoning, mimicking a planner's cognitive ability for morphological synthesis.
2. Interpretable "White-Box" reasoning: Advanced streetscape studies often employ secondary neural networks trained on segmented data to predict perceptual scores. However, these methodologies essentially remain "black-box" systems that rely on statistical correlations, lacking the logical "why" behind the results. In contrast, the agent in this study employs evidence-based inductive reasoning through a "reasoning and acting" (ReAct) process. It does not merely output a score but provides explicit natural-language justifications in outputs aligned with established evaluation rules in prompts. This transparency transforms the evaluation from an opaque prediction into a traceable process, which is far more actionable for planners than pure statistical fitting.
3. Zero-shot adaptability versus fragmented workflows: Existing large-scale SVI analytical workflows are often fragmented, requiring complex coordination between different platforms. The MLLM-agent operates as an integrated, automated workflow that leverages pre-trained large language models. Through prompt engineering, it can adapt flexibly to diverse urban contexts without task-specific retraining or large-scale manual annotation, while also reducing the need for complex cross-platform data conversion and transfer.
4. Lowering technical barriers for agent-based analysis: Beyond analytical performance, the proposed workflow lowers the technical threshold for developing and refining AI-based urban analysis agents. By relying on an AI platform to coordinate multimodal reasoning, structured prompting, and spatial visualization, it enables researchers to construct, test, and optimize agent-based workflows more rapidly and with less programming overhead.

Despite these structural advantages, the framework exhibits certain inherent limitations:

1. Inherent instability of generative AI: MLLMs possess an inherent randomness, and the stability of the image analysis can fluctuate, especially when evaluating physical indicators, such as accurately estimating a large spatial depth, while the corresponding key spatial references are ambiguous or missing in the SVI, or when the evaluation rules are formulated without sufficient detail.
2. Challenges in quantifying complex interfaces: Current metric models and sampling methods still face challenges in quantifying complex geometries such as arcades or overhanging structures, which necessitates the clear formulation of rigorous rules for semantic interpretation. Furthermore, boundary standards must be tailored for specific urban fabrics; for instance, the wall-defined boundaries of high-density cities like Shanghai require localized extraction rules that differ from those used for continuous building façades in European cities.

3. Prompt interference: We observed that complex reasoning rules for objective and subjective metrics can occasionally interfere with each other within the same task, subtly affecting output precision, which highlights a key area for future research to explore optimization strategies.
4. Balancing precision and adaptability in prompt design: While the MLLM agent strictly follows the logic provided in the prompts, crafting these instructions involves a delicate trade-off. Overly rigid rules may ensure consistent results for specific street types but often fail to adapt to the vast diversity of global urban contexts. Conversely, vague or broad instructions can lead to inconsistent scoring. Finding the optimal balance between rule-based precision and morphological adaptability remains a critical area for future optimization.

4.3. Future Research Directions

The establishment of this automated MLLM-agent framework opens several promising avenues for future urban science research:

1. Multi-source data integration: To overcome the drawbacks of relying solely on street-level images, future research should combine the AI's visual analysis with satellite data and GIS mapping. This integration would provide a more complete spatial context and ensure more reliable results for complex geometric measurements.
2. Context-aware boundary customization: Urban forms differ significantly across cultural and historical settings. Future research should refine how the "street boundary" is calculated to fit specific local conditions. Adjusting these computational rules to reflect regional spatial characteristics will improve the accuracy of context-sensitive assessments.
3. From morphological description to psychological experience: While this study introduces subjective evaluation, the current "subjective values" primarily remain at the level of describing objective physical quantities and their direct spatial effects. Future research should aim to bridge these morphological metrics with high-level human psychological experiences, such as a sense of safety, intimacy, or urban vibrancy.
4. Regional-wide morphological mapping: Beyond individual street analysis, this framework can be scaled up to map the morphological features of street interfaces across an entire region. By identifying the spatial distribution of these characteristics, future research could explore how these patterns influence pedestrian movement and urban vitality. These regional-scale "morphological maps" would provide a clear visual guide for planners, helping them make more informed decisions for urban renewal.

Ultimately, the continued exploration of these research avenues will provide a more rigorous framework for the future design and management of urban street interfaces, fostering more vibrant and human-centric urban environments.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/land15040610/s1, Supplementary File S1: Source materials for the urban analysis agent.

Author Contributions: Conceptualization, Y.W. and Y.Y.; methodology, Y.W. and Y.Y.; formal analysis, Y.W.; investigation, Y.W. and C.W.; resources, Y.W.; tool development, Y.W.; data curation, Y.W. and C.W.; writing—original draft preparation, Y.W.; writing—review and editing, Y.W., Y.Y. and C.W.; visualization, Y.W.; supervision, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Shanghai Pilot Program for Basic Research (22TQ1400300); the Research Project of Tongji Architectural Design (Group) Co., Ltd. 2023 (2023J-JB05); the Fundamental Research Funds for the Central Universities (2025-1-ZD-02); the Fundamental Research Funds for the Central Universities (2025-1-YB-04); and the China Postdoctoral Science Foundation (2025M781552).

Data Availability Statement: The reproducibility materials supporting this study are provided in Appendix A and the Supplementary Materials. Appendix A, Table A1 reports the geographic coordinates of the experimental sampling points. The Supplementary Materials include the full prompt package, model version, structured output schema, and the core reference scripts/pseudocode used for output generation and parsing. Due to third-party platform licensing and access restrictions, the complete street-view acquisition script is not publicly released. Other datasets supporting the conclusions of this article are available from the authors upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used the Google Gemini AI platform (gemini-3-flash-preview), including its Canvas module, to assist in developing the agent described in this paper. Street-view imagery used in this study was obtained from Baidu Map online services and was used solely for academic research purposes. Sensitive personal information such as faces and vehicle license plates was automatically blurred by the platform. The raw data presented in this study are derived from a combination of automated AI agent execution, expert subjective assessments, and manual field measurements. The authors reviewed and edited the content and take full responsibility for the final manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

LLMs	Large Language Models
MLLMs	Multimodal Large Language Models
AI	Artificial Intelligence
SVI	Street View Imagery
CV	Computer Vision
VLM	Vision-Language Model
RMSE	Root Mean Square Error
D/H	Width-to-Height Ratio
H/D	Height-to-Width Ratio
Acc	Qualified Rate
SID	Street Interface Depth

Appendix A

Table A1. Geographic coordinates of the experimental sampling points. Source: Authors' elaboration.

Zone	Street Segment	Longitude	Latitude
Hengfu Historic Area	Yueyang Road (West Jianguo Rd–Yongjia Rd)	121.459126	31.210209
		121.459052	31.21065
		121.458957	31.211091
		121.458874	31.211533
	Wukang Road (Wukang Mansion–Hunan Rd)	121.444564	31.210795
		121.445226	31.211977
		121.446096	31.21314
		121.446708	31.214336
	Wuxing Road (Hengshan Rd–Middle Huaihai Rd)	121.450044	31.20979
		121.449401	31.210628
		121.448792	31.211305
		121.448389	31.212131

Table A1. *Cont.*

Zone	Street Segment	Longitude	Latitude
Middle Huaihai Road Commercial District	Middle Huaihai Road (Madang Rd–South Xizang Rd)	121.483715	31.230185
		121.484671	31.23057
		121.485618	31.230919
		121.486529	31.231278
	Madang Road (Zizhong Rd–West Jinling Rd)	121.481032	31.224963
		121.480867	31.225837
		121.480395	31.227572
		121.480001	31.228423
	South Huangpi Road (Middle Jinling Rd–Zizhong Rd)	121.480907	31.229332
		121.481333	31.228475
		121.481713	31.227704
		121.481969	31.226884
Yangpu Workers' Village Area	Tieling Road (Zhangwu Rd–Benxi Rd)	121.518843	31.286197
		121.519376	31.285487
		121.519981	31.284712
		121.520595	31.283925
	Jinxi Road (Tieling Rd–Dahushan Rd)	121.520107	31.283179
		121.519248	31.282591
		121.51765	31.281449
		121.515233	31.279782
	Xuchang Road (Kongjiang Rd–Shuangliao Branch Rd)	121.518924	31.278554
		121.519169	31.277703
		121.519659	31.276893
		121.520506	31.276356

References

1. Madanipour, A. *Public and Private Spaces of the City*; Routledge: London, UK, 2003.
2. Bain, L.; Gray, B.; Rodgers, D. *Living Streets: Strategies for Crafting Public Space*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
3. Lynch, K. *The Image of the City*; MIT Press: Cambridge, MA, USA, 1960.
4. Sitte, C. City Planning According to Artistic Principles (1889). In *Historic Cities: Issues in Urban Conservation*; Rodríguez-Lores, J., Ed.; Routledge: London, UK, 2019; Volume 8, p. 307.
5. Jacobs, A.B. *Great Streets*; MIT Press: Cambridge, MA, USA, 1993.
6. Lo, R.H. Walkability: What is it? *J. Urban.* **2009**, *2*, 145–166. [CrossRef]
7. Gehl, J. *Life Between Buildings: Using Public Space*; Island Press: Washington, DC, USA, 2011.
8. Norberg-Schulz, C. *Genius Loci: Towards a Phenomenology of Architecture*; Rizzoli: New York, NY, USA, 1980.
9. Ashihara, Y. *The Aesthetic Townscape*; MIT Press: Cambridge, MA, USA, 1983.
10. Ye, Y.; Zeng, W.; Shen, Q.; Zhang, X.; Lu, Y. The visual quality of streets: A human-centred continuous measurement based on machine learning algorithms and street view images. *Environ. Plan. B Urban Anal. City Sci.* **2019**, *46*, 1439–1457. [CrossRef]
11. Yin, L. Street level urban design qualities for walkability: Combining 2D and 3D GIS measures. *Comput. Environ. Urban Syst.* **2017**, *64*, 288–296. [CrossRef]
12. Middel, A.; Lukasczyk, J.; Zakrzewski, S.; Arnold, M.; Maciejewski, R. Urban form and composition of street canyons: A human-centric big data and deep learning approach. *Landsc. Urban Plan.* **2019**, *183*, 122–132. [CrossRef]
13. Biljecki, F.; Ito, K. Street view imagery in urban analytics and GIS: A review. *Landsc. Urban Plan.* **2021**, *215*, 104217. [CrossRef]

14. Kang, S.; Chen, X.; Huang, C.; Li, J.; Yan, J.; Ye, Y. The exploration of street micro-renewal technology from the perspective of new urban science: A case study of Guzong Road in Nanhui New City. *New Archit.* **2023**, *6*, 52–57.
15. Ewing, R.; Handy, S. Measuring the unmeasurable: Urban design qualities related to walkability. *J. Urban Des.* **2009**, *14*, 65–84. [CrossRef]
16. Cullen, G. *The Concise Townscape*; Routledge: London, UK, 2012.
17. Ye, Y.; Richards, D.; Lu, Y.; Song, X.; Zhuang, Y.; Zeng, W.; Zhong, T. Measuring daily accessed street greenery: A human-scale approach for informing better urban planning practices. *Landsc. Urban Plan.* **2019**, *191*, 103434. [CrossRef]
18. Gaw, L.Y.; Chen, S.; Chow, Y.S.; Lee, K.; Biljecki, F. Comparing street view imagery and aerial perspectives in the built environment. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *10*, 49–56. [CrossRef]
19. Koo, B.W.; Hwang, U.; Guhathakurta, S. Streetscapes as part of servicescapes: Can walkable streetscapes make local businesses more attractive? *Comput. Environ. Urban Syst.* **2023**, *106*, 102030. [CrossRef]
20. Nagata, S.; Nakaya, T.; Hanibuchi, T.; Amagasa, S.; Kikuchi, H.; Inoue, S. Objective scoring of streetscape walkability related to leisure walking: Statistical modeling approach with semantic segmentation of Google Street View images. *Health Place* **2020**, *66*, 102428. [CrossRef] [PubMed]
21. Zhou, H.; He, S.; Cai, Y.; Wang, M.; Su, S. Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning. *Sustain. Cities Soc.* **2019**, *50*, 101605. [CrossRef]
22. Rui, Q.; Cheng, H. Quantifying the spatial quality of urban streets with open street view images: A case study of the main urban area of Fuzhou. *Ecol. Indic.* **2023**, *156*, 111204. [CrossRef]
23. Kelly, C.M.; Wilson, J.S.; Baker, E.A.; Miller, D.K.; Schootman, M. Using Google Street View to audit the built environment: Inter-rater reliability results. *Ann. Behav. Med.* **2013**, *45*, S108–S112. [CrossRef]
24. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015*; IEEE: New York, NY, USA, 2015; pp. 3431–3440.
25. Li, X.; Zhang, C.; Li, W.; Ricard, R.; Meng, Q.; Zhang, W. Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban For. Urban Green.* **2015**, *14*, 675–685. [CrossRef]
26. Seiferling, I.; Naik, N.; Ratti, C.; Proulx, R. Green streets—Quantifying and mapping urban trees with street-level imagery and computer vision. *Landsc. Urban Plan.* **2017**, *165*, 93–101. [CrossRef]
27. Gong, F.Y.; Zeng, Z.C.; Zhang, F.; Li, X.; Ng, E.; Norford, L.K. Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Build. Environ.* **2018**, *134*, 155–167. [CrossRef]
28. Kang, J.; Körner, M.; Wang, Y.; Taubenböck, H.; Zhu, X.X. Building instance classification using street view images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 44–59. [CrossRef]
29. Gong, Z.; Ma, Q.; Kan, C.; Qi, Q. Classifying street spaces with street view images for a spatial indicator of urban functions. *Sustainability* **2019**, *11*, 6424. [CrossRef]
30. Gebru, T.; Krause, J.; Wang, Y.; Chen, D.; Deng, J.; Aiden, E.L.; Fei-Fei, L. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 13108–13113. [CrossRef]
31. Ibrahim, M.R.; Haworth, J.; Cheng, T. Understanding cities with machine eyes: A review of deep computer vision in urban analytics. *Cities* **2020**, *96*, 102481. [CrossRef]
32. Zhou, Y. *Quantitative Research on Street Interface Morphology: Comparison Between Chinese and Western Cities*; Springer: Cham, Switzerland, 2022.
33. Zheng, M.; Tang, W.; Ogundiran, A.; Yang, J. Spatial simulation modeling of settlement distribution driven by random forest: Consideration of landscape visibility. *Sustainability* **2020**, *12*, 4748. [CrossRef]
34. Haddawy, P.; Wettayakorn, P.; Nonthaleerak, B.; Yin, M.S.; Wiratsudakul, A.; Schöning, J.; Laosiritaworn, Y.; Balla, K.; Euaungkanakul, S.; Quengdaeng, P.; et al. Large scale detailed mapping of dengue vector breeding sites using street view images. *PLoS Negl. Trop. Dis.* **2019**, *13*, e0007555. [CrossRef] [PubMed]
35. Naik, N.; Philipoom, J.; Raskar, R.; Hidalgo, C. Streetscore: Predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014*; IEEE: New York, NY, USA, 2015; pp. 779–785.
36. Dubey, A.; Naik, N.; Parikh, D.; Raskar, R.; Hidalgo, C.A. Deep learning the city: Quantifying urban perception at a global scale. In *Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Cham, Switzerland, 2016; pp. 196–212.
37. Pereira, M.F.; Almendra, R.; Vale, D.S.; Santana, P. The relationship between built environment and health in the Lisbon metropolitan area—Can walkability explain diabetes' hospital admissions? *J. Transp. Health* **2020**, *18*, 100893. [CrossRef]
38. Reisi, M.; Nadoushan, M.A.; Aye, L. Local walkability index: Assessing built environment influence on walking. *Bull. Geogr. Socio-Econ. Ser.* **2019**, *46*, 7–21. [CrossRef]
39. Wedyan, M.; Saeidi-Rizi, F. Assessing the impact of walkability indicators on health outcomes using machine learning algorithms: A case study of Michigan. *Travel Behav. Soc.* **2025**, *39*, 100983. [CrossRef]

40. Zeng, Q.; Gong, Z.; Wu, S.; Zhuang, C.; Li, S. Measuring cyclists' subjective perceptions of the street riding environment using K-Means SMOTE-RF Model and street view imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103739. [CrossRef]
41. Chen, N.; Wang, L.; Xu, T.; Wang, M. Perception of urban street visual color environment based on the CEP-KASS framework. *Landsc. Urban Plan.* **2025**, *259*, 105359. [CrossRef]
42. Ye, C.; Zhang, F.; Mu, L.; Gao, Y.; Liu, Y. Urban function recognition by integrating social media and street-level imagery. *Environ. Plan. B Urban Anal. City Sci.* **2021**, *48*, 1430–1444. [CrossRef]
43. Zhang, F.; Zu, J.; Hu, M.; Zhu, D.; Kang, Y.; Gao, S.; Zhang, Y.; Huang, Z. Uncovering inconspicuous places using social media check-ins and street view images. *Comput. Environ. Urban Syst.* **2020**, *81*, 101478. [CrossRef]
44. Helbich, M.; Yao, Y.; Liu, Y.; Zhang, J.; Liu, P.; Wang, R. Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. *Environ. Int.* **2019**, *126*, 107–117. [CrossRef] [PubMed]
45. Johansson, T.; Mangold, M.; Dabrock, K.; Donarelli, A.; Campo-Ruiz, I. Using street view images and visual LLMs to predict heritage values for governance support: Risks, ethics, and policy implications. *arXiv* **2025**, arXiv:2601.06056.
46. Wu, C.; Liang, Y.; Zhao, M.; Teng, M.; Yue, H.; Ye, Y. Perceiving the fine-scale urban poverty using street view images through a vision-language model. *Sustain. Cities Soc.* **2025**, *123*, 106267. [CrossRef]
47. Huang, W.; Wang, J.; Cong, G. Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *Int. J. Geogr. Inf. Sci.* **2024**, *38*, 1414–1442. [CrossRef]
48. Cai, C.; Kuriyama, K.; Gu, Y.; Biljecki, F.; Herthogs, P. Can a large language model assess urban design quality? Evaluating walkability metrics across expertise levels. *arXiv* **2025**, arXiv:2504.21040. [CrossRef]
49. Verma, D.; Mumm, O.; Carlow, V.M. Generative agents in the streets: Exploring the use of Large Language Models (LLMs) in collecting urban perceptions. *arXiv* **2023**, arXiv:2312.13126. [CrossRef]
50. Feng, J.; Liu, T.; Du, Y.; Guo, S.; Lin, Y.; Li, Y. CityGPT: Empowering urban spatial cognition of large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Toronto, ON, Canada, 3 August 2025*; ACM: New York, NY, USA, 2025; Volume 2, pp. 591–602.
51. Li, J.; Ma, M.; Lai, Y. Identifying street multi-activity potential (SMAP) and local networks with MLLMs and multi-view graph clustering. *Comput. Environ. Urban Syst.* **2025**, *122*, 102350. [CrossRef]
52. Tang, Y.; Qu, A.; Yu, X.; Deng, W.; Ma, J.; Zhao, J.; Sun, L. From street views to urban science: Discovering road safety factors with multimodal large language models. *arXiv* **2025**, arXiv:2506.02242. [CrossRef]
53. Al Mushayt, N.S.; Dal Cin, F.; Barreiros Proença, S. New lens to reveal the street interface: A morphological-visual perception methodological contribution for decoding the public/private edge of arterial streets. *Sustainability* **2021**, *13*, 11442. [CrossRef]
54. James, L.R.; Demaree, R.G.; Wolf, G. rwg: An assessment of within-group interrater agreement. *J. Appl. Psychol.* **1993**, *78*, 306. [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.