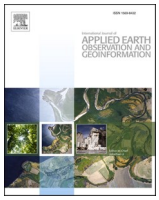
International Journal of Applied Earth Observation and Geoinformation 146 (2026) 105034

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



BuildingMultiView:powering multi-scale building characterization with large language models and Multi-perspective imagery☆

Zongrong Li [a,b], Yunlei Su [a], Filip Biljecki [c,d], Wufan Zhao [a,*]

[a] Thrust of Urban Governance and Design, The Hong Kong University of Science and Technology (Guangzhou), No.1 Duxue Rd, Nansha District, Guangzhou 511453, CN, China
[b] Spatial Sciences Institute, University of Southern California, Los Angeles, United States
[c] Department of Architecture, National University of Singapore, Singapore
[d] Department of Real Estate, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Building Characteristics Extraction
Multi-source Data fusion
Vision Language Models
Built Environment Analysis

ABSTRACT

Buildings play a crucial role in shaping urban environments, influencing their physical, functional, and aesthetic characteristics. However, urban analytics is frequently limited by datasets lacking essential semantic details as well as fragmentation across diverse and incompatible data sources. To address these challenges, we conducted a comprehensive *meta*-analysis of 6,285 publications (2019–2024). From this review, we identified 11 key visually discernible building characteristics grouped into three branches: satellite house, satellite neighborhood, and street-view. Based on this structured characteristic system, we introduce BuildingMultiView, an innovative framework leveraging fine-tuned Large Language Models (LLMs) to systematically extract semantically detailed building characteristics from integrated satellite and street-view imagery. Using structured image–prompt–label triplets, the model efficiently annotates characteristics at multiple spatial scales. These characteristics include swimming pools, roof types, building density, wall–window ratio, and property types. Together, they provide a comprehensive and multi-perspective building database. Experiments conducted across five cities in the USA with diverse architecture and urban form, San Francisco, San Diego, Salt Lake City, Austin, and New York City, demonstrate significant performance improvements, with an F1 score of 79.77% compared to the untuned base version of ChatGPT's 45.66%. These results reveal diverse urban building patterns and correlations between architectural and environmental characteristics, showcasing the framework's capability to analyze both macro-scale and micro-scale urban building data. By integrating multi-perspective data sources with cutting-edge LLMs, BuildingMultiView enhances building data extraction, offering a scalable tool for urban planners to address sustainability, infrastructure, and human-centered design, enabling smarter, resilient cities.

1. Introduction

Urban built environments have experienced unprecedented growth, shaped by an interplay of socio-cultural, economic, and technological factors, resulting in diverse urban landscapes (Liang et al., 2024; Coburn et al., 2017). At the core of these environments, building characteristics play a critical role in determining urban form, function, and performance. Accurately quantifying such characteristics is essential for analyzing spatial morphology, transportation efficiency, energy use, and social equity (Ashik et al., 2024; Li & Li, 2024). However, acquiring key building characteristics (e.g. number of floors, function) remains a major challenge for urban analytics. For example, Biljecki et al. (2021) investigated 140 open government geospatial datasets across 28 countries and found that only half included more than one building characteristic. These characteristics are essential for energy modeling, resilience analysis, and socio-spatial equity studies but are often missing or hard to access.

The usability of current building data is constrained by two key limitations. First, such information, even when available, is typically incomplete. For instance, OpenStreetMap (OSM), one of the most widely used open datasets, in most cases, beyond geometric footprints, it does not have any semantic information about buildings (Biljecki et al., 2023; Knezevic et al., 2022). Second, building-related data is fragmented across heterogeneous sources. These sources differ in format, geographic

☆ This article is part of a special issue entitled: 'Urban Digital Twins' published in International Journal of Applied Earth Observation and Geoinformation.

* Corresponding author.
E-mail addresses: zongrong@usc.edu (Z. Li), wufanzhao@hkust-gz.edu.cn (W. Zhao).

https://doi.org/10.1016/j.jag.2025.105034
Received 16 June 2025; Received in revised form 12 December 2025; Accepted 13 December 2025

coverage, and accessibility (Memduhoglu & Basaraner, 2023). These limitations pose significant barriers to consistent, scalable, and semantically rich building analysis. Therefore, establishing a comprehensive and extensible framework to systematically extract interpretable and multi-perspective building characteristics is essential.

To address these limitations, urban imagery, particularly satellite and street-view imagery, offers a promising avenue for acquiring richer and more consistent building information. These modalities provide complementary perspectives. Satellite imagery enables macro-scale monitoring of urban morphology and land cover, for example through vegetation indices or nighttime luminosity. In contrast, street-view data captures fine-grained ground-level details relevant to human experience and building facades (Mashala et al., 2023; Ashik et al., 2024). However, these sources are typically processed in isolation, resulting in fragmented workflows and limited semantic richness. Thus, integrating them is essential to build semantically rich and scalable building datasets.

Moreover, researchers have proposed a range of computer vision approaches to extract building characteristics from urban imagery. These include CNN-based methods for building footprint detection (e.g., Faster R-CNN, YOLO), segmentation networks for evaluating walkability and façade transparency (e.g., DeepLabV3 +), and image classification models for land-use inference (e.g., ResNet, VGGNet). Yet, such models demand large volumes of labeled data and frequent retraining, limiting their scalability and adaptability to new tasks or regions (Birgani et al., 2024). Furthermore, they often operate within a single data modality, failing to leverage the complementary strengths of heterogeneous sources thus reinforcing the fragmentation issue noted above. The advent of large language models (LLMs) opens new opportunities to tackle both the incompleteness and fragmentation challenges. Initially designed for textual tasks, LLMs have evolved to handle multimodal inputs through vision-language pretraining and alignment techniques. These models can interpret both structured imagery and textual metadata, offering a unified and adaptable architecture for cross-source analysis (Yan et al., 2023). Compared with traditional CNNs, LLMs provide greater scalability, require less task-specific retraining, and offer more flexible interaction interfaces (Zhang et al., 2024). However, these sources are typically processed in isolation, resulting in fragmented workflows and limited semantic richness. Bridging these modalities is essential for generating semantically rich, spatially explicit, and scalable building characteristic datasets.

To fill these gaps, we propose BuildingMultiView, a unified framework that leverages fine-tuned LLMs to extract detailed building characteristics in a holistic manner from both satellite and street-view imagery. Based on a *meta*-analysis, we identify 11 key characteristics across multiple spatial levels and design image–prompt–label triplets for task-specific fine-tuning of GPT-4o. This framework bridges fragmented data modalities and enhances semantic richness. The framework is also applied across five U.S. cities spanning diverse climate zones and urban forms, resulting in a representative multi-perspective building characteristics dataset.

Our contributions are threefold. First, we construct a structured characteristics system informed by a *meta*-analysis of the literature, and develop BuildingMultiView, a framework that systematically integrates satellite and street-view imagery to address the incompleteness of existing datasets by generating detailed and interpretable building characteristics across multiple spatial scales. Second, we fine-tune GPT-4o using multimodal image–prompt–label triplets and task-specific prompt engineering. This enables the model to learn from both visual and contextual cues, improving generalizability across geographic settings and spatial levels (house, neighborhood, street), while also helping to bridge fragmented data modalities and modeling processes. Third, we construct a large-scale building characteristics dataset consisting of over 110,000 annotations covering 10,000 buildings in five U.S. cities spanning diverse climate zones. Compared to existing platforms such as OSM, our dataset offers semantically richer, and interoperable building descriptors, facilitating downstream applications in energy modeling, climate adaptation, and human-centered urban planning.

2. Literature review

2.1. Perspectives and characteristics for comprehensive building analysis

Urban building characteristics extraction has primarily been approached from two perspectives: satellite remote sensing and street-view imagery. Satellite data enables large-scale mapping and spatial analysis, supporting tasks such as land use classification, structural delineation, and change detection (Zhao et al., 2021; Li et al., 2022). It allows the extraction of key spatial characteristics such as roof type, building footprint, parking availability, and swimming pool presence, which are closely tied to energy performance, land use efficiency, and urban heat regulation (Lee et al., 2014; Jadhav & Gore, 2016; Demir et al., 2021).

In contrast, street-view imagery offers a human-scale perspective that captures detailed architectural and functional attributes often invisible from above. This includes characteristics such as building type, window-to-wall ratio, architectural style, and floor count—factors crucial for understanding facade design, energy use, and regulatory compliance (Kang et al., 2018; Huang & Gurney, 2016; Alwetaishi & Benjeddou, 2021). The combination of both views enhances our understanding of buildings not only as spatial units but also as functional, culturally embedded, and energy-relevant structures.

While many studies have examined these characteristics in isolation, recent research highlights the need for semantic-rich, multi-perspective datasets that integrate both physical and functional attributes (Biljecki & Chow, 2022). For instance, multi-source fusion has been applied to detect abandoned buildings (Zou & Wang, 2022) or assess flood vulnerability (Xing et al., 2023), but a comprehensive, globally scalable building characteristics database remains largely absent. Bridging this gap calls for improved data fusion and automation techniques. Integrating satellite remote sensing and street-view data offers a scalable and effective approach for capturing both spatial structures and facade-level details in the built environment.

2.2. Advanced methods for building information extraction

In recent decades, scholars have shifted from traditional urban building studies to computer vision, automating tasks like building classification and information extraction (Starzyńska-Grześ et al., 2023; Wang et al., 2021). However, complex backgrounds and diverse exteriors make reusable building extraction challenging, such as CNN-based methods (e.g., Faster R-CNN, DeepLabV3+, ResNet) are widely used for specific tasks but often fail to transfer effectively across different cities (Yang et al., 2018). Transformer-based models such as Swin Transformer and SegFormer improve multi-scale reasoning and semantic segmentation accuracy, particularly in remote sensing, but remain computationally expensive and often lack generalization across data modalities (Wang et al., 2022). These challenges have spurred growing interest in vision-language models that support prompt-based, multi-modal learning.

Recent years, LLMs have been widely applied for building information extraction by integrating semantic, geometric, and regulatory insights (Rillig et al., 2023; Wang et al., 2024). They enable scalable automation in GIS data processing (Zhang et al., 2024), simulation modeling (Xiao & Xu, 2024; Zhu et al., 2024), and geospatial annotation (Li et al., 2024). A key application is regulatory compliance, where GPT-based frameworks parse building codes and align them with BIM/IFC models, reducing errors and expediting approvals (Peng & Liu, 2023). LLMs also improve document intelligence, extracting insights from construction reports and maintenance manuals to enhance risk management (Shahinmoghadam et al., 2024). Beyond text-based tasks, multimodal models such as BLIP-2 and PaLI integrate visual and

linguistic data for automated facade analysis and structural classification (Yao et al., 2025; Li et al., 2023). Additionally, combining LLMs with knowledge graphs supports urban planning by identifying correlations between building density, land use, and socioeconomic factors (Fu et al., 2024; Pusch and Tim, 2024). Extending this direction, the OpenFACADES framework leverages multimodal LLMs and street-level imagery to automatically enrich building profiles with semantic and geometric attributes, demonstrating robust performance across diverse cities and supporting fine-grained urban analysis (Liang et al., 2025). These advancements highlight LLMs' role in large-scale building data integration, driving efficient, sustainable, and data-driven built environment analysis.

3. Methodology

We develop BuildingMultiView, a framework for extracting building-centric characteristics using satellite and street-view imagery. The workflow includes three steps: (1) constructing a multi-level building characteristics system through a *meta*-analysis of recent literature; (2) fine-tuning large language models with multimodal data and prompts; and (3) developing an automated pipeline for data collection and annotation. The following sections provide a detailed breakdown of each component.

3.1. Building centric characteristics system construction through Meta-Analysis

3.1.1. Meta-Analysis of building centric characteristics

To identify and standardize the building characteristics that are most commonly used in building environment research, we conduct a systematic *meta*-analysis of scholarly works published between 2019 and 2024 to ensure methodological rigor and transparency. For that, we follow the typical approach of systematic literature reviews: we select relevant keywords and search for a set of papers, after which we filter papers relevant for our study, and then extract relevant information from them. An initial search is performed on November 15, 2024, within the Web of Science database using the keywords “building exterior,” “building characteristics,” and “building surroundings,” yielding 6,285 studies broadly related to building and urban form analysis. To capture the multidimensional nature of building-related research, these works are categorized into three conceptual themes—human, energy, and green—which together reflect how building characteristics relate to social, energy, and ecological aspects of the urban environment. Incorporating these additional keywords results in 4,683 relevant publications (544 under human, 4,115 under energy, and 523 under green).

In alignment with our study’s focus on image-based analytics, we further adopt a filtering criterion emphasizing visually discernible characteristics—that is, building characteristics that can be objectively identified or inferred from visual data such as satellite or street-view imagery (e.g., roof type, building density, wall–window ratio, and façade transparency). This approach follows established practices in similar multimodal urban analytics frameworks, such as those developed by Biljecki and Chow (2022) for building morphology characteristics, which adopt a similar literature review analytical framework to ensure reproducibility and interoperability. Through this systematic refinement, we identify 54 studies that explicitly examine visual or façade-level features of buildings and ultimately retain 39 publications providing quantifiable and visually observable characteristics suitable for integration into the characteristic system.

Given our comprehensive review of building centric characteristics, we have identified and summarized the key characteristics that researchers find intriguing, which are presented in Fig. 1.

The Satellite House branch represents high-resolution satellite imagery centered on each building, covering a spatial extent of 100 m × 100 m. This branch is designed to capture characteristics observable from above the top, providing contextual information around individual buildings. Based on this imagery, several top-level–related characteristics are extracted, including roof type, swimming pool presence, and green cover density, which together describe residential form, surface materials, and immediate landscape composition. The Satellite Neighborhood branch represents 1000 m × 1000 m satellite imagery centered on each building cluster, capturing the broader spatial organization and development context of the surrounding urban area. This branch focuses on neighborhood-scale characteristics that describe the intensity, structure, and function of built environments. Specifically, five key characteristics are derived: building density, large-building count, building-group pattern, land use, and road density. Together, these characteristics illustrate how buildings are spatially arranged and how different land functions interact within the urban fabric. The Street-View branch utilizes ground-level imagery to capture the built environment from a pedestrian perspective, emphasizing features directly perceived and experienced by humans. This branch focuses on architectural and visual characteristics observable from the street façade. Key characteristics include wall–window ratio, property type, and floor

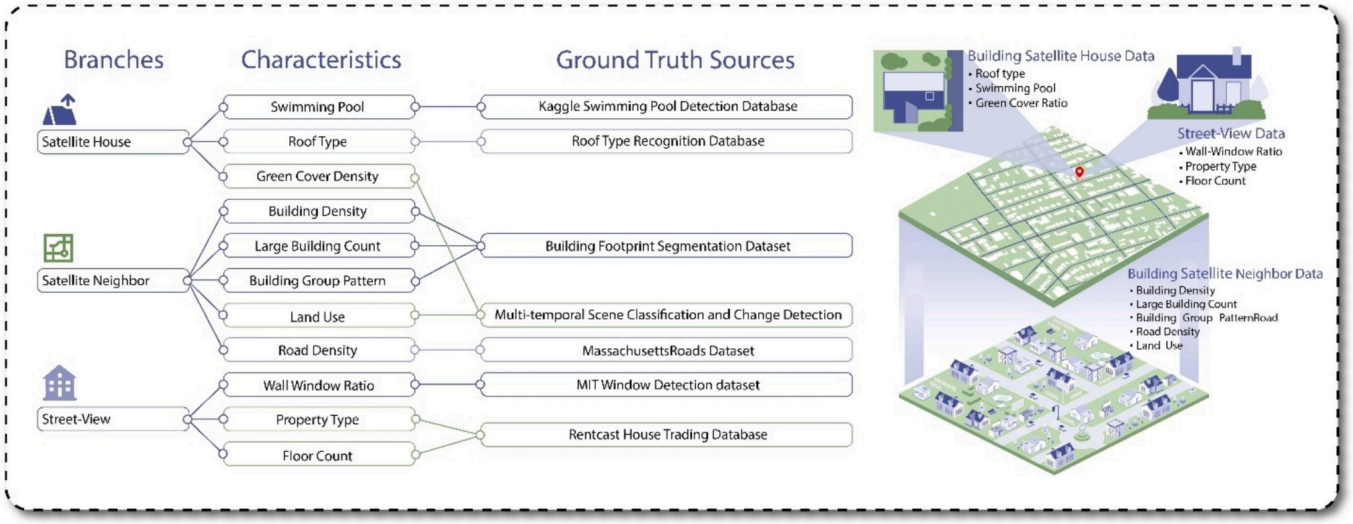


Fig. 1. The Characteristics and Ground Truth Source of Building-centric Database. Note: The left side lists 11 key building-related characteristics grouped under three visual data branches, Satellite House, Satellite Neighbor and Street-view, along with their corresponding ground truth data sources. These ground truth datasets are used to generate image–prompt–label triplets for fine-tuning the LLM. The right side illustrates the spatial scope and data types involved across the three branches.

count, which together describe structural form, building use, and vertical scale. These characteristics are essential for evaluating thermal performance, architectural style, and population density.

3.1.2. Ground truth construction based on Meta-Analysis

Based on the characteristic system identified through the previous *meta*-analysis, we collect and construct a corresponding ground truth dataset to support the subsequent fine-tuning of large language models (LLMs). This ground truth is organized across three major branches: (1) satellite house, (2) satellite neighbor, and (3) street-view level characteristics. Table 1 provides a summary of the characteristics, classification categories, and ground truth sources for each branch. Detailed descriptions and examples of each characteristic are presented in the following sections.

Satellite House Branch focuses on characteristics extracted from rooftop-level features visible in satellite imagery. Fig. 2 presents representative visual examples for each classification level under the three characteristics, highlighting their observable variations in satellite views.

Specifically, Swimming Pool (A) is sourced from the Kaggle Swimming Pool Detection Database (Coelho et al., 2021) and reflects the availability of recreational amenities, which are often associated with neighborhood socioeconomic status. Roof Type (B) is derived from the Roof Type Recognition Database (Alidoost & Arefi, 2018), capturing structural forms that adapt to climate and affect building performance (Wang et al., 2022; Zhang et al., 2021). Green Cover Density (C) is obtained from the Multi-temporal Scene Classification and Change Detection Dataset (Shao et al., 2020), representing the extent of vegetative cover and its role in supporting urban ecological health (Hami et al., 2019; Santamouris and Osmond, 2020).

Satellite Neighbor Branch includes five characteristics derived from building footprints, land cover, and road networks. These features describe spatial structure, urban morphology, and infrastructure distribution. Fig. 3 presents representative visual examples for each class within this branch, illustrating their observable patterns in overhead satellite imagery.

Building Density, Large Building Count, and Building Group Pattern (D) are derived from the Building Footprint Segmentation Dataset (Maggiori et al., 2017). Building density measures the proportion of land covered by buildings, indicating urbanization levels and land-use efficiency (Yang et al., 2021); large building count reflects the prevalence of commercial, industrial, or high-rise residential structures (Jaller et al., 2015)); and building group pattern classifies spatial arrangements, where clustered patterns indicate compact urban design, random patterns suggest unregulated development, and uniform patterns align with grid-based planning. Land Use (E), obtained from the Multi-temporal

Fig. 2. Visual Examples of Classification Levels in the Satellite House Branch. Note: This figure shows sample satellite images for three characteristics: (A) Swimming Pool, (B) Roof Type, and (C) Green Cover Density. Sources: Coelho et al., 2021, Alidoost & Arefi, 2018, Shao et al., 2020. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Scene Classification and Change Detection Dataset (Zhou et al., 2024), categorizes land into ten types for planning and regulatory purposes (Li et al., 2024). Finally, Road Density (F) is derived from the Massachusetts Roads Dataset (Ranzato et al., 2013), which serves as a training and reference dataset for defining the spatial distribution of road networks as a characteristic of transportation accessibility (Sahitya et al., 2020).The road density is calculated as the ratio of the road mask area to the total area, expressed as:

$$RoadDensity = \frac{RoadMaskArea}{TotalArea} \quad (1)$$

Street-view branch includes visual characteristics directly perceived from the pedestrian level, capturing façade elements and building typologies. Fig. 4 presents sample images across all classes of this branch, showcasing the observable differences in street-level appearance.

Table 1
The Overview of Characteristics and Class Labels Used for LLM Fine-Tuning with 11 Ground Truth Sources.

Branches	Characteristics	Class	Groundtruth Source
Satellite House	Swimming Pool	No 0, Yes 1	Kaggle Swimming Pool Detection Database
Satellite House	Roof Type	flat 0, gabled 1, hipped 2	Roof Type Recognition Database
Satellite House	Green Cover Density	0–10 % 0, 10–30 % 1, 30–60 % 2, 60 % and more 3	Multi-temporal Scene Classification and Change Detection
Satellite Neighbour	Building Density	0–10 % 0, 10–25 % 1, 25–100 % 2	Building Footprint Segmentation Dataset
Satellite Neighbour	Large Building Count	0 0, 1–5 1, 5–20 2, 20 and more than 3	Building Footprint Segmentation Dataset
Satellite Neighbour	Building Group Pattern	clustered 0, random 1, uniform 2	Building Footprint Segmentation Dataset
Satellite Neighbour	Land Use	agriculturalland 0, bareland 1, educationalland 2, greenspace 3, industrialland 4, publiccommercialland 5, residentialland 6, transportationland 7, waterbody 8, woodland 9	Multi-temporal Scene Classification and Change Detection
Satellite Neighbour	Road Density	0–10 % 0, 10–25 % 1, 25–100 % 2	Massachusetts Roads Dataset
Street-view	Wall Window Ratio	0–20 % 0, 20–40 % 1, 40–60 % 2, 60–100 % 3	MIT Window Detection Dataset
Street-view	Property Type	Single Family 0, Apartment 1, Multi-Family 2, Manufactured 3, Condo 4, Townhouse 5, other 6	Rentcast House Trading Database
Street-view	Floor Count	Numeric	Rentcast House Trading Database

Fig. 3. Visual Examples of Classification Levels in the Satellite Neighbor Branch. Note: This figure displays representative satellite images across five characteristics: (D) Building Density, Large Building Count, and Building Group Pattern; (E) Land Use; and (F) Road Density. Sources: Maggiori et al., 2017, Zhou et al., 2024.

Specifically, Wall Window Ratio (G), sourced from the MIT Window Detection Dataset (Simone et al., 2024), reflects the proportion of window area to wall surface, which affects ventilation, energy efficiency, and natural lighting conditions (Veillette et al., 2021).. The wall-window ratio is calculated as the ratio of the total window area to the total wall area, expressed as:

$$WallWindowRatio = \frac{WindowArea}{TotalWallArea} \quad (2)$$

Property Type and Floor Count (H) are obtained from the Rentcast House Trading Database (RentCast, 2020), which provides information

Fig. 4. Visual Examples of Classification Levels in the Street-view Branch. Note: This figure displays representative images for: (G) Wall Window Ratio, (H1) Property Type, and (H2) Floor Count. Sources: Simone et al., 2024, RentCast, 2020.

on building usage types and vertical scale. Property type is categorized into discrete classes, while floor count is recorded as a numerical value representing the number of stories in each building (Grace et al., 2004).

3.2. LLM Fine-Tuning

Fine-tuning LLMs is crucial for adapting pre-trained models to specialized tasks (Wei et al., 2023). In this study, we fine-tune GPT-4o on a multi-branch dataset comprising satellite house, satellite neighborhood, and street-view imagery, each labeled with structured ground-truth characteristics. For each characteristic, we design task-specific prompts to guide the model in interpreting spatial and visual characteristics. As shown in Fig. 5, the framework follows a two-step process: (1) pre-processing and formatting the data into image-prompt-label triples, and (2) fine-tuning GPT-4o with carefully designed task-specific prompts (Representative Prompt Samples are provided in Fig. A1).

First, we pre-process the ground truth datasets into standardized image–prompt–label triplets. Each sample contains a visual input, a task-specific natural language prompt, and a structured label. We design prompts to match specific characteristics: binary for swimming pool detection, categorical for roof type classification, and range-based for vegetation cover. To ensure semantic consistency, we constrain model outputs to a strict format (e.g., Filename: <name>, Type: <label >), simplifying downstream parsing. The dataset is split into training (60 %), validation (20 %), and test (20 %) sets, ensuring balanced coverage across all characteristics and cities.

In the second step, we fine-tune GPT-4o (base model: gpt-4o-2024–08-06) using a supervised multi-task learning scheme. The fine-tuned model was trained on approximately 110,000 image–prompt–label triplets across three data branches: satellite-house, satellite-neighborhood, and street-view. Rather than adopting a unified prompt or joint training scheme, we apply a branch-wise fine-tuning strategy in which each characteristic is treated as an independent task with customized prompts, label schemas, and objectives. The prompt templates (see Fig. A1. Representative Prompt Samples for Building-MultiView) are designed to guide the model's attention toward view-specific cues such as roof structure, façade materials, and spatial density, while maintaining a consistent output schema across branches. Fine-tuning is performed for three epochs using a batch size of 5, a learning rate of 1×10^{-5}, and the AdamW optimizer, with a fixed random seed (42) to ensure reproducibility. Model checkpoints are selected based on the highest validation F1 score on the held-out split. Evaluation metrics include accuracy, precision, recall, and F1 score, computed at the characteristic level.

3.3. Automated annotation framework for building characteristics

Building on the previous section, we develop a automate urban building characteristic generation framework using satellite and street-view imagery. As illustrated in Fig. 6, by integrating fine-tuned LLMs, it streamlines data collection, annotation, and analysis, producing a fine-grained building-centric characteristics dataset.

Our implementation begins by collecting geospatial inputs from multiple public APIs. OpenStreetMap (OSM) building footprints have been systematically evaluated for positional accuracy, completeness, and attribute reliability, showing high data quality in well-mapped regions such as North America and Western Europe, while global analyses of 13,189 urban centers found 1,848 cities exceeding 80 % completeness and 9,163 below 20 %, with higher coverage in Europe & Central Asia and North America (Herfort et al., 2023).These regions are among the best-mapped parts of the world, where continuous volunteer contributions and frequent updates ensure near-complete coverage of the built environment (Biljecki et al., 2023; Zhang & Zhu, 2018). Such well-established data quality makes OSM a reliable and widely accepted foundation for large-scale urban analysis and machine-learning–based annotation tasks. Accordingly, our case studies focus on cities in the United States and Western Europe, where OSM provides consistent and authoritative representations of building footprints suitable for robust model development and evaluation. Building footprints are retrieved from OpenStreetMap using the Overpass API, filtered by geometry type and minimum area thresholds to ensure urban relevance. Bounding boxes or city names are resolved into precise geometries, addresses, and heights via the Nominatim API. Each building polygon is associated with

Fig. 5. The Workflow and prompts of fine-tuning LLMs. Note: The diagram presents an LLM-based framework for satellite and street-view image analysis, with three branches (house, neighbor, and street-view) fine-tuned for extracting key information on buildings. For illustration purposes, three representative characteristics are selected from each branch to demonstrate the structure of the image–prompt–label triplets used in the fine-tuning process.

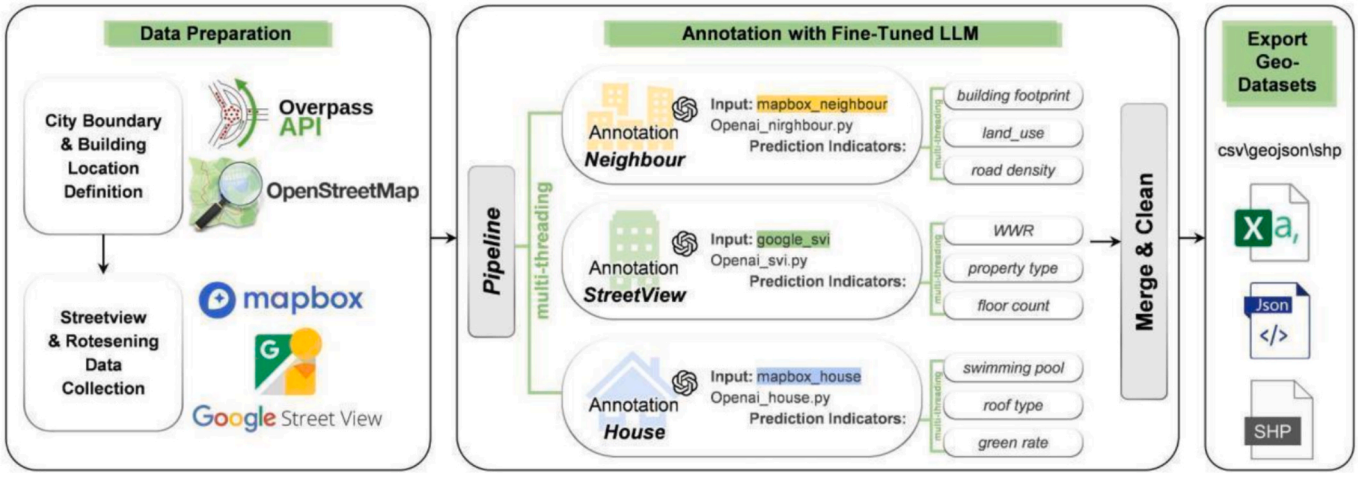


Fig. 6. BuildingMultiView pipeline with multithreaded annotation. Note: Each spatial branch uses dedicated scripts and prompts to extract specific characteristics. Outputs are automatically merged, cleaned, and exported into geo-formats for downstream analysis.

its centroid, which serves as the anchor point for image collection. For satellite imagery, we query Mapbox's static image API at dual resolutions (100 m2 for house level, 1 km2 for neighborhood level); for street-view imagery, the Google Street View API fetches panoramas within a 30-meter radius of each centroid, prioritizing front-facing facades using compass metadata when available. All retrieved data are structured into a JSONL-format annotation-ready dataset, indexed by building ID.

In addition, automatic annotation is performed via a multi-threaded GPT-4o pipeline, optimized for batch image-prompt processing. Each image is paired with a task-specific prompt template based on the characteristic type and image source. Post-processing includes spatial merging of annotations with raw geometry, removal of null or ambiguous predictions, and transformation into standardized geo-formats (CSV, GeoJSON, Shapefile). The annotation framework supports multi-threaded processing (via Python's concurrent.futures) and includes an error logging and retry mechanism to handle API limits or LLM timeouts. All outputs are publicly available on Hugging Face with version control, enabling reproducibility and future use in built environment studies.

4. Experiment

4.1. Study area and sampling data

To ensure consistent data sources and demonstrate the workflow's transferability, we select the United States as the study region for its high-quality, openly accessible, and standardized datasets (e.g., building information, climatic classifications, and street-view imagery). These resources enable comparability across cities, making the U.S. an effective testbed for method validation. Within this context, representative cities are selected across different climatic zones to capture diverse environmental and urban conditions. Following the U.S. Department of Energy's Building America Program (Antonopoulos et al., 2022) and high-GDP urban centers (U.S. BEA, 2022), five cities are chosen: San Francisco (marine), San Diego (hot-dry), Salt Lake City (cold), Austin (humid-hot), and New York City (mixed-humid). Fig. 7 presents the geographic distribution of the five cities across climate zones, with inset maps showing city boundaries and sampling locations. A total of 10,000 samples are collected, with 2,000 data points from each city.

4.2. Result and analysis

We present the results and analytical findings derived from the annotated building dataset covering five representative climate zones across the United States. We organize the analysis into three parts to comprehensively evaluate the framework and interpret the extracted building information. The first part focuses on the classification performance and validation results (Section 4.2.1), which confirm the accuracy and robustness of the automated annotation process. The second part summarizes the statistical profiles and spatial distributions of key building characteristics (Section 4.2.2), providing an overview of the constructed database. The third part explores the relationships between building characteristics and regional climatic conditions (Section 4.2.3), serving as a downstream analysis that demonstrates how the annotated data can be utilized to examine climate-responsive patterns in urban form.

4.2.1. Classification performance and validation

To evaluate our model's reliability, we compare the fine-tuned model against the original GPT-4o without task-specific tuning, as well as several competitive baselines including Vision Transformer, Gemini, and ResNet50, using a held-out test set. The dataset is split into 80 % for training and 20 % for testing, with no overlapping building instances. All characteristics are assessed using standard classification metrics, including accuracy, precision, recall, and F1 score.

In the fine-tuning process, we start from the base model gpt-4o-2024–08-06 and apply supervised multi-task learning across three branches—satellite-house, satellite-neighborhood, and street-view. Training is conducted for three epochs with a batch size of 5, a learning rate of 1×10^{-5}, and the AdamW optimizer under a fixed random seed (42). The best checkpoint is selected based on validation F1 score. For the vision-based baselines, both ViT and ResNet50 are implemented using the TIMM framework with pretrained ImageNet weights. ViT adopts the vit_base_patch16_224 architecture, and ResNet uses resnet50. Both models are trained for 10 epochs with a batch size of 32, learning rate 3×10^{-4}, and AdamW optimization. Input images are resized to 224 × 224 pixels and normalized to [0.5, 0.5, 0.5]. The Gemini baseline is configured in zero-shot mode under equivalent input settings for cross-model comparability.

As shown in Table 2, the fine-tuned model achieves significant improvements: accuracy increases from 55.04 % to 80.83 %, and F1 score rises from 45.66 % to 79.77 %, consistently outperforming both the zero-shot GPT-4o baseline and other competitive vision models. This confirms the effectiveness of fine-tuning in enhancing predictive performance and underscores the robustness of our approach.

At the branch level, performance varies across characteristic groups. In the satellite house branch, swimming pool detection achieves 96.00 % accuracy (F1: 96.18 %), while roof type and green cover ratio both exceed 85 %, reflecting the visual clarity and distinctiveness of property-

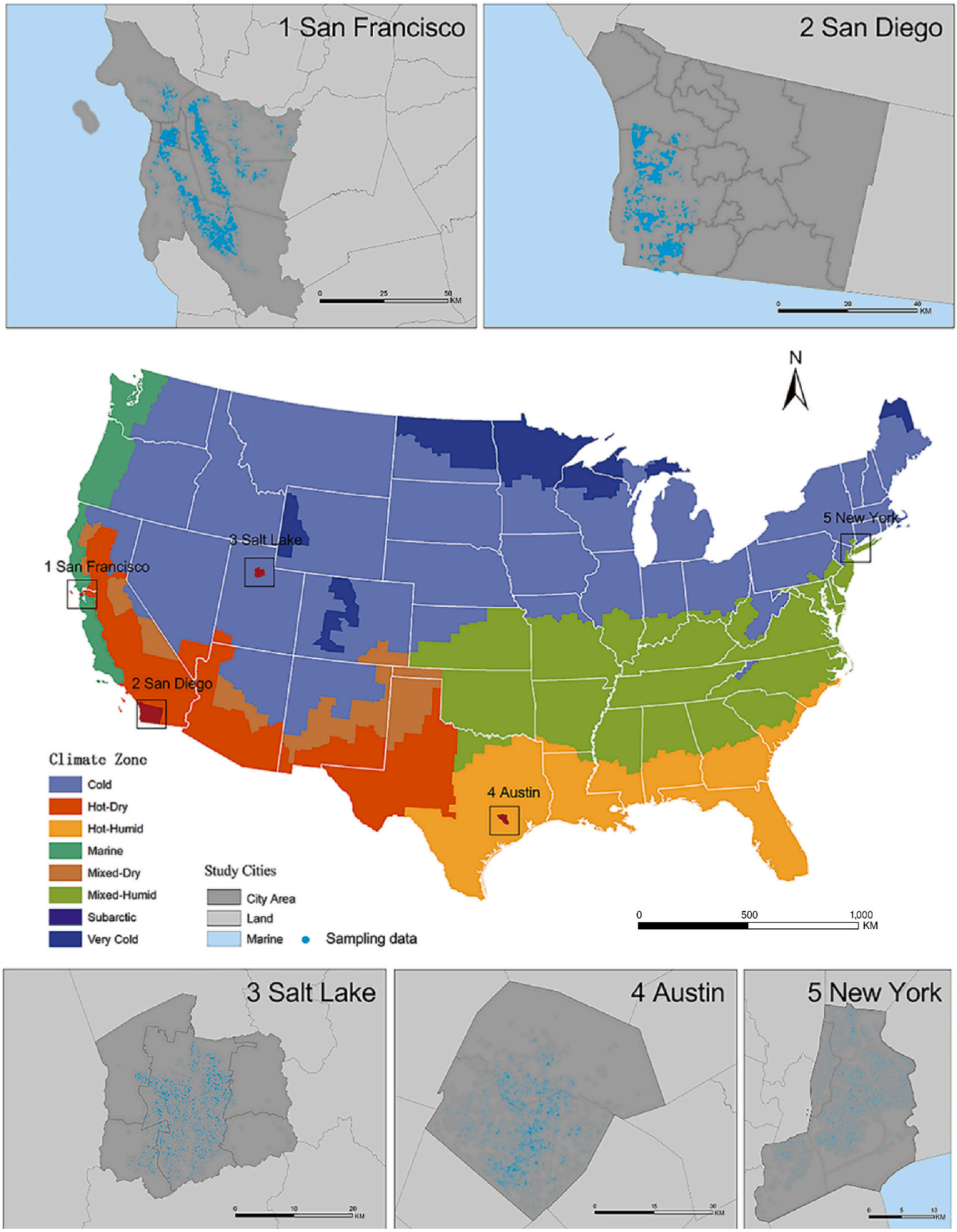


Fig. 7. Study Area of BuildingMultiView. Note: This figure shows the five study cities—San Francisco, San Diego, Salt Lake City, Austin, and New York City—selected for their diverse climates and economic profiles. The main map highlights their locations and climate zones, with insets showing city boundaries and sample distributions (2,000 buildings per city).

Table 2
Performance of Different Task-Driven LLM Fine-Tuning Models and Competitive Baselines. This table summarizes the performance of fine-tuned models across 11 building characteristics Evaluation metrics include Accuracy, Precision, Recall, and F1 Score, all calculated on a held-out test set. "Manual Interpretation" refers to the accuracy of human labeled results on sampled subsets. Fine-tuned models achieve substantial gains over the base GPT-4o and other competitive baselines (Vision Transformer, Gemini, ResNet50).

Branches	Characteristics	Accuracy	Precision	Recall	F1 Score	Manual Interpretation
Satellite House Level	Swimming Pool	96.00 %	96.52 %	96.00 %	96.18 %	91.32 %
	Roof Type	84.78 %	84.86 %	84.78 %	84.76 %	86.48 %
	Green Cover Ratio	83.75 %	85.58 %	83.75 %	84.33 %	78.81 %
Satellite Neighbour Level	Building Density	92.50 %	93.37 %	92.50 %	92.30 %	83.37 %
	Large Building Count	72.50 %	76.67 %	72.50 %	72.07 %	78.35 %
	Neighbor Group Pattern	85.00 %	87.25 %	85.00 %	85.69 %	83.37 %
	Land Use	81.30 %	78.66 %	80.31 %	79.16 %	76.31 %
	Road Density	94.98 %	94.94 %	94.98 %	94.96 %	87.68 %
Street-view Level	Wall Window Ratio	77.80 %	75.95 %	77.81 %	76.24 %	76.25 %
	Property Type	81.62 %	80.68 %	81.62 %	80.44 %	81.92 %
	Floor Count	83.08 %	82.24 %	83.08 %	82.64 %	88.08 %
Fine-Tuning Models Avg		**80.83 %**	**79.62 %**	**80.84 %**	**79.77 %**	**82.90 %**
ChatGPT-4o		55.04 %	63.29 %	52.32 %	45.66 %	N/A
Vision Transformer		72.80 %	64.28 %	72.80 %	66.63 %	N/A
Gemini		52.86 %	55.68 %	50.14 %	13.27 %	N/A
ResNet50		78.55 %	75.17 %	78.55 %	76.62 %	N/A

level features in aerial imagery. In the satellite neighborhood branch, road density (83.41 %) and building group pattern perform solidly, but large building count is lower (72.50 %), likely due to inter-city variation in high-rise distribution and dataset imbalance. For the street-view branch, floor count (88.08 %) and property type (84.12 %) perform strongly, whereas wall-to-window ratio (76.25 %) and land use (76.31 %) are less accurate, affected by occlusion, angle distortion, and semantic ambiguity in street-level imagery. These results indicate that while the model captures building characteristics effectively, refining characteristic definitions and expanding training datasets could further improve accuracy.

To further validate predictive performance, we conduct a manual inspection benchmark using 1,000 randomly sampled instances across five cities. Model predictions are compared to manually verified labels

Fig. 8. Representative Audit Samples from Five U.S. Cities.Note: This figure presents satellite and street-level imagery for five representative buildings selected across different U.S. cities as part of a manual interpretation audit. Red bounding boxes highlight the targe buildings evaluated for prediction accuracy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for an objective accuracy assessment. As also shown in Table 2 "Manual Interpretation" line, the model demonstrates strong accuracy across most characteristics, with swimming pool detection (91.32 %), floor count (88.08 %), and road density (87.68 %) performing well. However, the wall-to-window ratio (76.25 %) and land use (76.31 %) show lower accuracy due to classification ambiguities and dataset variability.

To further address concerns regarding variability and statistical significance, we conducted additional experiments under 10 different random seeds (42–51) and applied stratified bootstrap resampling (B = 2000) to estimate 95 % confidence intervals. Compared with the baseline model (55.04 %), the fine-tuned model consistently outperformed with statistical significance (one-sided test, $p < 0.001$). Across all seeds, the average F1 score reached 76.35 % (95 % CI [74.19 %, 80.08 %]), demonstrating robust and stable improvements over the baseline.

To complement these large-scale validation metrics, we also perform a detailed manual interpretation audit on five representative buildings across distinct U.S. cities. As shown in Fig. 8, we compile satellite and street-view imagery across the three branches (satellite house, satellite neighborhood, and street-view) for each sample. Table 3.1 presents the predicted labels generated by the BuildingMultiView framework, while Table 3.2 documents the discrepancies identified through manual interpretation, and characteristics are marked in red reflect prediction errors in building characteristics.

The manual interpretation audit shows that the framework performs reliably across most building characteristics, demonstrating strong generalizability in diverse urban contexts. However, several misclassifications emerge and are worth highlighting. In Austin, the predicted property type is labeled as "Single Family," yet the presence of a Greek-letter fraternity sign ("ΦΚΕ") strongly suggests the building functions as a fraternity house—beyond what standard imagery alone can confidently resolve. In New York City, the model underestimates the floor count, likely due to limited perspective and occlusion common in narrow street-views, which hinder full facade visibility. The case in San Diego represents an extreme outlier: the target building is a high-end, secluded estate distinctly different from typical residential structures. The swimming pool is not detected, possibly because it is partially outside the 100 × 100 m aerial tile used for prediction. Additionally, street-view images of this property are blurred, contributing to inaccuracies in wall-to-window ratio and property type prediction.

Despite challenging cases, the framework correctly labels most characteristics even under atypical visual and environmental conditions, demonstrating strong robustness. The results also highlight the importance of manual audits in identifying nuanced errors that may be overlooked by aggregate metrics. Overall, the BuildingMultiView framework effectively improves building characteristic extraction through multi-perspective data integration and fine-tuned learning.

4.2.2. Distribution and correlation analysis of building characteristics

This subsection summarizes the statistical profiles and spatial distributions of key building characteristics, providing an overview of the constructed building- characteristic database. The analysis aims to describe the overall composition and variability of the automatically generated characteristics, thereby assessing the representativeness and interpretability of the dataset. By examining how these characteristics are distributed within and across cities, we identify major morphological patterns and validate whether the framework captures meaningful urban structures. The results presented in Figs. 9–11 illustrate the spatial distributions of eleven building characteristics in New York City, cross-city comparisons among five representative urban regions, and the correlation structure among all characteristics.

Spatial Distributions of Characteristics. This study collects street-view and satellite imagery to extract 11 key building-related characteristics across five U.S. cities, offering a comprehensive view of urban form and function. Among them, New York City is chosen for visualization due to its diverse and vertical urban structure (Fig. 9). The spatial distributions of these characteristics reveal sharp contrasts between the dense urban core and peripheral areas. In Manhattan, building density and floor count reach their highest levels, flat roofs dominate, and large buildings cluster tightly, reflecting commercial and high-rise residential land use. Wall-to-window ratios are also elevated, indicating façade openness consistent with glass-intensive architecture. By contrast, outer boroughs such as Staten Island and parts of Queens exhibit more gabled and hipped roofs, lower densities, and higher green cover. Swimming pools, though sparse overall, appear more frequently in these low-density residential areas.

Property types also vary across space: central districts are dominated by apartments and condominiums, while peripheral neighborhoods display a more diverse mix, including single-family homes and townhouses. Land use transitions gradually from public and commercial zones in the core to green, transportation, and lower-density residential areas in the suburbs. Road density mirrors this gradient, with dense street networks in Manhattan giving way to more fragmented layouts outward. Building group patterns also shift from uniform or clustered forms in dense cores to more dispersed arrangements in peripheral zones.

Comparative Distribution. Fig. 10 compares urban building and environmental characteristics across five cities. Austin shows the highest green cover (60 %+), a high share of gabled and hipped roofs, frequent swimming pools, and predominantly single-family homes. Its building density is moderate, with few large buildings and a mostly uniform group pattern, reflecting suburban characteristics; New York City stands out with high building and road density, a dominance of flat roofs, frequent 3–5 story buildings, and high WWR values. It is primarily composed of apartments and condos, and has the largest buildings, typical of a dense vertical urban core; San Francisco exhibits mixed roof

Table 3.1
Predicted Building Characteristic Labels for Five Representative Building Samples. Red text highlights characteristics with prediction errors identified during manual interpretation.

City	Austin	New York	San Diego	San Francisco	Salt Lake City
OSM ID	380,917,039	241,842,474	558,916,442	267,111,803	462,574,596
Latitude	30.2852095	40.6451466	32.8792244	37.7635182	40.7854052
Longitude	−97.7474545	−73.9615802	−117.2498552	−122.4703935	−111.9215708
Swimming Pool	No	No	**No**	No	No
Roof Type	Hipped	Flat	Hipped	Flat	Hipped
Green Cover Density	10–30 %	10–30 %	10–30 %	10–30 %	30–60 %
Building Density	25–100 %	25–100 %	0–10 %	25–100 %	10–25 %
Large Building Count	5–20	5–20	1–5	5–20	1–5
Building Group Pattern	Uniform	Uniform	Clustered	Uniform	Uniform
Land Use	Residential land	Residential land	Residential land	Transportation land	Residential land
Road Density	10–25 %	10–25 %	0–10 %	10–25 %	10–25 %
Wall Window Ratio	0–20 %	0–20 %	**0–20 %**	0–20 %	0–20 %
Property Type	**Single Family**	Apartment	**Single Family**	Single Family	Single Family
Floor Count	2	**2**	1	2	1

Table 3.2
Manual Audit Findings: Discrepancies Between Predicted and Observed Characteristics. Each row highlights the predicted label, visual assessment, and a concise explanation of the likely cause of misclassification.

City	Characteristic	Predicted	Actual (Visual)	Notes
Austin	Property Type	Single Family	Shared Housing (Fraternity)	Fraternity signage indicates non-single family usage
New York	Floor Count	2	≥3	Undercounted; upper floors obscured in street view
San Diego	Swimming Pool	No	Yes	Missed due to partial pool visibility outside the 100 × 100 m tile
San Diego	Wall Window Ratio	0–20 %	Higher (estimated visually)	Underestimated due to occlusion and poor street-view angle
San Diego	Property Type	Single Family	Likely Estate/Other	Luxurious estate, possibly not typical single family

types, moderate green space, and a combination of residential and parkland land uses. It also features noticeable townhouse presence, high road density, and mostly clustered group patterns; San Diego shows balanced distributions in roof types and green cover, and land use favors residential and green areas, with a slight presence of swimming pools; Salt Lake City has the lowest building and road density, dominant sloped roofs, mostly single-family homes, and minimal large buildings. Its urban fabric is highly uniform, with relatively high vegetation and low vertical development.

Correlation Patterns. To explore the relationships among key Characteristics across five cities, we conduct a correlation analysis (Fig. 11). Several consistent patterns emerge across cities. Large building count strongly correlates with density (0.75–0.79), reflecting the concentration of high-rises. Roof type aligns with building group pattern (0.38–0.42), indicating design consistency in clustered areas. Land use also correlates with group pattern (0.60–0.67), showing zoning's impact on spatial layout. Wall-to-window ratio negatively correlates with density (−0.28 to −0.35), implying reduced facade openness in compact, energy-conscious zones.

In terms of city-specific patterns, New York City exhibits a strong correlation between floor count and wall-to-window ratio (0.62), indicating that taller buildings tend to have more enclosed facade styles. San Francisco shows a high correlation between roof type and land use (0.55), potentially due to its architectural controls and zoning constraints. In Salt Lake City, green cover density has a relatively weak correlation with building density (0.32), which diverges from the general trend and reflects its dispersed, low-density development. San Diego displays strong alignment between building group pattern and large building count (0.61), suggesting its denser built zones are spatially clustered. In Austin, swimming pool presence is notably correlated with property type (0.59), which is consistent with its dominance of single-family housing and warm climate conditions.

4.2.3. Climate-Driven analysis of architectural and environmental characteristics

We then explore the relationships between building characteristics and regional climatic conditions, serving as a downstream analysis that demonstrates how the annotated data can be utilized to examine climate-responsive patterns in urban form. Since climate fundamentally shapes building form and urban environments, this analysis aims to test whether the proposed framework can capture such climate-driven regularities. Establishing this linkage is critical for evaluating the transferability and explanatory power of the workflow beyond the chosen case studies. Therefore, we cluster the five representative cities based on temperature and precipitation and incorporate LLM-inferred climate-responsive keywords to interpret architectural and environmental characteristics. By connecting the derived characteristics to exogenous climate conditions, this analysis verifies that the framework captures climate-consistent semantics and enhances interpretability and generalizability across regions (Figs. 12–13).

Clustering Based on Temperature and Precipitation. Our clustering analysis (Fig. 12) examines architectural and environmental characteristics across five cities, focusing on temperature and precipitation.

Warmer cities such as Austin and San Diego have higher green cover density, providing natural cooling and supporting sustainability goals. Cities with higher rainfall, such as San Francisco and New York City, adapt through efficient drainage systems and sloped roofs for durability in humid conditions. High precipitation also correlates with greater vegetation, emphasizing urban greenery's role in climate adaptability. Roof types vary by climate, with flat roofs common in warmer regions for heat management and sloped roofs in colder areas for snow removal and structural stability.

Word Clouds of Climate-Responsive Characteristics. To explore the relationship between urban climates and building characteristics, we generate word clouds for five representative U.S. cities using LLM-based reasoning (Fig. 13). Each city reflects climate-adaptive architectural strategies. San Francisco (marine climate) emphasizes wind resistance, corrosion control, and compact, low-rise buildings suited to coastal and seismic conditions. San Diego (hot-dry) features water conservation, drought-tolerant landscaping, and passive cooling. Salt Lake City (cold climate) highlights insulation, snow mitigation, and heating efficiency. Austin (hot-humid) focuses on ventilation, reflective materials, and humidity control. New York City (mixed-humid) integrates insulation, ventilation, and energy-efficient design to address seasonal variability.

The consistency between extracted keywords and known climatic characteristics provides indirect validation of our framework's reasoning capacity. These results suggest that the fine-tuned LLM captures underlying climatic logic in built-environment semantics, revealing not only expected architectural adaptations (e.g., thermal insulation in cold zones) but also subtler patterns, such as the co-occurrence of wind and corrosion terms in marine cities or moisture-control language in dry and humid areas.

5. Discussion

5.1. Optimizing annotation accuracy through prompt engineering

Prompt engineering plays a key role in improving annotation accuracy and data quality in large-scale built environment analysis. Well-structured prompts help language models capture nuanced characteristics while minimizing errors. In our previous work, the BuildingView framework (Li et al., 2024), we used a single merged prompt for zero-shot annotation, which minimized manual effort but increased model and pipeline complexity. While simple characteristics such as window color and floor count remained accurate, complex ones, involving architectural style and roof materials, suffered from misclassification due to the lack of targeted instructions.

To address these challenges, BuildingMultiView shifts from a single-prompt approach to a characteristic-specific strategy. Each characteristic receives a dedicated prompt with refined instructions and tailored examples to improve accuracy. This modular approach isolates tasks, reducing confusion and preventing errors in one characteristic from affecting others. For instance, a solar panel detection prompt can incorporate technical thresholds and examples of partially obscured panels, while a building style prompt can distinguish between historical and contemporary designs. By refining prompts, BuildingView-Turbo enhances precision and interoperability across building analysis.

This methodological shift from a unified prompt to a characteristic-specific approach raises the question of whether the observed

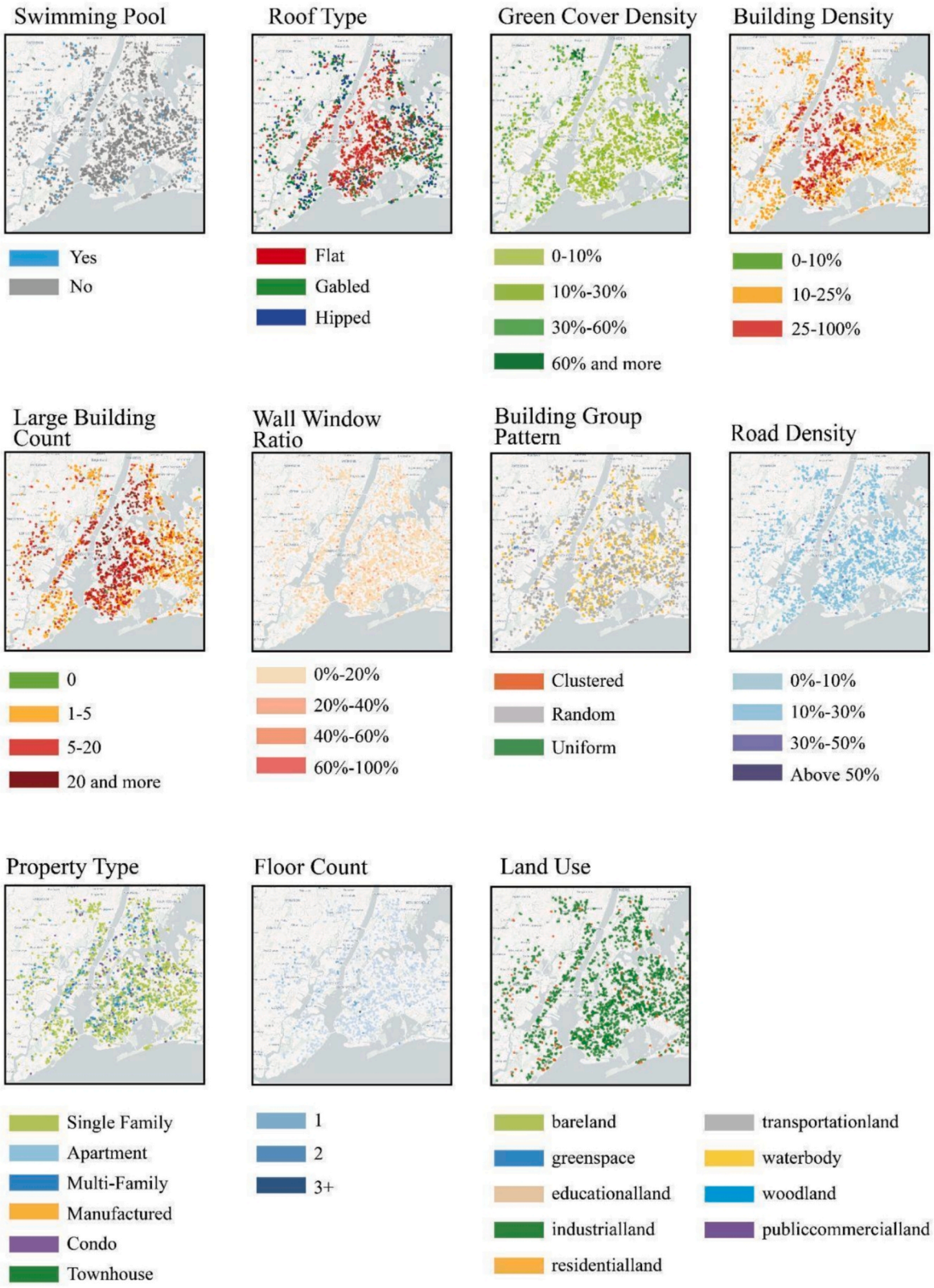


Fig. 9. Distribution of 11 Key Building Characteristics in New York City.

improvements can be attributed to structural design rather than incidental factors. To provide quantitative evidence, we conducted ablation experiments comparing single-branch (characteristic-specific) and unified models under both base and fine-tuned settings (Table 4).

The results clearly show that branch-wise fine-tuning achieves superior accuracy, precision, recall, and F1 score, supporting our claim that characteristic-specific prompts lead to more reliable and transferable annotations.

Another key advantage of this characteristic-specific approach is that it embodies a broader methodological contribution to multimodal urban

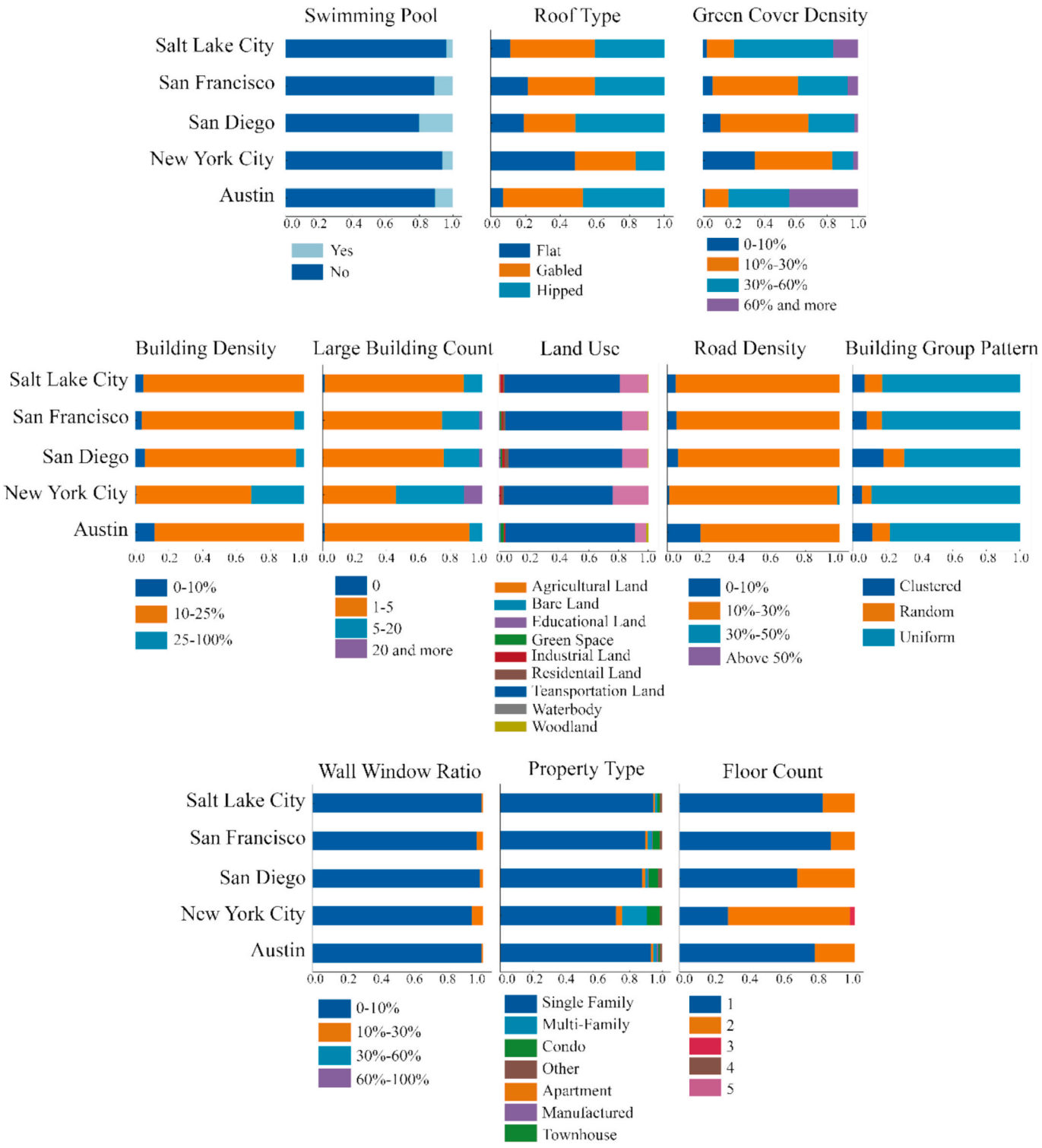


Fig. 10. Comparative Distribution of Architectural and Environmental Characteristics in Five Cities.

analytics. Existing studies have explored vision–language approaches for building analysis, but none have proposed a unified and interpretable workflow for multi-scale building characteristic evaluation. For instance, Pan et al. (2024) demonstrated zero-shot building attribute extraction using vision–language models, but their approach was limited to single-view inference without structured adaptation. Yao et al. (2024) focused on façade condition assessment restricted to visual degradation, while Chen et al. (2025) introduced a multimodal framework for city-scale spatial intelligence that did not address micro-scale or building-level characterization. In contrast, our framework explicitly models multi-view complementarity, hierarchical fine-tuning, and prompt-controlled semantic alignment, establishing a systematic and reproducible workflow for cross-scale semantic reasoning. By enforcing branch-wise fine-tuning and semantic alignment across satellite and street-view imagery, the proposed design transforms multimodal adaptation from task-specific optimization into a generalizable learning principle. This methodological rigor, coupled with comprehensive data design and interpretability, positions the framework as a bridge between

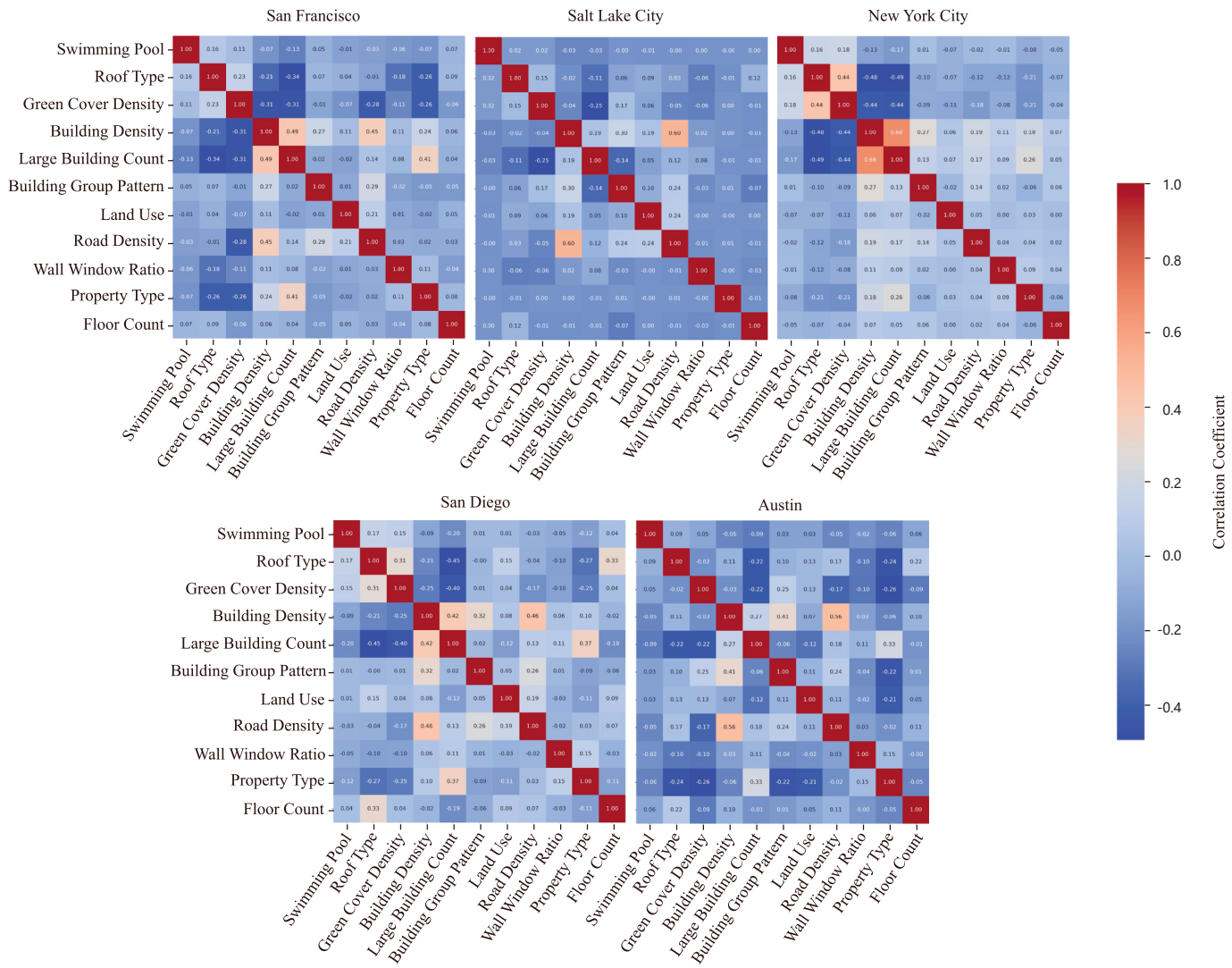


Fig. 11. Correlation Analysis Between Architectural and Environmental Characteristics in Five Cities.

engineering implementation and analytical advancement in urban informatics.

Beyond accuracy and adaptability, the proposed framework also demonstrates computational efficiency and environmental sustainability. Across five representative U.S. cities, approximately 10,000 buildings were processed for 11 characteristics, involving around 20,000 Mapbox imagery requests and 10,000 Google Street View queries. The end-to-end workflow, including multi-attribute inference with OpenAI-4o, incurred an estimated total cost of about USD 700, corresponding to 70–110 million tokens, or roughly 640–1,000 tokens (≈0.6 cents) per building–characteristic unit. Following established methodologies for estimating AI energy consumption and emissions (Henderson et al., 2020; Jegham et al., 2025), this workload equates to approximately 5–15 kWh of electricity use and 2–6 kg CO_2e in total, including both inference and imagery retrieval. These values indicate that the framework maintains a modest computational and carbon footprint while scaling efficiently across cities, supporting its practical and sustainable deployment for large-scale urban analysis (Samsi et al., 2023).

5.2. Limitations and Considerations

Our study presents a comprehensive framework for urban building characteristics extraction, incorporating multiple data perspectives and advanced predictive modeling. Throughout the development and validation process, we have carefully examined key challenges and their potential impact.

First, while the proposed framework demonstrates strong performance and generalizability, several potential directions for improvement remain, including enhancing multi-label representation, addressing data coverage and temporal inconsistencies, and expanding the framework's analytical adaptability. The framework is inherently extensible due to its hierarchical and standardized structure, which organizes building-centric characteristics into three analytical levels—building, block, and urban—each following a consistent three-step process: (1) defining the visual data source (e.g., satellite, aerial, or street-view imagery); (2) extracting relevant visual features (such as façade openness, greenery proportion, or roof reflectance); and (3) transforming these features into standardized quantitative characteristics within a unified range. Because every characteristic follows unified definition, extraction, and normalization procedures, each serves as an independent analytical module that can be directly extended. This modular design allows new visually discernible characteristics—such as façade texture complexity, shading distribution, or color composition—to be seamlessly integrated by defining their data source and transformation rule, without altering the existing analytical logic. Furthermore, the hierarchical organization supports flexible aggregation and weighting of characteristics across spatial levels, while standardized references ensure compatibility across multiple imagery types

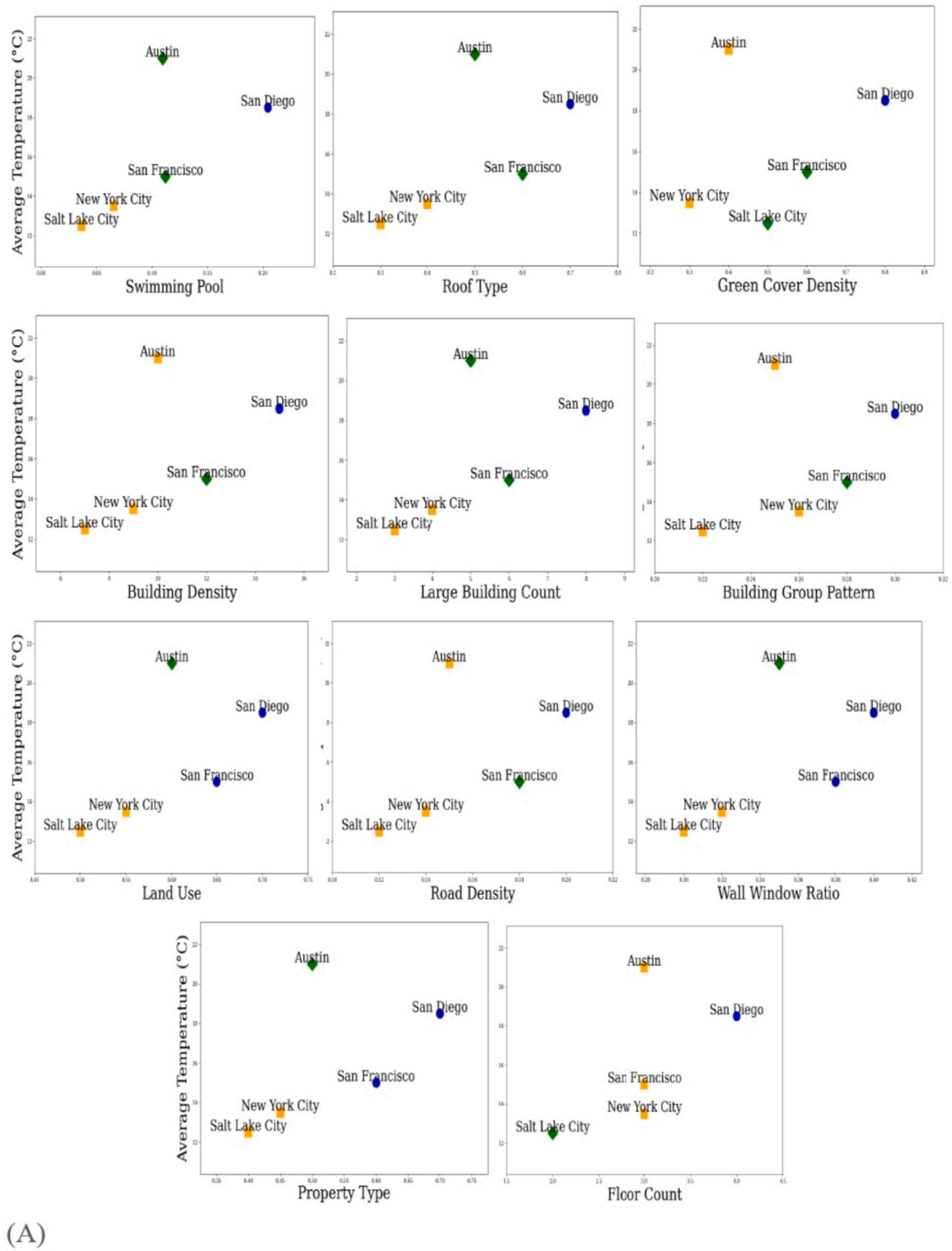
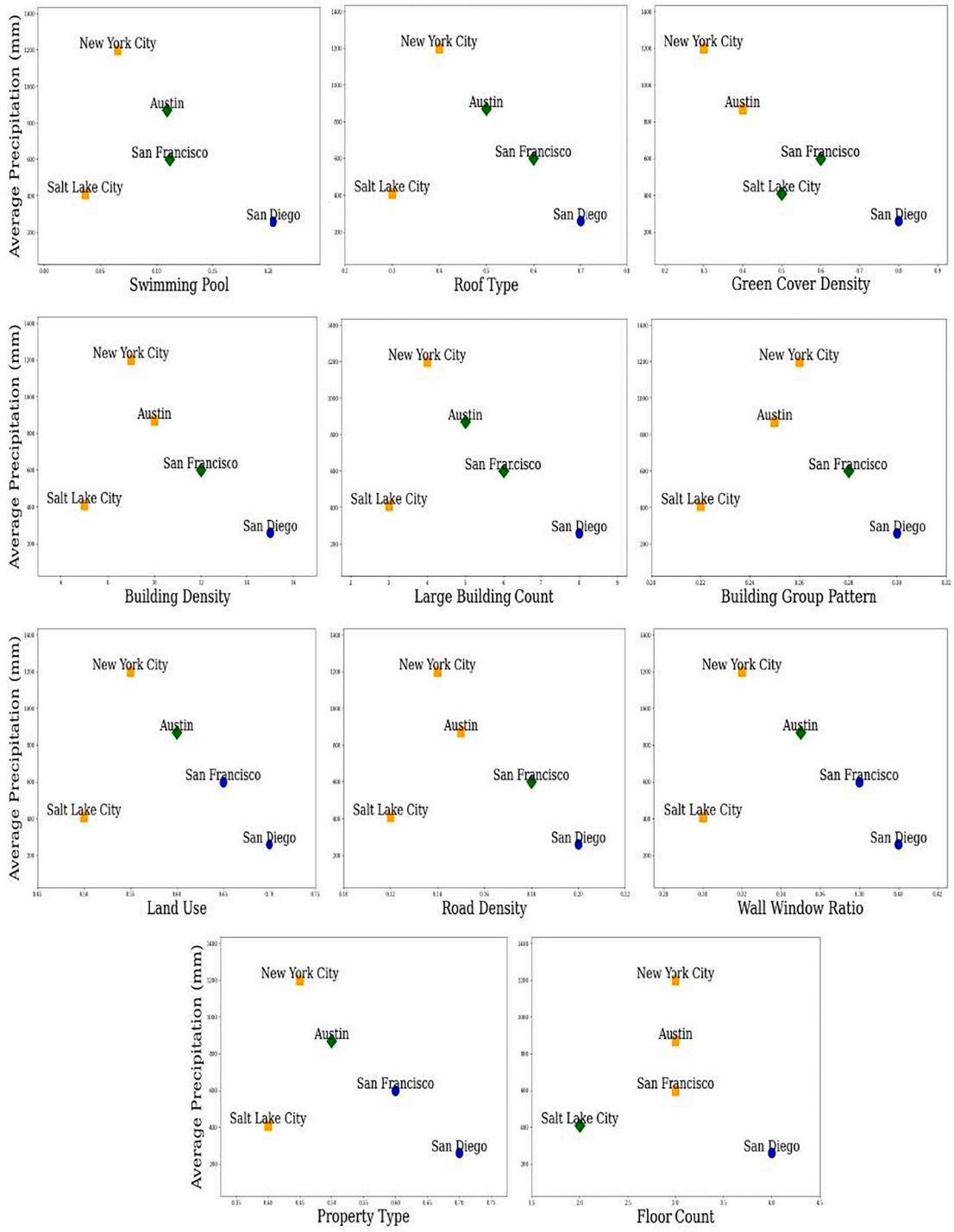


Fig. 12. Clustering Analysis Based on Climate Factors: (A) Clustering by Annual Average Temperature, (B) Clustering by Annual Average Precipitation. Note: Colors denote the three clusters derived from the climate-factor clustering results, where cities sharing the same color are grouped into the same cluster in both panel (A) and panel (B).

and resolutions. As new visual data and computational techniques continue to emerge, the framework can readily evolve to incorporate additional or non-visual characteristics, thereby enhancing its scalability, systematic comprehensiveness, and long-term adaptability for future multimodal urban analytics.

Another limitation concerns the treatment of multi-label building characteristics, as a single building may simultaneously exhibit multiple functional or physical attributes (e.g., a residential building with a swimming pool and a green roof). BuildingMultiView captures these diverse attributes through a three-branch, multi-level design: the



(B)

Fig. 12. (*continued*).

Satellite House branch focuses on individual-level features (e.g., roof type, swimming pool, green cover density), the Satellite Neighbor branch captures neighborhood-scale characteristics (e.g., building density, land use, road density), and the Street View branch extracts façade and functional properties (e.g., property type, floor count, wall–window ratio). Although explicit multi-label annotation is not applied to every single building, the land use characteristic is modeled as a multi-label classification task based on the Multi-temporal Scene Classification and Change Detection Dataset (Zhou et al., 2024), which includes ten land-use categories such as Agricultural land, Residential land, Commercial land, and Green space. This multi-level structure already enables the framework to recognize coexisting attributes at different spatial scales. In future work, we will further improve the handling of multi-label and mixed-use cases by incorporating multimodal transformer-based fusion and multi-label learning strategies (Zhou et al., 2023) to jointly model correlated features across remote sensing, POI,

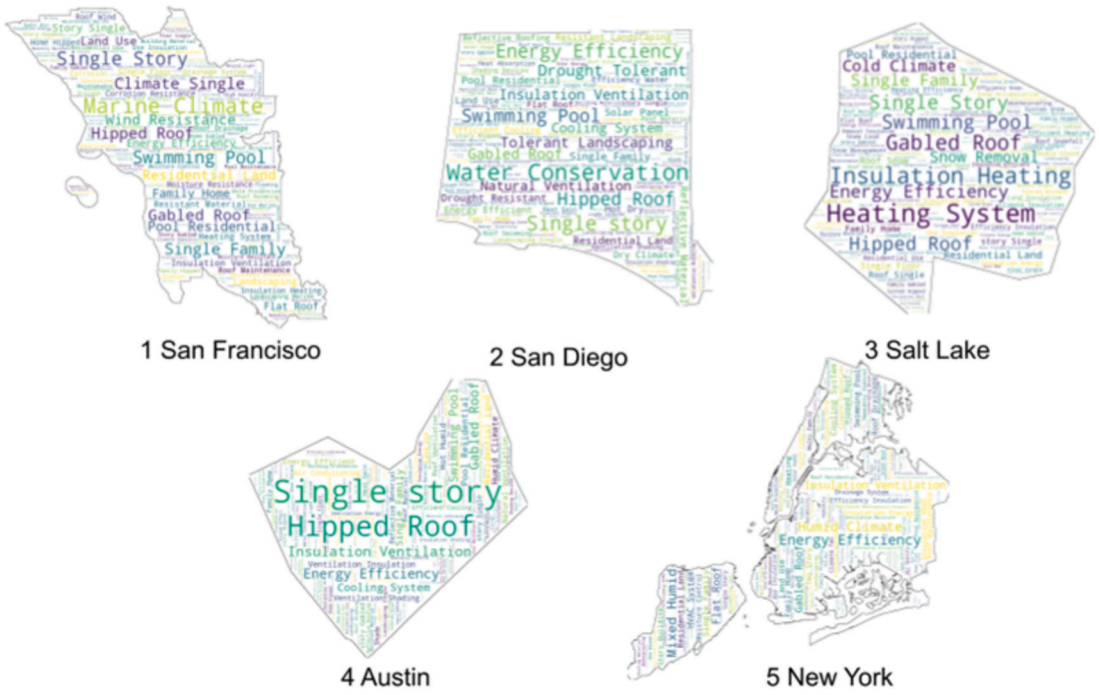


Fig. 13. Climate-Responsive Word Clouds.

Table 4
Ablation study comparing single-branch and unified models under base and fine-tuned settings.

Model	Accuracy	Precision	Recall	F1 Score
Single-Branch (fine-tuning)	**80.83 %**	**79.62 %**	**80.84 %**	**79.77 %**
Single-Branch (base model)	55.04 %	63.29 %	52.32 %	45.66 %
Unified Model (fine-tuning)	72.46 %	65.42 %	59.57 %	59.74 %
Unified Model (base model)	27.75 %	9.19 %	10.16 %	9.02 %

Note: Results demonstrate that branch-wise fine-tuning consistently outperforms the unified approach across accuracy, precision, recall, and F1 score, highlighting the advantages of characteristic-specific prompts for reliable and transferable building annotation.

and street-view modalities, thereby enhancing the scalability and generalization of the proposed framework.

A further consideration involves the uneven global coverage of OSM building footprints and street-view imagery. While OSM exhibits substantial regional disparities, with higher completeness in Europe and North America and lower coverage across many regions in the Global South, and street-view data show similar spatial heterogeneity, these limitations are less critical in our experiments because the study area is restricted to U.S. cities, where both datasets are relatively complete and reliable (Herfort et al., 2023; Fan et al., 2025).When extending the framework to a global scale, however, this issue should not be viewed merely as a constraint requiring fallback mechanisms. Instead, it highlights the extensibility of the multi-perspective design. A key strength of BuildingMultiView is that the Satellite House and Satellite Neighbor branches rely on satellite imagery, one of the most universally available data sources, which allows the workflow to function robustly in regions with limited street-view coverage. More importantly, the modular and standardized image–prompt–label architecture enables seamless integration of emerging global datasets such as the Global Building Atlas (Zhu et al., 2025) and OpenBuildingMap (Oostwegel et al., 2025). These datasets combine OpenStreetMap, Microsoft's Global ML Footprints, and Google Open Buildings to provide harmonized, semantically rich, and near-global building coverage with uniform completeness and positional accuracy. Incorporating such datasets not only fills spatial gaps but also strengthens the framework's geographic transferability beyond well-mapped regions. For areas lacking street-view imagery, synthetic street-level perspectives derived from high-resolution remote sensing and complementary crowdsourced platforms such as Mapillary and KartaView can provide additional façade information, particularly in places where proprietary sources like Google Street View are unavailable (Hou & Biljecki, 2022). Together, these capabilities demonstrate that the BuildingMultiView framework is not limited to a specific set of inputs but is capable of evolving alongside future data ecosystems, enhancing its scalability and adaptability for global multimodal urban analytics.

Finally, an important consideration for further improvement is the temporal inconsistency among OSM, satellite, and street-view imagery, since these sources follow independent update cycles and may lag behind real-world changes. As a result, localized discrepancies may occur in characteristics such as roof type, wall–window ratio, or floor counts when buildings are renovated or re-purposed. These local errors can propagate to downstream analyses: for example, energy simulation results may be biased if envelope or fenestration attributes are wrong (Nouri et al., 2024); land-use classification may mislabel building functions when the functional use attribute is outdated; and socio-economic inference models that rely on built-form proxies could be skewed by such attribute errors. Nevertheless, both our empirical results and recent multi-source building dataset research suggest that while temporal mismatches introduce noise at the attribute level, their influence on macro-scale morphology or pattern-level metrics is limited. For instance, the CMAB dataset (Zhang et al., 2025) employs multi-source imagery and street-view data to derive multi-attribute building characteristics and demonstrates stable performance across cities despite temporal heterogeneity. Similarly, the HISDAC-US project (Leyk & Uhl, 2018) integrates parcel, remote sensing, and building datasets across decades and acknowledges temporal misalignment while preserving consistency in large-scale settlement analyses. In our case, despite the heterogeneous temporal provenance of the data, the fine-tuned model maintains stable predictive performance, with an average F1 score of 76.35 % (95 % CI [74.19 %, 80.08 %]) across ten random seeds, and strong accuracy for key characteristics such as swimming pool (91.32 %) and floor count (88.08 %). These results demonstrate that the

BuildingMultiView framework is resilient to temporal inconsistencies when applied to large-scale urban morphology and functional pattern analysis. Looking ahead, integrating temporal metadata (e.g., capture dates) to weight attribute confidence, as well as incorporating automated change-detection modules to identify potentially outdated annotations, would strengthen the framework's robustness.

6. Conclusion

This study introduces BuildingMultiView, a unified framework that integrates satellite and street-view imagery with fine-tuned large language models to extract multi-scale, building-centric characteristics. By leveraging structured image–prompt–label triplets and tailored fine-tuning strategies, the model enables accurate, transferable, and automated annotation of 11 key characteristics across spatial levels. Applied across five U.S. cities spanning distinct climate zones, the framework achieves substantial improvements in predictive performance and reveals spatial, functional, and climate-responsive patterns in the built environment. BuildingMultiView also demonstrates the feasibility of combining vision-language models with multi-perspective urban imagery for large-scale, standardized building analysis. Its modular pipeline and open dataset offer a scalable foundation for future applications in urban planning, energy modeling, and climate adaptation. This work contributes both a methodological advance in characteristic extraction and a reproducible infrastructure to support data-driven urban research and decision making.

7. Author Agreement

I, Zongrong Li, certify that all authors have seen and approved the final version of the manuscript titled "BuildingMultiView: Powering Multi-Scale Building Characterization with Large Language Models and Multi-Perspective Imagery." We warrant that this manuscript is the author's original work, has not been previously published, and is not under consideration for publication elsewhere.

Appendix A. Appendix

8. Conflict of interest Statement

I, Zongrong Li, declare that there are no conflicts of interest regarding the research and publication of this paper titled "BuildingMultiView: Powering Multi-Scale Building Characterization with Large Language Models and Multi-Perspective Imagery." I have no financial, personal, or professional relationships that could be perceived to influence the findings or interpretations in this work.

CRediT authorship contribution statement

Zongrong Li: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yunlei Su:** Writing – original draft, Validation, Software, Investigation. **Filip Biljecki:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Wufan Zhao:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 42401567), the Tertiary Education Scientific Research Project of Guangzhou Municipal Education Bureau (Grant No. 2024312159), the Guangzhou Municipal Science and Technology Bureau Program (Grant No. 2025A03J3640), the Open Fund of the Technology Innovation Center for 3D Real Scene Construction and Urban Refined Governance, Ministry of Natural Resources (Grant No. 2024PF-1), and the AI Research and Learning Base of Urban Culture, Guangdong Provincial Department of Education (Grant No. 2023WZJD008).

Satellite House Branch

TASK=SwimmingPool

System:

You are given a remote sensing image. Determine the Vegetation Cover Density class for the image. The Vegetation Cover Density classes are as follows: 0 (0 - 10%), 1 (10-30%), 2 (30-60%), 3 (60% and more). Output format: 'Filename: <filename>, Vegetation_Cover_Class: <class>'. Only output the filename and Vegetation Cover Density Class.

TASK=RoofType

System:

You are given a remote sensing image. Determine the Roof type class for the image. The Roof type classes are as follows: 0 (flat), 1 (gabled), 2 (hipped). Output format: 'Filename: <filename>, Type_Class: <class>'. Only output the filename and Type class.

TASK=GreenRatio

System:

You are given a remote sensing image. Determine the Vegetation Cover Density class for the image. The Vegetation Cover Density classes are as follows: 0 (0 - 10%), 1 (10-30%), 2 (30-60%), 3 (60% and more). Output format: 'Filename: <filename>, Vegetation_Cover_Class: <class>'. Only output the filename and Vegetation Cover Density Class.

Satellite Neighbour Branch

TASK=Footprint (Building Density、Large Building Count、Neighbor Grounp Pattern)

System:

Classify some building footprint types for a remote sensing image considering the following parameters: Building Density 0 (0-10%), 1 (10-25%), 2 (25%-100%); Large Building Count:0(0), 1(1-5),2 (5-20), 3 (20 and more than); Building Distribution Patterns: 0 (clustered), 1(random), 2 (uniform). Output format: Filename: <filename>, BD: <density_class>, LB: <building_count_class>, BDP: <Patterns_class>'.

TASK=Landuse

System:

Classify the land use type for a remote sensing image. Possible classes: 0 (agriculturalland), 1 (bareland), 2 (educationalland), 3 (greenspace), 4 (industrialland), 5 (publiccommercialland), 6 (residentialland), 7 (transportationland), 8 (waterbody), 9 (woodland). Output: 'Filename: <filename>, Type_Class: <class>'. Each image can have multiple classes.

TASK=Road

System:

Classify some Road Network types for a remote sensing image, considering the following parameters: Road Coverage Ratio(RCR)0 (0%-10%),1 (10%-30%),2 (30%-50%),3 (Above 50%). Output format: 'Filename: <filename>, RCR: <rcr_class>''.

Street-view Branch

TASK=WallWindowRatio

System:

You are given a street view image. Determine the WWR class for the image based on its window-to-wall ratio (WWR). The WWR classes are as follows: 0 (0-20%), 1 (20-40%), 2 (40-60%), 3 (60-100%). Output format: 'Filename: <filename>, WWR_Class: <class>'. Only output the filename and WWR class.

TASK=PropertyType

System:

You are given a streetview image. Determine the Building Property type class for the image. The Building Property type classes are as follows: Single Family 0, Apartment 1, Multi-Family 2, Manufactured 3, Condo, 4 Townhouse 5, other 6. Output format: 'Filename: <filename>, Type_Class: <class>'. Only output the filename and Type class.

TASK=FloorCount

System:

You are given a street view image. Determine the floor count of the building in the image. Output format: 'Filename: <filename>, FloorCount: <Count>'. Only output the filename and floorcount.

Fig. A1. Representative Prompt Samples for BuildingMultiView.

Data availability

Data will be made available on request.

References

Alidoost, F., Arefi, H., 2018. A CNN-based approach for automatic building detection and recognition of roof types using a single aerial image. PFG – J. Photogramm. Remote Sens. Geoinf. Sci. 86 (5–6), 235–248. https://doi.org/10.1007/s41064-018-0060-5.

Alwetaishi, M., Benjeddou, O., 2021. Impact of window to wall ratio on energy loads in hot regions: a study of building energy performance. Energies 14 (4), 1080. https://doi.org/10.3390/en14041080.

Antonopoulos, C.A., Gilbride, T.L., Margiotta, E.R., Kaltreider, C.E., 2022. Guide to determining climate zone by county: Building America and IECC 2021 updates. Report No. PNNL-33270, Pacific Northwest National Laboratory (PNNL), Richland, WA, United States.

Biljecki, F., Chow, Y.S., 2022. Global building morphology indicators. Comput. Environ. Urban Syst. 95, 101809. https://doi.org/10.1016/j.compenvurbsys.2022.101809.

Biljecki, F., Chow, Y.S., Lee, K., 2023. Quality of crowdsourced geospatial building information: a global assessment of OpenStreetMap attributes. Build. Environ. 237, 110295. https://doi.org/10.1016/j.buildenv.2023.110295.

Biljecki, F., Zheng, L., Milojevic-Dupont, N., Creutzig, F., 2021. Open government geospatial data on buildings for planning sustainable and resilient cities. *arXiv preprint*.

Birgani, S.A., Zadeh, S.S., Davari, D.D., Ostovar, A., 2024. Deep learning applications for analysing concrete surface cracks. Int. J. Appl. Data Sci. Eng. Health 1 (2), 69–84. https://ijadseh.com/index.php/ijadseh/article/view/16.

Chen, J., Wang, H., Li, J., Liu, Y., Dong, Z., & Yang, B. (2025). SpatialLLM: From Multi-modality Data to Urban Spatial Intelligence. ArXiv.org. https://arxiv.org/abs/2505.12703.

Coburn, A., Vartanian, O., Chatterjee, A., 2017. Buildings, beauty, and the brain: a neuroscience of architectural experience. J. Cogn. Neurosci. 29 (9), 1521–1531. https://doi.org/10.1162/jocn_a_01146.

Coelho, C., Costa, M.F.P., Ferrás, L.L., Soares, A.J., 2021. Object detection with RetinaNet on aerial imagery: the Algarve landscape. In: Gervasi, O. et al. (Eds.), *Comput. Sci. Appl. – ICCSA 2021*, Lect. Notes Comput. Sci., vol. 12953, Springer, Cham, pp. 501–516. doi: 10.1007/978-3-030-86960-1_35.

Demir, S., Basaraner, M., Taskin Gumus, A., 2021. Selection of suitable parking lot sites in megacities: a case study for four districts of Istanbul. Land Use Policy 111, 105731. https://doi.org/10.1016/j.landusepol.2021.105731.

Fan, Z., Feng, C.-C., Biljecki, F., 2025. Coverage and bias of street view imagery in mapping the urban environment. Comput. Environ. Urban Syst. 117, 102253. https://doi.org/10.1016/j.compenvurbsys.2025.102253.

Ashik, F.R., Sreezon, A.I.Z., Rahman, M.H., Zafri, N.M., Labib, S.M., 2024. Built environment influences commute mode choice in a global south megacity context: insights from explainable machine learning approach. J. Transp. Geogr. 116, 103828. https://doi.org/10.1016/j.jtrangeo.2024.103828.

Fu, J., Han, H., Su, X., Fan, C., 2024. Towards human-AI collaborative urban science research enabled by pre-trained large language models. *Urban*. Inform 3 (1). https://doi.org/10.1007/s44212-024-00042-y.

Grace Wong, K.M., 2004. Vertical cities as a solution for land scarcity: the tallest public housing development in Singapore. Urban Des. Int. 9 (1), 17–30. https://doi.org/10.1057/palgrave.udi.9000108.

Hami, A., Abdi, B., Zarehaghi, D., Maulan, S.B., 2019. Assessing the thermal comfort effects of green spaces: a systematic review of methods, parameters, and plants' attributes. Sustain. Cities Soc. 49, 101634. https://doi.org/10.1016/j.scs.2019.101634.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J., 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. J. Mach. Learn. Res. 21 (248), 1–43. https://www.jmlr.org/papers/v21/20-312.html.

Herfort, B., Lautenbach, S., Porto, J., Anderson, J., Zipf, A., 2023. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. Nat. Commun. 14 (1). https://doi.org/10.1038/s41467-023-39698-6.

Hou, Y., Biljecki, F., 2022. A comprehensive framework for evaluating the quality of street view imagery. Int. J. Appl. Earth Obs. Geoinf. 115, 103094. https://doi.org/10.1016/j.jag.2022.103094.

Huang, J., Gurney, K.R., 2016. The variation of climate change impact on building energy consumption to building type and spatiotemporal scale. Energy 111, 137–153. https://doi.org/10.1016/j.energy.2016.05.118.

Jadhav, A., Gore, N.G., 2016. Cost optimization of roof top swimming pool. Int. Res. J. Eng. Technol. 3, 1320–1322.

Jaller, M., Wang, X., Holguín-Veras, J., 2015. Large urban freight traffic generators: opportunities for city logistics initiatives. J. Transp. Land Use 8 (1), 51–67.
Jegham, N., Abdelatti, M., Elmoubarki, L., & Hendawi, A. (2025). *How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference*. ArXiv.org. https://arxiv.org/abs/2505.09598.
Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building instance classification using street view images. ISPRS J. Photogramm. Remote Sens. 145, 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006.
Knezevic, M., Donaubauer, A., Moshrefzadeh, M., Kolbe, T.H., 2022. Managing urban digital twins with an extended catalog service. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* X-4/W3-2022, 119–126. doi: 10.5194/isprs-annals-x-4-w3-2022-119-2022.
Lee, K.E., Williams, K.J.H., Sargent, L.D., Farrell, C., Williams, N.S., 2014. Living roof preference is influenced by plant characteristics and diversity. Landsc. Urban Plan. 122, 152–159. https://doi.org/10.1016/j.landurbplan.2013.09.011.
Leyk, S., Uhl, J.H., 2018. HISDAC-US, historical settlement data compilation for the conterminous United States over 200 years. Sci. Data 5 (1). https://doi.org/10.1038/sdata.2018.175.
Li, J., Huang, X., Tu, L., Zhang, T., Wang, L., 2022a. A review of building detection from very high resolution optical remote sensing images. Gisci. Remote Sens. 59 (1), 1199–1225. https://doi.org/10.1080/15481603.2022.2101727.
Li, J., Li, C., 2024. Characterizing urban spatial structure through built form typologies: a new framework using clustering ensembles. Land Use Policy 141, 107166. https://doi.org/10.1016/j.landusepol.2024.107166.
Li, J., Li, D., Savarese, S., Hoi, S., 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. Int. Conf. Mach. Learn. (ICML), Proc. Mach. Learn. Res. 202, 19730–19742. https://proceedings.mlr.press/v202/li23q.
Li, Y., Peng, L., Wu, C., Zhang, J., 2022b. Street view imagery (SVI) in the built environment: a theoretical and systematic review. Buildings 12 (8), 1167. https://doi.org/10.3390/buildings12081167.
Li, Z., Su, Y., Zhu, C., Zhao, W., 2024. BuildingView: constructing urban building exteriors databases with street view imagery and multimodal large language model. *arXiv preprint* arXiv:2409.19527.
Li, Z., Xu, J., Wang, S., Wu, Y., Li, H., 2024. StreetviewLLM: extracting geographic information using a chain-of-thought multimodal large language model. *arXiv preprint* arXiv:2411.14476.
Liang, X., Chang, J.H., Gao, S., Zhao, T., Biljecki, F., 2024. Evaluating human perception of building exteriors using street view imagery. Build. Environ. 263, 111875. https://doi.org/10.1016/j.buildenv.2024.111875.
Liang, X., Xie, J., Zhao, T., Stouffs, R., Biljecki, F., 2025. OpenFACADES: an open framework for architectural caption and attribute data enrichment via street view imagery. ISPRS J. Photogramm. Remote Sens. 230, 918–942. https://doi.org/10.1016/j.isprsjprs.2025.10.014.
Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? The INRIA aerial image labeling benchmark. In: Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), pp. doi: 10.1109/igarss.2017.8127684.
Mashala, M.J., Dube, T., Mudereri, B.T., Ayisi, K.K., Ramudzuli, M., 2023. A systematic review on advancements in remote sensing for assessing and monitoring land use and land cover changes impacts on surface water resources in semi-arid tropical environments. Remote Sens. 15 (16), 3926. https://doi.org/10.3390/rs15163926.
Memduhoglu, A., Basaraner, M., 2023. Semantic enrichment of building functions through geospatial data integration and ontological inference. Environ. Plan. B Urban Anal. City Sci. 51 (4), 923–938. https://doi.org/10.1177/23998083231206165.
Nouri, A., van Treeck, C., Frisch, J., 2024. Sensitivity assessment of building energy performance simulations using MARS meta-modeling in combination with sobol' method. Energies 17 (3), 695. https://doi.org/10.3390/en17030695.
Oostwegel, L.J.N., Schorlemmer, D., Guéguen, P., 2025. From footprints to functions: a comprehensive global and semantic building footprint dataset. Sci. Data 12 (1). https://doi.org/10.1038/s41597-025-06132-z.
Pan, F., Jeon, S., Wang, B., Mckenna, F., Yu, S.X., 2024. Zero-shot building attribute extraction from large-scale vision and language models. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2022, 8632–8641. https://doi.org/10.1109/wacv57701.2024.00845.
Peng, J., Liu, X., 2023. Automated code compliance checking research based on BIM and knowledge graph. Sci. Rep. 13 (1). https://doi.org/10.1038/s41598-023-34342-1.
Pusch, L., Tim, C., 2024. Combining LLMs and knowledge graphs to reduce hallucinations in question answering. *arXiv preprint* arXiv:2409.04181.
Ranzato, M., Mnih, V., Susskind, J.M., Hinton, G.E., 2013. Modeling natural images using gated MRFs. IEEE Trans. Pattern Anal. Mach. Intell. 35 (9), 2206–2222. https://doi.org/10.1109/tpami.2013.29.
RentCast, 2020. Rental property rates and market trends. https://www.rentcast.io/.
Rillig, M.C., Ågerstrand, M., Bi, M., Gould, K.A., Sauerland, U., 2023. Risks and benefits of large language models for the environment. Environ. Sci. Technol. 57 (9), 3464–3466. https://doi.org/10.1021/acs.est.3c01106.
Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., & Gadepally, V. (2023). *From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference*. ArXiv.org. https://arxiv.org/abs/2310.03003.
Sahitya, K.S., Prasad, C.S.R.K., 2020. Evaluation of opportunity based urban road network accessibility using GIS. Spat. Inf. Res. 28 (4), 487–493. https://doi.org/10.1007/s41324-019-00309-6.
Santamouris, M., Osmond, P., 2020. Increasing green infrastructure in cities: impact on ambient temperature, air quality and heat-related mortality and morbidity. Buildings 10 (12), 233. https://doi.org/10.3390/buildings10120233.
Shahinmoghadam, M., Kahou, S.E., Motamedi, A., 2024. Neural semantic tagging for natural language-based search in building information models: implications for practice. Comput. Ind. 155, 104063. https://doi.org/10.1016/j.compind.2023.104063.
Shao, Z., Zhou, W., Deng, X., Zhang, M., Cheng, Q., 2020. Multilabel remote sensing image retrieval based on fully convolutional network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 318–328. https://doi.org/10.1109/jstars.2019.2961634.
Simone, D., Biswas, S., Wu, O., 2024. Window to wall ratio detection using SegFormer. *arXiv preprint* arXiv:2406.02706.
Starzyńska-Grześ, M.B., Roussel, R., Jacoby, S., Asadipour, A., 2023. Computer vision-based analysis of buildings and built environments: a systematic review of current approaches. ACM Comput. Surv. 55 (13s). https://doi.org/10.1145/3578552.
Veillette, D., Rouleau, J., Gosselin, L., 2021. Impact of window-to-wall ratio on heating demand and thermal comfort when considering a variety of occupant behavior profiles. Front. Sustain. Cities 3, 700794. https://doi.org/10.3389/frsc.2021.700794.
Wang, C., Antos, S.E., Triveno, L.M., 2021. Automatic detection of unreinforced masonry buildings from street view images using deep learning-based image segmentation. Autom. Constr. 132, 103968. https://doi.org/10.1016/j.autcon.2021.103968.
Wang, L., Fang, S., Meng, X., Li, R., 2022a. Building extraction with vision transformer. IEEE Trans. Geosci. Remote Sens. 60, 1–11. https://doi.org/10.1109/TGRS.2022.3186634.
Wang, P., Liu, Z., Zhang, X., Zhang, H., Chen, X., Zhang, L., 2022b. Adaptive building roof combining variable transparency shape-stabilized phase change material: application potential and adaptability in different climate zones. Build. Environ. 222, 109436. https://doi.org/10.1016/j.buildenv.2022.109436.
Wang, S., Hu, T., Xiao, H., Li, Y., Zhang, C., Ning, H., Zhu, R., Li, Z., Ye, X., 2024. GPT, large language models (LLMs) and generative artificial intelligence (GAI) models in geospatial science: a systematic review. Int. J. Digit. Earth 17 (1). https://doi.org/10.1080/17538947.2024.2353122.
Wang, X., Li, H., Sodoudi, S., 2022c. The effectiveness of cool and green roofs in mitigating urban heat island and improving human thermal comfort. Build. Environ. 217, 109082. https://doi.org/10.1016/j.buildenv.2022.109082.
Wang, Y., Li, S., Teng, F., Lin, Y., Wang, M., Cai, H., 2022d. Improved Mask R-CNN for rural building roof type recognition from UAV high-resolution images: a case study in Hunan Province. China. Remote Sens. 14 (2), 265. https://doi.org/10.3390/rs14020265.
Wei, L., Jiang, Z., Huang, W., Sun, L., 2023. InstructionGPT-4: a 200-instruction paradigm for fine-tuning MiniGPT-4. *arXiv preprint* arXiv:2308.12067.
Xiao, T., Xu, P., 2024. Exploring automated energy optimization with unstructured building data: a multi-agent based framework leveraging large language models. Energ. Buildings 322, 114691. https://doi.org/10.1016/j.enbuild.2024.114691.
Xing, Z., Yang, S., Zan, X., Dong, X., Yao, Y., Liu, Z., Zhang, X., 2023. Flood vulnerability assessment of urban buildings based on integrating high-resolution remote sensing and street view images. Sustain. Cities Soc. 92, 104467. https://doi.org/10.1016/j.scs.2023.104467.
Yan, Y., Wen, H., Zhong, S., Chen, W., Chen, H., Wen, Q., Zimmermann, R., Liang, Y., 2023. UrbanCLIP: learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. *arXiv preprint* arXiv:2310.18340.
Yang, H.L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A.D., Bhaduri, B.L., 2018a. Building extraction at scale using convolutional neural network: mapping of the United States. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (8), 2600–2614. https://doi.org/10.1109/jstars.2018.2835377.
Yang, J., Shi, B., 2021. Scale or size? an analysis of the factors that affect building density: evidence from high-density central urban zones in Asia. no pagination Urban Des. Int. 26. https://doi.org/10.1057/s41289-021-00165-7.
Yang, J., Su, J., Xia, J., Jin, C., Li, X., Ge, Q., 2018. The impact of spatial form of urban architecture on the urban thermal environment: a case study of the Zhongshan District, Dalian, China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 11(8), 2709–2716. DOI: 10.1109/jstars.2018.2808469.
Yao, S., Ghorbany, S., Forstchen, M., Korotasz, A., 2025. *Leveraging multimodal LLMs for building condition assessment from street-view imagery*. University of Notre Dame. https://academicweb.nd.edu/~cwang11/papers/isvc25-bca.pdf.
Zhang, G., Wu, Q., He, B.-J., 2021. Variation of rooftop thermal environment with roof typology: a field experiment in Kitakyushu, Japan. Environ. Sci. Pollut. Res. 28 (22), 28415–28427. https://doi.org/10.1007/s11356-021-12799-9.
Zhang, G., Zhu, A-X., 2018. The representativeness and spatial bias of volunteered geographic information: a review. Ann. GIS 24 (3), 151–162. https://doi.org/10.1080/19475683.2018.1501607.
Zhang, L., Yang, Q., Agrawal, A., 2024. Assessing and learning alignment of unimodal vision and language models. *arXiv preprint* arXiv:2412.04616.
Zhang, Y., Wei, C., He, Z., Yu, W., 2024b. GeoGPT: an assistant for understanding and processing geospatial tasks. Int. J. Appl. Earth Obs. Geoinf. 131, 103976. https://doi.org/10.1016/j.jag.2024.103976.
Zhang, Y., Zhao, H., Long, Y., 2025. CMAB: a multi-attribute building dataset of China. Sci. Data 12 (1). https://doi.org/10.1038/s41597-025-04730-5.
Zhao, W., Persello, C., Stein, A., 2021. Building outline delineation: from aerial images to polygons with an improved end-to-end learning framework. ISPRS J. Photogramm. Remote Sens. 175, 119–131. https://doi.org/10.1016/j.isprsjprs.2021.02.014.
Zhou, W., Liu, J., Peng, D., Guan, H., Shao, Z., 2024. MtSCCD: land-use scene classification and change-detection dataset for deep learning. Natl. Remote Sens. Bull. 28 (2), 321–333. https://doi.org/10.11834/jrs.20243210.
Zhou, W., Persello, C., Li, M., Stein, A., 2023. Building use and mixed-use classification with a transformer-based network fusing satellite images and geospatial textual information. Remote Sens. Environ. 297, 113767. https://doi.org/10.1016/j.rse.2023.113767.

Zhu, J., Dang, P., Cao, Y., Lai, J., Guo, Y., Wang, P., Li, W., 2024. A flood knowledge-constrained large language model interactable with GIS: enhancing public risk perception of floods. Int. J. Geogr. Inf. Sci. 38 (4), 1–23. https://doi.org/10.1080/13658816.2024.2306167.

Zhu, X.X., Chen, S., Zhang, F., Shi, Y., Wang, Y., 2025. *GLOBALBUILDINGATLAS: an open global and complete dataset of building polygons, heights and LoD1 3D models*. ArXiv.org. https://arxiv.org/abs/2506.04106.

Zou, S., Wang, L., 2022. Mapping individual abandoned houses across cities by integrating VHR remote sensing and street view imagery. Int. J. Appl. Earth Obs. Geoinf. 113, 103018. https://doi.org/10.1016/j.jag.2022.103018.