Vision-language model (VLM)-enabled street view analytics: A systematic literature review

Ziyu Peng, Weisheng Lu*, Hongda An, Xianhua Xia, Yi Zhang, Fan Xue, and Junjie Chen
Department of Real Estate and Construction, The University of Hong Kong, Hong Kong

Abstract

Purpose. Street view analytics (SVA) is an emerging field focusing on the systematic analysis of street-level imagery to understand urban environments, which has rapidly advanced with the advent of vision language models (VLMs). Despite the significant advancements, a critical review of the applications of VLMs for SVA is lacking. This paper aims to fill this gap by providing a comprehensive literature review on VLM-enabled SVA.

Design/methodology/approach. This study adopts a Preferred Reporting Items for Systematic Reviews and Meta-Analyses-guided systematic literature review. After keyword retrieval, literature collection, thematic screening and a five-domain quality assessment (data representativeness, ground truth validity, model design and/or analytic rigor, validation and/or generalization and reporting and/or reproducibility), 69 VLM-enabled SVA studies (2020–2025) were selected. Five reviewers independently extracted and synthesized evidence, and inter-rater reliability was quantified to verify consistency.

Findings. The systematic analysis underscores the transformative potential of VLMs in SVA, emphasizing their multimodal data handling and open-domain knowledge integration. However, key challenges, while rooted in broader SVA limitations, manifest distinctly in VLM contexts: temporal dynamics, contextual reliance, annotation inconsistencies, computational demands and process transparency. Handling remains task-dependent, with future research focusing on city- and year-held-out temporal evaluation, robustness to street-level variability, retrieval-augmented generation for consistency, hybrid edge-cloud models and chain-of-thought prompting.

Originality/value. This study contributes to the field by synthesizing the latest development of VLMs for SVA, identifying avenues for future research and ultimately proposing an integrated workflow for enhancing VLMs' applications in SVA tasks.

1. Introduction

Street view analytics (SVA) refers to the systematic analysis of street-level imagery to recognize patterns in the physical environment, human behavior, or infrastructure (Biljecki & Ito, 2021). It has been widely applied in urban planning (Chen et al., 2022), transportation and mobility (Zhou et al., 2024), public health and safety (Keralis et al., 2020), social and behavioral research (Koo et al., 2022), and many others. The rapid growth of SVA is largely attributed to two phenomenal developments in recent years. Firstly, it is the rapid growth of the available street view imageries (e.g., Google Street View [GSV]), together with metadata of GIS, sensors, and other sources. Secondly, it is the powerful computing infrastructure and intelligent algorithms to mine the urban big data. The available methods continue to evolve with incremental advancements until the recent disruption of Vision Language Models (VLMs).

VLMs are defined here as an ecosystem covering artificial intelligence (AI) models and systems for vision–language analytics (Zhou et al., 2022). Built largely on transformers and trained on large-scale image–text pairs, this scope spans core VLMs that take visual inputs (e.g., CLIP) and multimodal large language models (LLMs) that couple a visual encoder with a generative model (e.g., BLIP-2, LLaVA, GPT-4V/4o). While traditional LLMs are text-centric, they are included in this review as non-visual components supporting VLM workflows. There are multiple studies to harness the power of these VLMs for SVA. To confirm the research gap, a pilot search was conducted, as described in Section 3.1, across Scopus, Web of Science, and other web sources. This yielded no dedicated literature reviews focused specifically on VLM-enabled SVA, though broader reviews on VLMs in urban studies (e.g., Liu et al., 2025) mention related multimodal tasks without in-depth synthesis for SVA applications. This absence highlights the need for a targeted review. This study serves as a pivotal reference for future research by synthesizing the state-of-the-art applications of VLMs in SVA, highlighting persistent challenges, and proposing an integrated framework that guides the field in advancing multimodal SVA.

The aim of this research is to provide a comprehensive literature review on VLM-enabled SVA. Beyond consolidation, this study advances SVA in three respects: (1) a model-centric synthesis that organizes dispersed work around linkages among research domains, data sources, models, applications, and limitations; (2) an integrated framework that aligns data sources, model training, prompting, and evaluation into a coherent workflow; and (3) a concise delineation of VLMs' advantages for SVA, current challenges, and future directions.

2. Conceptual background and SVA research field

SVA is defined as the process of acquiring information and analyzing patterns about urban areas using street view imagery (SVI) (Biljecki & Ito, 2021; He & Li, 2021; Li et al., 2022). SVI is typically captured through panoramic photographs taken at regular intervals along roadways (Rzotkiewicz et al., 2018). These datasets provide a rich, ground-level visual information that reveals physical, social, and environmental characteristics of urban areas.

- **Research domains**: Automated Annotation; Metadata Identification; Multi-source Data Fusion; Instance Segmentation; Semantic Segmentation; Object Detection; Scene Understanding; Change Detection
- **Data sources**: Google Street View; Mapillary; Tencent Maps; Open Street Map; Baidu Street View; Open-access Datasets; Exclusive Datasets; Not Specified
- **Models**: GPT Series; BERT; LLaMA; CLIP; BLIP; Other
- **Applications**: Urban Morphology; Environmental Perception; Socioeconomic Analysis; Urban Planning; Pedestrian Behavior Analysis; Environmental Monitoring; Transportation Management;
- **Limitations**: Data Coverage, Quality & Bias; Model Performance & Generalization; Prompt & Textual Constraints; Geographic & Cultural Constraints; Temporal & Real-time Aspects; Methodological & Validation Gaps;

The above five key aspects of SVA serve as a guide (a) when analyzing the selected publications; and (b) when reporting the data analysis results and findings. The process was developed based on the process outlined in Section 3.3. Their interdependence is evident: domains dictate the tasks that require specific data sources, which in turn influence the choice of models, leading to diverse applications. Limitations highlight areas needing refinement across the process. This study will investigate how the involvement of VLMs facilitates and transforms the SVA studies considering these aspects.

Under each aspect, sub-areas are identified for comprehensive reviews. Research domain focuses on concrete tasks that VLMs take to transform SVA. Automated Annotation accelerates the foundational labeling process. Scene Understanding integrates these insights for broader context, while Change Detection tracks temporal shifts, relying on fused data from multiple sources. Data Sources, including GSV, Mapillary, etc., supply the raw material for SVA tasks. Then, the Models aspect included five common model series. These models will then lead to seven typical application scenarios. Lastly, limitations, as divided into the six dimensions, address barriers to SVA's success.

3. Methodology

This research adopts the systematic literature review (SLR) methodology. Over the years, SLR has been widely employed in the fields of SVA (Biljecki & Ito, 2021; He & Li, 2021; Li et al., 2022), indicating a mature and sufficient body of studies and data sources.

3.1 Collection and filtering

At the outset of research, a pilot study was conducted to gain an initial understanding of the body of literature connected to VLM-enabled SVA. The main focus here was on how VLM transforms and augments the traditional SVA tasks. The time span covered January 1, 2020, to May 1, 2025. To start, keywords were examined used in previous systematic literature reviews (Bardhan et al., 2024; Dai et al., 2024; Hou et al., 2024a). The keyword selection was intentionally broad to encompass the overall research area, yet specific enough to ensure the relevance of the literature found. To achieve this, the study adopted the comparable systematic literature reviews (Haddaway et al., 2015). The study initially applied this set of keywords to search Google Scholar, and after analyzing the first 100 results, the study refined the keyword set, as presented in Section A.3, Appendix A.

This study then identified several databases for gathering academic literature. The study includes literature from three databases: Scopus, Google Scholar, and Citation Referring. The research then developed search strings using a final set of keywords linked with Boolean operators (OR and AND). The strategy involved connecting synonyms within a single item using OR. Then, different items were then combined with AND. The completed search query can be found in Section A.4, Appendix A. To improve search accuracy and manage the review workload, the study focused on the title and abstract. The research established inclusion and exclusion criteria as follows:

Inclusion criteria:

1) The paper published between January 1st 2020 and May 1st 2025;
2) The paper states that it employs a VLM;
3) The paper indicates that the research includes an SVA task;
4) The paper's full text is publicly accessible.
5) The paper is classified as at least moderate quality within the quality assessment process (Section A.1, Appendix A)

Exclusion criteria:

1) Redundant papers or similar works by the same authors in different versions;
2) Tool demonstrations and editorial pieces;
3) Papers presented at workshops or doctoral symposia;
4) Grey literature, such as technical reports;
5) Literature not written in English;
6) Papers are classified as low quality within the quality assessment process (Section A.1, Appendix A)

3.2 Collection and extraction

In this stage, PRISMA framework presented in Figure 1 (a) is adopted to structure our review. This framework provides clear guidelines for conducting and documenting SLRs, improving methodological rigor and reporting transparency.

As it turns out, 69 relevant publications were selected. This sample size reflects the emerging nature of VLM-enabled SVA, which only gained traction post-2022 with the advent of accessible generative AI tools like ChatGPT (Nah et al., 2023). The full list of literature can be found in the Appendix B. Figure 1(b) summarizes basic statistics. As it shows, the emergence of VLM-enabled SVA as a research field came with the launch of ChatGPT in 2022 (Nah et al., 2023). The Sankey diagram (see Figure 1[c]) illustrates a flow across SVA aspects. Prominent domains include Scene Understanding and Multi-source Data Fusion evenly distributing high values (234 and 198 respectively). Data sources then connect uniformly to models, with GSV and open-access datasets being the largest flows to models. Models feed into applications with Other and GPT models showing the strongest contributions to areas like Urban Planning and Environmental Perception. Finally, each application branches evenly to all six limitations, with Urban Planning being the highest total flow. At last, most of researchers came from the United States, China, and Singapore as shown in Figure 1(d).

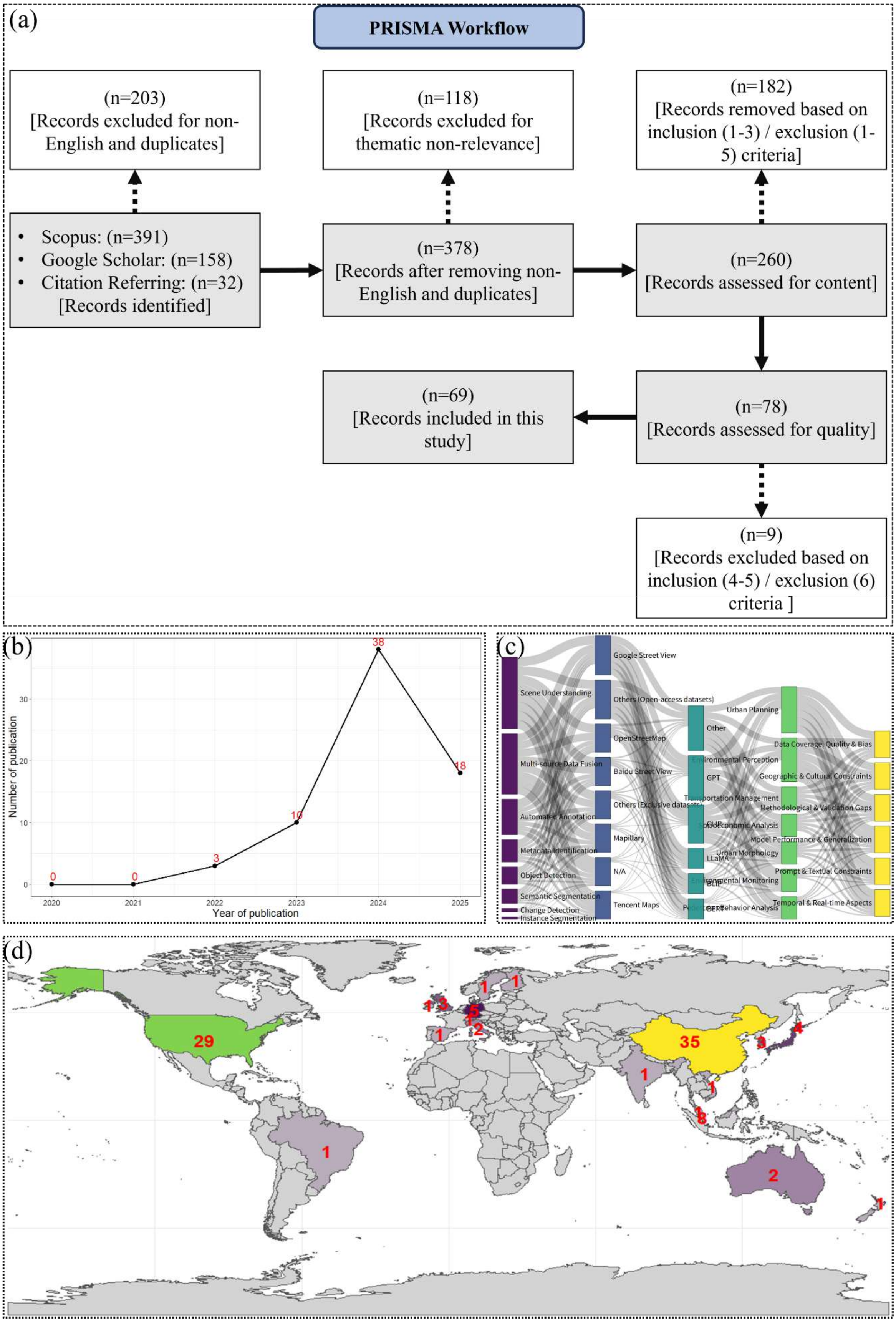
(a)
PRISMA Workflow
(n=203)
[Records excluded for non-English and duplicates]
(n=118)
[Records excluded for thematic non-relevance]
(n=182)
[Records removed based on inclusion (1-3) / exclusion (1-5) criteria]
• Scopus: (n=391)
• Google Scholar: (n=158)
• Citation Referring: (n=32)
[Records identified]
(n=378)
[Records after removing non-English and duplicates]
(n=260)
[Records assessed for content]
(n=69)
[Records included in this study]
(n=78)
[Records assessed for quality]
(n=9)
[Records excluded based on inclusion (4-5) / exclusion (6) criteria]
(b)
Number of publication
0
10
20
30
Year of publication
2020
2021
2022
2023
2024
2025
0
0
3
10
38
18
(c)
Scene Understanding
Multi-source Data Fusion
Automated Annotation
Metadata Identification
Object Detection
Semantic Segmentation
Change Detection
Instance Segmentation
Google Street View
Others (Open-access datasets)
OpenStreetMap
Baidu Street View
Others (Exclusive datasets)
Mapillary
N/A
Tencent Maps
Other
GPT
CLIP
LLaMA
BLIP
BERT
Urban Planning
Environmental Perception
Transportation Management
Socioeconomic Analysis
Urban Morphology
Environmental Monitoring
Pedestrian Behavior Analysis
Data Coverage, Quality & Bias
Geographic & Cultural Constraints
Methodological & Validation Gaps
Model Performance & Generalization
Prompt & Textual Constraints
Temporal & Real-time Aspects
(d)
29
35
1
3
1
5
1
1
1
2
1
3
4
1
1
3
1
2
1
1

Figure 1 Flow diagram for literature selection, publication years, Sankey diagram, and countries of selected literature

3.3 Synthesis and review

Then, the synthesis process for identifying sub-areas under each theme is conducted. This study uses Liu et al. (2023) as an example to demonstrate how sub-areas are determined (see Section A.5 in Appendix A for more details). One reviewer will read the selected literature thoroughly and summarize the paper under each of the five aspects by applying two-step thematic coding techniques: (1) first identifying and extracting key elements such as methodologies, data details, model specifics, practical uses, and challenges, and then (2) categorizing them to align with the predefined key aspects. This involved iterative coding to ensure concise summaries that captured the paper's core contributions. Then, the summaries of each selected paper will be further distilled into keywords. After comparing the commonalities given 69 studies, the final keywords of five aspects (see Section 2) that can be representation key sub-areas under each aspect will be synthesized.

Based on the SVA key aspects and their contained sub-areas, five human reviewers were invited and independently examined the selected literature. The reviewed contents were organized using spreadsheet formats, where each literature was mapped against the aspects and sub-areas (Appendix B). The coding tasks will be fulfilled manually to identify the matching of themes as proposed in the analytic framework. Their outputs are evaluated using the Inter-Rater Reliability (IRR) procedure to quantify the degree of agreement. The details concerning evaluation process and results can be found in Section A.2, Appendix A.

4. Review results and findings

This section reports the domains, data sources, specific VLMs, applications, and limitations associated with the application of VLMs in SVA research.

4.1 SVA domains

The applications of VLMs in specific SVA domains have become evident across multiple domains, with research spanning technical foundations, methodological innovations, and domain-specific problem-solving. The average number of references in each domain category across the five reviewers is presented by the Figure B1 in the Appendix B.

Scene understanding is the most widely studied in VLM-enabled SVA. This field includes interpreting and analyzing visual data to understand the background of the urban environment. For example, Liu et al. (2023) used a multimodal learning model to evaluate walkability. Their work proved how LLM could interpret street view images to evaluate the quality of urban design that affects pedestrian experience. The CLIP model in this paper enhanced SVA by using its zero-shot learning and contrastive learning capabilities to assess perceived walkability from SVI without requiring extensive labeled datasets. In addition, Chen et al. (2024) developed a semi-supervised prediction VLM: SPHINX-V2, combining street view images with text data to comprehend urban streets.

Multi-source data fusion is the second most studied field, which integrates different data types to understand the urban environment. For example, Ramalingam & Kumar (2025) developed a hybrid model, which combined the text and elevation features in SVI to predict the use of buildings. Specifically, they adopted the GPT-3.5 Turbo LLM to interpret textual information from building signage, which improved building usage classification. Zhang et al. (2024) combined street view images with spatial environment to evaluate the vitality of child-friendly cities, and showed how data fusion can provide information for urban policy and design.

Automated annotation, as the third important field, focuses on automatic labeling of visual elements in SVI. This process identifies objects, scenes, and their attributes so as to simplify large-scale urban analysis. Bleč et al. (2024) annotated the walkability-related features of street view images through multimodal VLM. It did so by providing both quantitative walkability scores and qualitative linguistic explanations. Ouyang et al. (2024a) used Health CLIP to predict the depression rate by extracting health-related features from satellite and street view images. They used GPT-4 model to generate health-related captions for SVI, allowing the fine-tuned CLIP model to extract depression-related environmental features.

4.2 Vision language models

Diverse models have been applied in the literature to address VLM-enabled SVA tasks. A large share of studies adopts discriminative VLMs (e.g., CLIP/BLIP) that align images and text, while a growing subset employ multimodal LLMs (e.g., GPT-4V/LLaVA) that integrate visual inputs into a generative language model. These models are typically pre-trained on massive datasets of image-text pairs (e.g., billions for foundation models like CLIP) and then adapted for SVA via prompt engineering for zero- or few-shot inference or fine-tuning on labeled datasets. These labeled datasets often involve thousands of samples for domain adaptation. Among the 69 reviewed studies, GPT-series emerges as the most frequently employed model, featured in over 30 articles, followed by CLIP, LLaMA, BLIP, and BERT. The average number of references in each model category is presented by the Figure B2 in the Appendix B. As summarized in Table 1, GPT-4V offers superior reasoning and explanations but can be cost-inefficient at city scale; lighter GPT-3.5 + vision pipelines reduce cost yet are weaker on complex spatial reasoning, whereas CLIP-family models remain cost-efficient for retrieval/tagging but are ill-suited to free-form explanation.

Table 1 SVA Applications, strengths, limitations of VLMs

Family	SVA Applications	Strengths	Limitations
GPT-4V/o	Urban Planning; Socioeconomic Analysis; Transportation Management; Environmental Perception	Best reasoning and explanation; strong zero/few-shot; flexible prompting	High cost/latency at city scale; less effective for strict geometry or precise localization

GPT-3.5 + vision encoders	Environmental Monitoring; Socioeconomic Analysis; Urban Morphology	Lower cost; scalable	Weaker on complex spatial reasoning; sensitive to prompt design
LLaVA / BLIP-2	Environmental Perception; Urban Planning; Transportation Management; Pedestrian Behavior Analysis	On-prem, tunable; moderate cost	Falls short of GPT-4V on long-range reasoning; needs curated multimodal tuning
CLIP / ALIGN	Urban Morphology; Environmental Perception; Socioeconomic Analysis; Environmental Monitoring; Transportation Management	Fast; cost-efficient; excellent zero-shot matching and retrieval	Limited for chained reasoning and free-form textual explanation; non-generative outputs
BLIP	Environmental Monitoring; Urban Morphology; Socioeconomic Analysis	Good weak labels; speeds downstream tasks	Fine semantics may be missed
BERT / LLaMA	Socioeconomic Analysis; Urban Planning	Very low cost; strong for structured text	No direct visual reasoning

4.3 Data sources

Data sources for VLM-enabled SVA tasks are characterized by reliance on GSV, which is adopted in 34 out of 69 studies. This is notable but lower than the baseline in general SVA literature, where GSV is used in 87% of studies (Biljecki & Ito, 2021), reflecting the emerging field's greater use of diverse or regional sources. Regional alternatives such as Baidu Street View and Tencent Maps collectively accounted for about 10 studies, with a strong emphasis on China-centric research.

A significant trend in VLM-driven SVA is the integration of open-access datasets with street view platforms to address domain-specific analytical challenges. Studies have combined Baidu Street View with OpenStreetMap to enhance spatial topology modeling, as seen in Chen et al. (2024) and del Castillo et al. (2023), where road networks and POI data enriched urban walkability predictions. Similarly, 7 studies fused SVI with satellite data for cross-view analysis, such as Zhang et al. (2025), which predicted urban inequality indicators using satellite and SVI with fine-tuned chain-of-thought-enabled VLM. The reliance on exclusive datasets was evident in 8 studies, including López-Otero et al. (2024), which merged GSV with online housing price data, and Wang et al. (2024), which combined Baidu Street View with corporate registries and patent databases.

To use these data sources to support VLM tasks, a special preprocessing process is needed to map SVI to specific SVA applications. The typical process begins with sampling and acquisition through API, then stitching to create panoramic view and distortion correction, and finally normalizing to deal with illumination or perspective changes (Hou et al., 2024b). Annotation protocol plays a key role in environmental monitoring or target detection in scene understanding, which involves a structured codebook used to mark elements such as road damage (Ren et al., 2024). These protocols enhance VLM fine tuning by providing consistent ground truth. In the case of limited coverage of real SVI, synthetic data generation can expand the dataset through modeling or generative AI, and map it to tasks such as object detection (Turkcan et al., 2024).

4.4 Applications

Integrating VLMs into SVA has been able to achieve a series of diverse tasks, which reflect the adaptability of VLMs in interpreting multimodal urban data, integrating text and visual information, and generating operable aids for geospatial and socio-economic applications. the average number of identified citations for each application domain by the Figure B4 in the Appendix B.

Urban Planning stands out to be the most common application of VLMs in SVA, it leverages VLM to evaluate and predict urban features and functions. Cheng et al. (2024) fed SVI into an LLM to evaluate urban safety, taking risk factors like poor lighting or narrow sidewalks into account when giving out a safety score. Feng et al. (2024) introduced a framework called CityBench, together with various VLMs including GPT-4o, LLaVA-NeXT, and Qwen-VL-plus to complete 8 specific urban tasks related to perception and decision-making. These models are applied to scenarios such as image geolocalization and outdoor visual-language navigation. Shihab et al. (2024) focused on sidewalk detection to ensure accessibility in urban layouts during the planning process.

Environmental perception and monitoring are the second and third most studied application fields. These applications include evaluating the subjective quality of urban space and tracking their changes. VLMs stood out in this field, simulating and quantifying the perception usually collected during human on-site experience by explaining the visual clues in SVI. Lyu et al. (2024) developed a vision-to-language framework for training-free visual place recognition, enhancing localization through coarse-to-fine spatial reasoning. These prove VLMs' ability to connect quantitative image analysis with qualitative human perception.

Lastly, VLMs in SVA have been proven to be accurate in analyzing human behavior and socio-economic status. Socio-economic analysis uses VLMs to infer economic and social indicators from SVI. López-Otero et al. (2024) analyzed the qualitative information and street view data in Wikipedia to study immigration and isolated space. VLMs were also able to simulate human perceptions and behaviors. Verma et al. (2023) explored the use of generative agents powered by VLMs to collect urban perceptions and model pedestrian experiences in street environments.

4.5 Limitations

These studies also reported the limitations of using VLMs for SVA, which are rooted in data coverage, quality, and bias, as well as geographic and cultural constraints. The average number of references is presented by the Figure B5, Appendix B.

The most frequently mentioned limitation comes from the coverage, quality, and bias of data. Studies have reported that the geographical scope of training and evaluation datasets is narrow. This spatial bias not only limits the generalizability of underrepresented areas but also makes the unfairness of the algorithm exist for a long time. For instance, Wu and Huang (2024) revealed that pure English text encoder degraded geo-localization accuracy in non-English regions. These issues are compounded by data quality dependencies: adverse weather conditions (Yao et al., 2024; Li et al., 2024) and low-resolution SVI (Cepeda et al., 2023; Pan et al., 2024).

Another significant limiting factor is the technical limitation of model performance. VLMs are very sensitive to the prompts, which means even a small change in words may lead to inconsistent scores. For example, zero-shot urban function inference (Huang et al., 2024) relied heavily on manual prompting design and required domain expertise to make culture-related descriptors. This may lead to limited scalability in urban tasks.

Lastly, methodological and validation gaps emerged in about 50% of studies, where synthetic data (Feng et al., 2024) replaced ground-truth metrics. Roberts et al. (2024)'s geographic assessments correlated weakly with precise localization tasks due to training data biases and limited API access for robust validation. The open-vocabulary classification algorithms developed by de Moraes et al., 2024 illustrated an inconsistent performance in pathway and surface material detection due to hallucinations in CLIP-based model outputs. To address these issues, integrating participatory sensing, e.g., community surveys in Zhang et al. (2024), and hybrid AI-human workflows (Malekzadeh et al. 2025), is essential.

5. The integrated framework for applying VLMs in SVA

The integrated framework is derived through a synthesis of key insights from the 69 reviewed studies, focusing on those that propose enhancements to VLM applications in SVA rather than mere usage. Specifically, it is anchored in three foundational works: IM2CITY (Wu & Huang, 2022) for multi-modal geo-localization at the city level; PIGEON (Haas et al., 2024) for precise coordinate prediction and bias mitigation; and VELMA (Schumann et al., 2024) for language-guided navigation. These were selected based on their emphasis on improving VLM robustness, as identified in our coding process (Section 3.3). Insights from other studies were incorporated to address prevalent limitations (Section 4.5), such as temporal dynamics (e.g., Yin et al., 2025) and prompt constraints (e.g., Huang et al., 2024). This bottom-up synthesis is implemented to ensure the framework is grounded in existing literature while extending it to diverse SVA tasks based on VLMs. However, since they were proposed to fit into specific SVA scenarios, the framework is enhanced based on other reviewed works to cope with diverse SVA tasks and challenges.

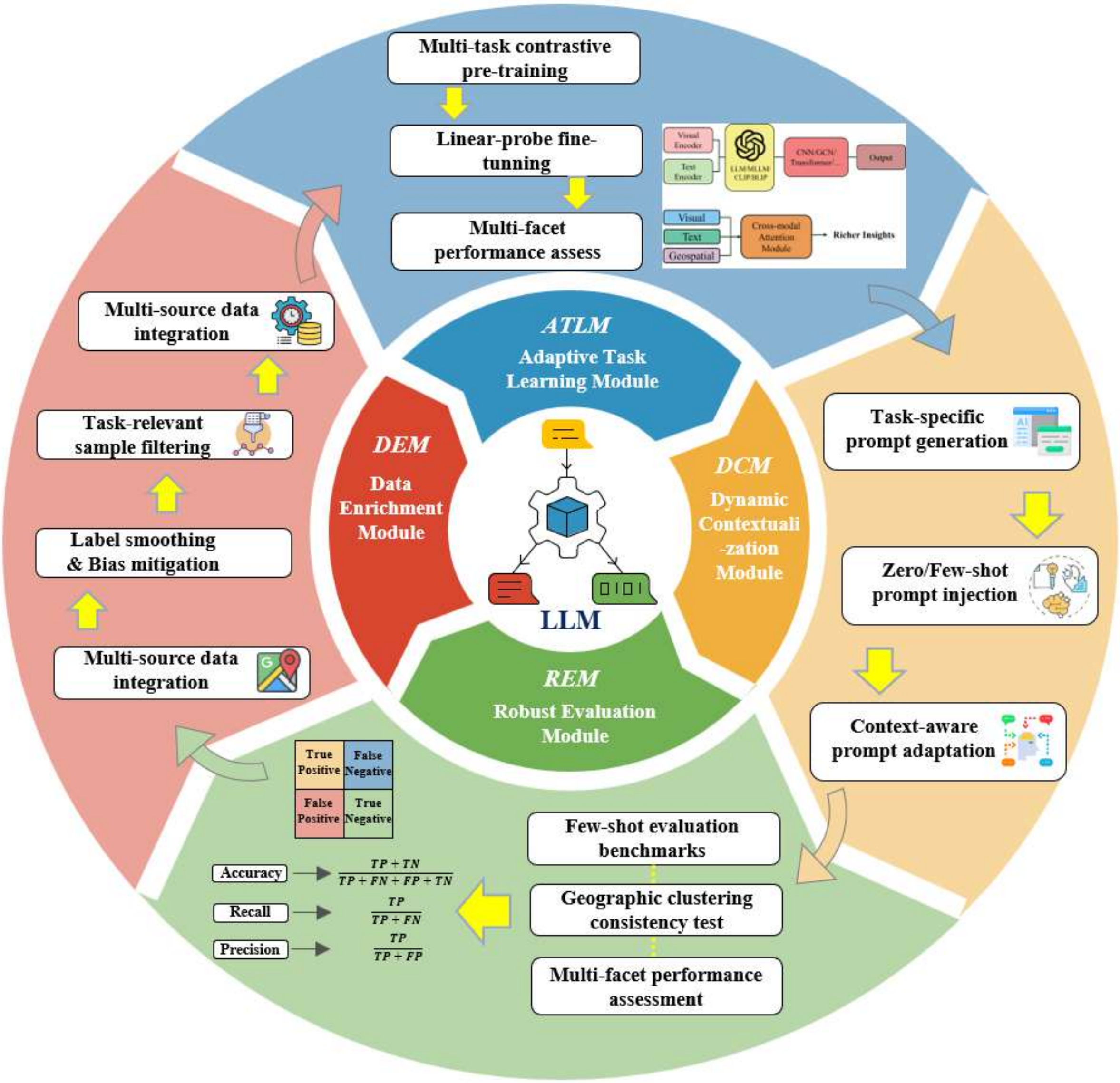
Multi-task contrastive pre-training
Linear-probe fine-tunning
Multi-facet performance assess
Visual Encoder
Text Encoder
LLM/MLLM/ CLIP/BLIP
CNN/GCN/ Transformer/...
Output
Visual
Text
Geospatial
Cross-modal Attention Module
Richer Insights
ATLM
Adaptive Task Learning Module
DEM
Data Enrichment Module
DCM
Dynamic Contextualization Module
REM
Robust Evaluation Module
LLM
Multi-source data integration
Task-relevant sample filtering
Label smoothing & Bias mitigation
Multi-source data integration
Task-specific prompt generation
Zero/Few-shot prompt injection
Context-aware prompt adaptation
Few-shot evaluation benchmarks
Geographic clustering consistency test
Multi-facet performance assessment
True Positive
False Negative
False Positive
True Negative
Accuracy
Recall
Precision
$\frac{TP+TN}{TP+FN+FP+TN}$
$\frac{TP}{TP+FN}$
$\frac{TP}{TP+FP}$

Figure 2 The four-module integrated framework for VLMs in SVA

As illustrated in Figure 2, the foundation of this framework lies in Data Enrichment Module (DEM). This foundational unit is implemented to ensure that VLM in SVA can handle diverse and representative datasets that will not suffer from the issues of data coverage, quality, bias, and temporal changes. DEM draws lessons from PIGEON’s semantic geographic unit creation method that uses open data sources such as SVI and satellite images to construct rich data units. DEM will center around the label smoothing technique to reduce the deviation by balancing geographical and cultural representation, and ensures that the model will not be biased towards over-represented regions or demographic data (Haas et al., 2023). Lastly, in order to cope with the dynamic changes of time, DEM introduces time-stamped data, which analyzes historical trends and real-time conditions in parallel (Yin et al., 2025).

Then, the Adaptive Task Learning Module (ATLM) will be integrated to handle all kinds of SVA tasks through a flexible learning process, so as to meet the challenges of model performance and generalization. ATLM is inspired by the Generalized Embedding Model (GEM) in IM2CITY, which adds a lightweight linear layer to the pre-trained CLIP model for fine-tuning VLM in response to specific tasks. This will make ATLM quickly adapt to new tasks with small samples.

Then, Dynamic Contextualization Module (DCM) will be applied to handle prompts and text constraints by tailoring context-aware prompts for each SVA task. DCM designs prompts containing specific task requirements to make VLM's output relevant and operable. DCM also equips with the zero sample GEM that enables the framework to perform certain SVA tasks without a lot of retraining by using pre-training knowledge. The module links to ATLM by providing tips consistent with its fine-tuning model, and enters the final module by providing the text framework needed to evaluate the actual performance.

Finally, Robust Evaluation Module (REM) is designed to ensures the framework's reliability and generalizability across diverse tasks. The REM adopts the small sample learning evaluation method. This will allow the proposed framework to maintain a good performance under in low data scenarios. Also, REM has the location clustering and retrieval functions suggested by PIGEON to group similar urban environments. At last, REM adopts a multi-faceted evaluation approach to ensure robustness. It will connect back to DEM by providing improved feedback of iterative data, evaluates the accuracy of ATLM, and tests the relevance of DCM prompts.

Table 2 illustrates this mapping, showing how each module systematically addresses identified challenges and draws from specific sources. Nonetheless, it is acknowledged that this framework remains conceptual. Validation is needed given real-world scenarios.

Table 2 Principled mapping of framework modules to challenges and literature

Module	Addressed challenges	Key derivations from literature
Data Enrichment Module (DEM)	Data coverage, quality & bias; Temporal & real-time aspects	PIGEON (Haas et al., 2024) for semantic units and bias mitigation; Yin et al. (2025) for time-stamped data
Adaptive Task Learning Module (ATLM)	Model performance & generalization; Prompt & textual constraints	IM2CITY (Wu & Huang, 2022) for linear-probing GEM; VELMA (Schumann et al., 2024) for few-shot learning; PIGEON for multi-task pre-training
Dynamic Contextualization	Geographic & cultural constraints; Methodological	IM2CITY for zero-shot GEM; Huang et al. (2024) for prompt engineering

Module (DCM)	& validation gaps	
Robust Evaluation Module (REM)	Model performance & generalization; Methodological & validation gaps	VELMA for small-sample evaluation; PIGEON for clustering; Feng et al. (2024) for benchmarking

6. Discussion

6.1 VLMs as a strong force for SVA

Our review showed that VLM-based approached were strong alternatives to traditional computer vision methods in SVA. VLMs present three-fold significant advantages in SVA tasks. First and most intuitively, compared with traditional CV-based SVA pipelines that rely heavily on manual annotation and supervised training, VLM-enabled SVA is able to reduce labelling and training requirements for many tasks. This reduction is often achieved through pre-trained multimodal models and prompt-based inference that enable zero-shot or few-shot capabilities (Blečić et al., 2024; Ouyang et al., 2024; Verma et al., 2023). However, manual annotation is not universally saved for all SVA-related tasks. Tasks including fine-grained attribute labelling, domain-specific feature extraction, and high-precision localization still require substantial labelled data and task-specific fine-tuning.

Secondly, the multimodal capabilities of VLMs were found particularly suited to the heterogeneous demands of SVA. In SVA, understanding urban environments often requires interpreting visual data together with contextual information. For instance, Yang et al. (2024) developed V-IRL which integrates visual data from street view imagery and contextual information from geospatial APIs to enable AI agents to navigate and perform tasks in real-world urban environments. Chen et al. (2024) used LLMs to combine visual and textual representations for predicting urban street functions and socioeconomic indicators. Their abilities to handle unstructured, multimodal data and generate human-like interpretations not only enhance the depth of analysis but also scale to large datasets.

The last significant advantage lies in the open-domain knowledge capabilities of VLMs, which focus on vast, cross-disciplinary information. This adaptability is useful in integrating street view images into the complex urban framework. For example, in the pedestrian detection model developed by Park et al. (2024), multimodal VLMs inferred appearance elements from diverse textual descriptions, enhancing visual cue integration for improved detection performance. This work showed that the integration of VLMs and SVA made it possible to adjust the extensive knowledge of VLMs to specific SVA requirements and provide richer explanations than narrow tools in specific fields.

6.2 Challenges in the status quo

Although the branch of VLMs-enabled SVA is becoming a strong competitor in this field, it also has certain limitations and needs further research.

(1) Temporal sensitivity

It is of high importance in SVA to capture the time evolution of urban environment for tasks such as change detection, urban planning and infrastructure monitoring. While SVIs provide rich ground-level images, the reliance on static snapshots in literature can hinder temporal generalization. In VLM-enabled SVA, explicit handling of temporal variation is still highly task-dependent. Some works begin to address time-stamped data, change detection, or time-aware evaluation, but comprehensive temporal validation remains limited. As demonstrated by Gan et al. (2024), who found that cross-domain differences between template and real traffic signs reduce recognition accuracy under varying conditions, such as physical wear.

(2) Contextual dependence

As a general challenge in SVA and CV, the excessive dependence on context patterns derived from training data, which proved to be particularly problematic in view of the dynamic nature of street view images. Due to weather, lighting and seasonal changes, the urban scenes captured in street view data set show significant variability, and all these factors will significantly change the visual clues. The instability of this background weakens VLM's ability. Moreover, explicit handling of contextual variation tends to be task-dependent, such as weather-related metadata or time-aware evaluation to simulate varied conditions. For example, Zeng et al. (2025) noted that the DVBench framework has a strong reliance on safety-critical driving videos. This contextual dependence may harm the consistency of performance in autonomous driving.

(3) Annotation inconsistency

The inconsistency of annotations brings challenges to VLMs of SVA, especially when automatically labeling elements in SVI. In a complex urban environment, subtle changes in visual cues, such as changes in lighting conditions or camera angles, usually lead to unstable or unclear labels on the same features in different images by VLM. This undermines the reliability of data sets that are crucial for SVA applications. For example, in the task of scene understanding, VLM may correctly identify the cracked sidewalk as 'poor quality' in one image, but ignore the same defect in another image due to shadow interference, thus undermining the efforts to draw a pedestrian safety map consistently.

(4) Computational demands

The literature illustrated a clear trend towards multi-modalities, integrating text with images, videos, etc. While this direction opens up a new way for understanding and interpreting the work, it requires a high amount of computation. This is because each data type has its own unique structure and processing requirements, such as word sequence of text data, color-valued pixel grid of image data and so on. In order to analyze multiple data types at the same time, the algorithm has to go beyond single processing, cross-modal integration, and correlation information. Because of that, both Hao et al. (2024) and Ramalingam & Kumar (2025) have emphasized the computational intensity required to train VLMs on large-scale street view datasets. These computational requirements may limit the accessibility of multimodal research.

(5) Process transparency

Literature has showed that the opacity in VLMs has become a key limitation of SVA. Cheng et al. (2024) and Blečić et al. (2024) demonstrated that VLM can generate useful output from SVI, such as safety or walkability scores, but the reasoning behind these predictions was hidden, making the results unconvincing to the public or policymakers. This opacity stems from the complex levels and parameters of the model, and the way they transform input into output is difficult to explain directly. This problem is more obvious in multimodal VLM which integrates visual and text data.

6.3 Navigation for future research

In face of the challenges discussed above, this section presents potential research avenues for improving VLMs in SVA. The details are discussed as follows.

(1) City- and year-held-out temporal evaluation and supervision

Future work should operationalize temporal dynamics in street view analytics by enforcing city- and year-held-out splits, with explicit leakage checks and a geographic split map. Beyond static snapshots, VLMs can be weakly supervised with time-aligned urban event logs: roadwork permits, completion records, sign-replacement logs, and maintenance reports (Badi et al., 2017). Consecutive panoramas from the same location provide contrastive pairs for temporal pretraining, while prompts conditioned on events tie changes to known interventions. Evaluation should include object-level and segment-level change detection, temporal calibration, and degradation under increasing gaps. A minimal protocol benchmarks the same model on the same-year, adjacent-year, and cross-year settings across more than two cities, with confidence intervals and error attribution to urban processes.

(2) Robustness to weather, lighting, and seasonal variability in street scenes

To reduce context over-reliance within specific street views, training should couple physics-guided augmentation with measured metadata. Synthesize weather/illumination via render-based relighting and generative translation that preserve geometry and signage legibility, tag each augmented image with physical parameters (rain rate, sun elevation, ground wetness). VLMs should ingest meteorological observations as auxiliary inputs and be regularized for invariance across matched locations and conditions. Pose normalization via structure from motion stabilizes viewpoint changes for façade-level reasoning. Benchmarks should report per-condition slices (day/night; rain/fog; summer/winter) and uncertainty intervals. Ablations must separate gains from augmentation, metadata conditioning, and prompt design to reveal robustness sources.

(3) Maintain annotation inconsistency based on Retrieval-Augmented Generation (RAG)

RAG integrates a retriever (accessing the knowledge base of related examples) with a generator (using this information to generate consistent and accurate annotations). In the context of SVA, the retriever can extract high-quality annotations from a variety of SVI databases, covering various camera angles, surrounding environments, etc. When labeling new images, the model can retrieve similar images or features that have been correctly labeled in the past, thus providing key background information to guide their decision-making. For example, if the shadow blocks the sidewalk, the model can retrieve other images with similar shadows, but the sidewalk is still accurately marked as 'poor quality'. This retrieval step reduces the influence of visual changes and makes the model consistent between different images with the same feature.

(4) Incorporating a hybrid edge-cloud model

To meet the computational challenge of multimodal SVA, future research can use edge computing to optimize processing efficiency. Edge computing place the data processing closer to its sources, such as camera in urban built environment or private data sources. This measure will reduce the dependence on centralized cloud servers (Shi et al., 2016). Not only that, by deploying lightweight algorithms on edge devices, initial data preprocessing and analysis can be done locally, and only necessary insights can be sent to the central server. This reduces the computational burden and network pressure of training large-scale models. In this case, a hybrid edge-cloud model, in which edge devices handle real-time processing and cloud manages complex computation, provides a balanced solution.

(5) Using chain of thought to break down the decision-making

The last future research direction is adopting the method of chain of thought (CoT) (Li et al., 2024) during the application of VLM in SVA. It does so by decomposing complex decisions into clear sequential steps. Additionally, by fine-tuning on the dataset provided by human experts with final evaluation and detailed reasoning, VLMs can be trained to generate natural language explanations while outputting. Finally, for VLM that processes both visual and text data, visual attention mechanism (Guo et al., 2022) can reveal which specific image regions or text segments have the greatest influence on prediction, thus establishing a practical connection between input and output.

6.4 Comparative positioning relative to prior similar thematic SLRs

To situate the contribution, this review is contrasted with two similar thematic SLR studies (see Table 3). Biljecki & Ito (2021) provide a comprehensive survey of SVI in urban analytics, screening 619 papers and classifying 250 studies across ten application domains, thereby documenting Google Street View's dominance and the breadth of SVI applications. Li et al. (2022) focused on the architectural environment, put forward an innovative decision-making framework to organize the adoption and implementation of SVI, and pointed out the key success factors. Together, these works establish SVI's value and scope while emphasizing applications and adoption workflow, thereby motivating a focused synthesis on language–vision methods now emerging in SVA.

Table 3 Comparative synthesis of this study and previous reviews

Item	**This review**	**Biljecki & Ito (2021)**	**Li et al. (2022)**
Primary scope	An AI model-centric SLR on VLM enabled SVA, covering data sources, model training, prompting, and evaluation protocols specific to language–vision pipelines.	A field-wide survey of SVI in urban analytics, classifying applications across ten domains and documenting platform usage.	A systematic review of SVI adoption in the built environment, structured by an innovative decision-making framework to surface enablers, barriers, and implementation guidance.
Corpus & Time span	69 papers; 2022-2025	250 papers; 2018-2020;	263 papers; 2007-2022
Methods	Conventional SLR;	Conventional SLR;	Conventional SLR;
Contributions	1. Systematic synthesis of VLM-enabled SVA 2. Integrated framework for VLM application in SVA 3. Advantages, challenges, and future navigations of VLMs in SVA	1. Comprehensive application mapping across 10 domains 2. The quantification of provider dominance, data sources, and capture modes 3. Synthesis of methods and metrics, identification of deficiencies in geographic and temporal coverage and access	1. Innovation–decision framework 2. Synthesizes enablers/barriers, provider selection, data-integration workflows, and practitioner-oriented guidance for implementation

Building on that foundation, this study delivers three significant pieces of progress. Firstly, this paper systematically summarizes the VLM-enabled SVA literature, and integrates scattered papers into a model-aware perspective. Secondly, this review puts forward an integrated framework of applying VLM in SVA, which integrates data sources, model training, prompting, and evaluation protocols into a coherent workflow. Thirdly, this paper points out the advantages and current challenges of VLM in SVA, and defines the future development direction. Collectively, these contributions reposition the discussion from broad SVI application mapping toward a reproducible, AI model-centric agenda for VLM-enabled SVA.

7. Conclusion

This paper provides a comprehensive review of the application of VLMs in SVA for various value-added applications. It reveals that VLMs for SVA are still in its nascent stage, considering the relatively small but growing body of literature compared to their traditional SVA counterparts. Nevertheless, the review highlights the transformative impact of VLMs on SVA, noting their superior performance in typical SVA applications. Three main advantages of VLM-enabled SVA are identified: (a) minimal annotation, (b) multimodal capabilities, and (c) open-domain knowledge. However, integrating VLMs into SVA is not without challenges. Problems related to temporal dynamics, contextual reliance, annotation inconsistencies, computational demands, and process transparency are identified. To address these issues, the study proposed future research directions including city- and year-held-out temporal evaluation, physics-guided robustness to street-level variability, retrieval-augmented generation, edge-cloud computing, and chain of thought. However, this review has some limitations. First and most intuitively, it is constrained by the nascent state of the field, relying on 69 papers. Second, the conceptual nature of the proposed framework lacks empirical validation, potentially limiting its immediate applicability without further testing in diverse SVA scenarios. Future research could extend this review by conducting updates to capture rapid VLM advancements, and empirically evaluating the integrated framework through case studies to refine its practicality.

The potential of VLMs to completely change SVA is enormous. They may be strong competitors of traditional methods such as computer vision or segmentation. Future research should focus on improving the accuracy and efficiency of VLM through fine-tuning and retrieval enhancement generation, integrating different data sources to improve coverage and resolution, and developing privacy protection technologies to reduce network security risks. By overcoming these obstacles, VLM can fully realize its potential to support the development of smart cities and evidence-based decision-making and ultimately contribute to a more sustainable, livable, and flexible urban environment.

Practically, this review equips urban planners and policymakers with a consolidated resource on VLM applications, enabling more efficient integration of multimodal datasets and patterns arising from them into SVA for better urban environments, transportation planning, and public health. Furthermore, the proposed integrated framework provides a ready-to-adapt workflow for practitioners, facilitating the deployment of VLMs in real-time SVA tasks while mitigating common challenges like computational demands and data biases. Lastly, by identifying targeted future directions, the study empowers AI developers and researchers to innovate scalable solutions that can drive evidence-based urban policies.

Funding

This research is supported by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) 2024-25 (Project No.: 17202224).

References

*Refers to the ones included in the literature review.

Bardhan, M., Li, F., Browning, M. H., Dong, J., Zhang, K., Yuan, S., ... & Helbich, M. (2024). From space to street: A systematic review of the associations between visible greenery and bluespace in street view imagery and mental health. *Environmental Research*, *263*(3), 120213. https://doi.org/10.1016/j.envres.2024.120213

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, *215*, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

*Blečić, I., Saiu, V., & A. Trunfio, G. (2024, July). Enhancing Urban Walkability Assessment with Multimodal Large Language Models. In *International Conference on Computational Science and Its Applications* (pp. 394-411). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-65282-0_26

Chen, L., Lu, Y., Ye, Y., Xiao, Y., & Yang, L. (2022). Examining the association between the built environment and pedestrian volume using street view images. *Cities*, *127*, 103734. https://doi.org/10.1016/j.cities.2022.103734

*Chen, M., Li, Z., Huang, W., Gong, Y., & Yin, Y. (2024, August). Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 319-328). https://doi.org/10.1145/3637528.3671918

*Cheng, Y., Yin, Z., Li, D., & Li, Z. (2024, October). Assessing Urban Safety: A Digital Twin Approach Using Streetview and Large Language Models. In *2024 IEEE 100th Vehicular Technology Conference* (VTC2024-Fall) (pp. 1-5). IEEE. https://doi.org/10.1109/VTC2024-Fall63153.2024.10757666

*Cepeda, V. V., Nayak, G. K., & Shah, M. (2023). Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, *36*, 8690-8701.

Dai, S., Li, Y., Stein, A., Yang, S., & Jia, P. (2024). Street view imagery-based built environment auditing tools: a systematic review. *International Journal of Geographical Information Science*, *38*(6), 1136-1157. https://doi.org/10.1080/13658816.2024.2336034

*de Moraes V. K., Phillipi C. S., Brovelli, M. A., & Rodrigues dos Santos, D. (2024). Investigating the Performance of Open-Vocabulary Classification Algorithms for Pathway and Surface Material Detection in Urban Environments. *ISPRS International Journal of Geo-Information*, *13*(12), 422. https://doi.org/10.3390/ijgi13120422

*del Castillo, N. D., Neri, I., & Bogdanović, R. (2023). CLIP and the City: Addressing the Artificial Encoding of Cities in Multimodal Foundation Deep Learning Models. OpenReview.

*Feng, J., Zhang, J., Liu, T., Zhang, X., Ouyang, T., Yan, J., ... & Li, Y. (2024). CityBench:

Evaluating the Capabilities of Large Language Models for Urban Tasks. arXiv preprint. https://doi.org/10.48550/arXiv.2406.13945

*Gan, Y., Li, G., Togo, R., Maeda, K., Ogawa, T., & Haseyama, M. (2024, October). Cross-domain few-shot in-context learning for enhancing traffic sign recognition. In *2024 IEEE International Conference on Image Processing (ICIP)* (pp. 2564-2570). IEEE. https://doi.org/10.1109/ICIP51287.2024.10647129

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, *8*(3), 331-368. https://doi.org/10.1007/s41095-022-0271-y

Haddaway, N. R., Woodcock, P., Macura, B., & Collins, A. (2015). Making literature reviews more reliable through application of lessons from systematic reviews. *Conservation Biology*, *29*(6), 1596-1605. https://doi.org/10.1111/cobi.12541

*Haas, L., Alberti, S., & Skreta, M. (2023). Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv preprint*. https://doi.org/10.48550/arXiv.2302.00275

*Haas, L., Skreta, M., Alberti, S., & Finn, C. (2024). Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12893-12902).

*Hao, X., Chen, W., Yan, Y., Zhong, S., Wang, K., Wen, Q., & Liang, Y. (2024). UrbanVLP: Multi-Granularity Vision-Language Pretraining for Urban Socioeconomic Indicator Prediction. *arXiv preprint*. https://doi.org/10.48550/arXiv.2403.16831

He, N., & Li, G. (2021). Urban neighbourhood environment assessment based on street view image processing: A review of research trends. *Environmental Challenges*, *4*, 100090. https://doi.org/10.1016/j.envc.2021.100090

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., ... & Wang, H. (2024a). Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, *33*(8), 1-79. https://doi.org/10.1145/3695988

Hou, Y., Quintana, M., Khomiakov, M., Yap, W., Ouyang, J., Ito, K., ... & Biljecki, F. (2024b). Global Streetscapes—A comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. *ISPRS Journal of Photogrammetry and Remote Sensing*, *215*, 216-238. https://doi.org/10.1016/j.isprsjprs.2024.06.023

*Hu, Y., Ou, D., Wang, X., & Yu, R. (2023, December). Enabling vision-and-language navigation for intelligent connected vehicles using large pre-trained models. In *2023 IEEE International Conferences on Internet of Things and IEEE Green Computing & Communications and IEEE Cyber, Physical & Social Computing and IEEE Smart Data and IEEE Congress on Cybermatics* (pp. 390-396). IEEE. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics60724.2023.00083

*Huang, W., Wang, J., & Cong, G. (2024). Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *International Journal of Geographical Information Science*, *38*(7), 1414-1442. https://doi.org/10.1080/13658816.2024.2347322

*Juhász, L., Mooney, P., Hochmair, H. H., & Guan, B. (2023). ChatGPT as a mapping assistant: A novel method to enrich maps with generative AI and content derived from street-level photographs. *arXiv preprint*. https://doi.org/10.25436/E2ZW27

Keralis, J. M., Javanmardi, M., Khanna, S., Dwivedi, P., Huang, D., Tasdizen, T., & Nguyen, Q. C. (2020). Health and the built environment in United States cities: Measuring associations using GSV-derived indicators of the built environment. *BMC Public*

Health, *20*, 1-10. https://doi.org/10.1186/s12889-020-8300-1

Koo, B. W., Guhathakurta, S., & Botchwey, N. (2022). How are neighborhood and street-level walkability factors associated with walking behaviors? A big data approach using street view images. *Environment and Behavior*, *54*(1), 211-241. https://doi.org/10.1177/00139165211014609

Li, Y., Peng, L., Wu, C., & Zhang, J. (2022). 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., Xu, J., Wang, S., Wu, Y., & Li, H. (2024). StreetviewLLM: Extracting Geographic Information Using a Chain-of-Thought Multimodal Large Language Model. *arXiv preprint*. https://doi.org/10.48550/arXiv.2411.14476

*Liang, H., Zhang, J., Li, Y., Wang, B., & Huang, J. (2024). Automatic Estimation for Visual Quality Changes of Street Space Via Street-View Images and Multimodal Large Language Models. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3408843

*Liu, X., Haworth, J., & Wang, M. (2023, November). A New Approach to Assessing Perceived Walkability: Combining Street View Imagery with Multimodal Contrastive Learning Model. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications* (pp. 16-21). https://doi.org/10.1145/3615888.3627811

*López-Otero, J., Obregón-Sierra, Á., & Gavira-Narváez, A. (2024). Migration and Segregated Spaces: Analysis of Qualitative Sources Such as Wikipedia Using Artificial Intelligence. *Social Sciences*, *13*(12), 664. https://doi.org/10.3390/socsci13120664

*Lyu, Z., Zhang, J., Lu, M., Li, Y., & Feng, C. (2024). Tell me where you are: Multimodal llms meet place recognition. arXiv preprint arXiv:2406.17520.

*Ma, Y., Zhang, T., & Zhan, G. (2024, June). An LLM-based Intelligent System for the Evaluation of Property Geographical Environment. In *2024 International Symposium on Intelligent Robotics and Systems (ISoIRS)* (pp. 258-262). IEEE. https://doi.org/10.1109/ISoIRS63136.2024.00057

*Malekzadeh, M., Willberg, E., Torkko, J., & Toivonen, T. (2025). Urban attractiveness according to ChatGPT: Contrasting AI and human insights. *Computers, Environment and Urban Systems*, *117*, 102243. https://doi.org/10.1016/j.compenvurbsys.2024.102243

Nah, F. F. H., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277-304. https://doi.org/10.1080/15228053.2023.2233814

*Ouyang, T., Zhang, X., Han, Z., Shang, Y., & Li, Y. (2024, May). Health CLIP: Depression Rate Prediction Using Health Related Features in Satellite and Street View Images. In *Companion Proceedings of the ACM Web Conference 2024* (pp. 1142-1145). https://doi.org/10.1145/3589335.3651451

*Pan, F., Jeon, S., Wang, B., Mckenna, F., & Yu, S. X. (2024). Zero-shot Building Attribute Extraction from Large-Scale Vision and Language Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 8647-8656).

*Park, S., Kim, H., & Ro, Y. M. (2024). Integrating language-derived appearance elements with visual cues in pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(9), 7975-7985. https://doi.org/10.1109/TCSVT.2024.3383914

*Ramalingam, S. P., & Kumar, V. (2025). Building usage prediction in complex urban scenes by fusing text and facade features from street view images using deep learning.

Building and Environment, *267*, 112174. https://doi.org/10.1016/j.buildenv.2024.112174

Ren, M., Zhang, X., Zhi, X., Wei, Y., & Feng, Z. (2024). An annotated street view image dataset for automated road damage detection. *Scientific Data*, *11*(1), 407. https://doi.org/10.1038/s41597-024-03263-7

*Roberts, J., Lüddecke, T., Sheikh, R., Han, K., & Albanie, S. (2024). Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 554-563).

*Schumann, R., Zhu, W., Feng, W., Fu, T. J., Riezler, S., & Wang, W. Y. (2024, March). Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(17), 18924-18933. https://doi.org/10.1609/aaai.v38i17.29858

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, *3*(5), 637-646. https://doi.org/10.1109/JIOT.2016.2579198

*Shihab, I. F., Bhagat, S. R., & Sharma, A. (2024). Precise and Robust Sidewalk Detection: Leveraging Ensemble Learning to Surpass LLM Limitations in Urban Environments. *arXiv preprint*. https://doi.org/10.48550/arXiv.2405.14876

*Tian, Y., Carballo, A., Li, R., Thompson, S., & Takeda, K. (2025). Query by example: Semantic traffic scene retrieval using LLM-based scene graph representation. *Sensors*, *25*(8), 2546. https://doi.org/10.3390/s25082546

Turkcan, M. K., Li, Y., Zang, C., Ghaderi, J., Zussman, G., & Kostic, Z. (2024). Boundless: Generating Photorealistic Synthetic Data for Object Detection in Urban Streetscapes. *arXiv preprint*. https://doi.org/10.48550/arXiv.2409.03022

*Verma, D., Mumm, O., & Carlow, V. M. (2023). Generative agents in the streets: Exploring the use of Large Language Models (LLMs) in collecting urban perceptions. *arXiv preprint*. https://doi.org/10.48550/arXiv.2312.13126

*Wang, S., Liang, C., Gao, Y., Liu, Y., Li, J., & Wang, H. (2024, October). Decoding Urban Industrial Complexity: Enhancing Knowledge-Driven Insights via IndustryScopeGPT. In *Proceedings of the 32nd ACM International Conference on Multimedia* (pp. 4757-4765). https://doi.org/10.1145/3664647.3681705

*Wen, L., Yang, X., Fu, D., Wang, X., Cai, P., Li, X., ... & Qiao, Y. (2023). On the Road with GPT-4V(ision): Early Explorations of Visual-Language Model on Autonomous Driving. arXiv preprint. https://doi.org/10.48550/arXiv.2311.05332

*Wu, M., & Huang, Q. (2022, November). Im2city: image geo-localization via multi-modal learning. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (pp. 50-61). https://doi.org/10.1145/3557918.3565868

*Yang, J., Ding, R., Brown, E., Qi, X., & Xie, S. (2024, September). V-IRL: Grounding Virtual Intelligence in Real Life. In *European Conference on Computer Vision* (pp. 36-55). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72995-9_3

*Yao, J., Li, J., Li, Y., Zhang, M., Zuo, C., Dong, S., & Dai, Z. (2024). A Vision–Language Model-Based Traffic Sign Detection Method for High-Resolution Drone Images: A Case Study in Guyuan, China. *Sensors*, *24*(17), 5800. https://doi.org/10.3390/s24175800

*Yin, W., Xue, Y., Liu, Z., Li, H., & Werner, M. (2025). LLM-enhanced disaster geolocalization using implicit geoinformation from multimodal data: A case study of Hurricane Harvey. *International Journal of Applied Earth Observation and Geoinformation*, *137*, 104423. https://doi.org/10.1016/j.jag.2025.104423

*Zeng, T., Wu, L., Shi, L., Zhou, D., & Guo, F. (2025). Are vision llms road-ready? a comprehensive benchmark for safety-critical driving video understanding. arXiv preprint arXiv:2504.14526.

*Zhang, D., Song, K., & Zhao, D. (2024a). Leveraging Multi-Source Data for the Trustworthy Evaluation of the Vibrancy of Child-Friendly Cities: A Case Study of Tianjin, China. *Electronics*, *13*(22), 4564. https://doi.org/10.3390/electronics13224564

*Zhang, Y., Ma, R., Zhang, X., & Li, Y. (2025c, April). Perceiving urban inequality from imagery using visual language models with chain-of-thought reasoning. In *Proceedings of the ACM on Web Conference 2025* (pp. 5342-5351). https://doi.org/10.1145/3696410.3714536

*Zhang, Y., Zhang, F., & Chen, N. (2022). Migratable urban street scene sensing method based on vision language pre-trained model. *International Journal of Applied Earth Observation and Geoinformation*, *113*, 102989. https://doi.org/10.1016/j.jag.2022.102989

*Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., & Wang, X. (2025, April). Drivedreamer-2: LLM-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(10), 10412-10420. https://doi.org/10.1609/aaai.v39i10.33130

Zhou, H., Wang, J., Widener, M., & Wilson, K. (2024). Examining the relationship between active transport and exposure to streetscape diversity during travel: A study using GPS data and street view imagery. *Computers, Environment and Urban Systems*, *110*, 102105. https://doi.org/10.1016/j.compenvurbsys.2024.102105

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9), 2337-2348. https://doi.org/10.1007/s11263-022-01653-1