Vision-Language Models for Urban Digital Twins

Amirhossein Nourbakhshrezaei[1], Saeed Abbasi[1], Mojgan Jadidi[1,*]

[1]Lassonde School of Engineering, York University, 4700 Keele Street, Toronto, ON M3J 1P3, ON, Canada,
amirnbr2@yorku.ca, saeedabb@yorku.ca, mjadidi@yorku.ca,

Keywords: Digital Twin, Vision-Language Model, Urban Sensing, Multi-modal Language Models, Semantic Scene Understanding

Abstract

Urban digital twins are virtual city replicas that can greatly support urban planning by simulating infrastructure and mobility scenarios. However, keeping a digital twin up-to-date with fine-grained, real-world urban conditions is challenging. This paper proposes a novel system that leverages multi-modal AI models to bridge the gap between physical urban data collection and a 3D city digital twin. In proposed approach, ordinary smartphones carried in vehicles act as mobile sensors, continuously capturing multi-modal data (road images, GPS coordinates, and speed). Advanced vision-language models then analyze the data to automatically extract information from the traffic and road infrastructure and detect road anomalies. The extracted information such as the locations of traffic signs, traffic signals, road surface, and potential blind spots at intersections is geo-tagged and streamed into vision-language models to interpret data and stream human readable insights into the digital twin model. The case study is the digital twin of the city of Toronto. By aggregating data from many drivers and analyzing it (in post-processing for high accuracy), the digital twin evolves into a living model of the built environment. This enriched and dynamic twin provides urban planners with up-to-date insights on traffic signage, road conditions, and other relevant road infrastructure elements, enabling proactive maintenance and informed decision-making for city planning.

1. Introduction

Digital twins are complete virtual representations of real-world environments, often used by cities as "living" models of urban infrastructure. By integrating large datasets about a city, a digital twin can mirror the state of physical assets and even simulate the impact of changes in a risk-free virtual space. Thus, city planners can test scenarios and visualize outcomes (e.g, new traffic patterns or road layouts) before implementing them in reality. Despite their promise, current urban digital twins typically rely on fixed sensor infrastructures. This limits their ability to capture the dynamic, fine-grained changes occur in built environments (Jones and Smith, 2022).

Multi-modal AI advances now allow us to fully exploit these rich data streams. Traditional computer vision methods for detecting road features (signs, defects, etc.) typically require training on specific object classes and large labeled datasets. In contrast, the latest vision-language models (also known as multi-modal language models) can interpret images in a more flexible, general-purpose way by leveraging learned associations between visual content and natural language (Radford et al., 2021). For example, OpenAI's GPT-4 model is a large-scale multi-modal Transformer that accepts image inputs and produces text outputs, demonstrating near human-level ability to describe and analyze visual scenes (OpenAI, 2023). Such models can recognize a virtually unlimited set of visual concepts by name or description, without needing a bespoke classifier for each new object. Recent research in remote sensing confirms the value of multi-modal models for geospatial analysis, they have been used for tasks like scene description, object detection, and change detection on imagery, proving useful for applications in environmental monitoring and urban planning (Saidi et al., 2024).

The proposed platform leverages foundation models to extract rich semantic descriptions from street-level imagery, moving beyond simple object labels to capture contextual details, such as identifying a "faded stop sign with graffiti" or a "longitudinal road crack near the curb." These descriptions, when combined with geospatial and sensor metadata, enable the digital twin to understand not only the 3D coordinates of detected events or objects, but also their descriptive attributes, resulting in a more informative and situational awareness of urban representation.

2. Problem Definition

Urban digital twins require continuous, fine-grained, and context-aware updates to accurately represent the evolving state of real-world environments. However, existing approaches to maintain such models face several fundamental challenges. First, most sources of data are based on periodically updated information derived from surveys, satellite imagery, or fixed sensor infrastructures. Although these sources provide reliable baseline information, they fail to capture the high-frequency, localized, and transient changes that occur at the street level, such as temporary road hazards, degraded infrastructure, obstructed signage, or evolving traffic conditions. As a result, the digital twin gradually diverges from real-world conditions, reducing its effectiveness for planning and operational decision-making. Second, traditional sensing approaches lack scalability and spatial coverage. Fixed sensors (such as traffic cameras or IoT devices) are expensive to deploy and maintain, and their coverage is inherently limited to predefined locations. Third, existing computer vision and geospatial analysis pipelines primarily focus on structured outputs such as object detection labels or numerical sensor readings. While useful, these outputs lack semantic richness and contextual understanding. Finally, integrating heterogeneous, crowd-sourced data into a unified spatial framework introduces challenges related to geolocation accuracy, data alignment, and semantic consistency. Raw GPS data collected from mobile devices is often noisy or imprecise, particularly in dense urban environments,

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-425-2026 |

leading to spatial inaccuracies when mapping observations to the digital twin.

To address these challenges, this paper focuses on the problem of enabling scalable, continuous, and semantically rich urban data acquisition using mobile phone application, and integrating this information into a digital twin. Specifically, the objective is to capture street-level observations using widely available devices, extract meaningful semantic information from visual data using vision-language models, improve spatial accuracy through map-aware geolocation techniques, and represent the resulting insights in a form that enhances both machine processing and human interpretability within the digital twin environment.

3. Background

Zhang et al (Zhang et al., 2025) argue that mobile crowd-sensed data information collected via ubiquitous citizen-carried devices can be a "game-changing resource" for digital twins providing high-resolution, context-rich updates. Integrating anonymized, aggregated data from smartphones (e.g., geotagged images, GPS traces, motion sensors) can thus enrich a digital twin with timely insights on urban conditions and human activities, enabling more adaptive and responsive planning decisions. For instance, phone 's accelerometer and gyroscope have been used to measure pavement roughness and classify defects, with machine learning models enhancing the accuracy of anomaly detection (Zareei et al., 2025). Nourbakhshrezaei et al. (Nourbakhshrezaei et al., 2023) developed a situational awareness system within intelligent transportation systems (ITS) to improve the safety of vulnerable road users. Their framework combines a mobile application, a server-side processing system, and a database management layer to collect and analyze both static and dynamic data sources. The system identifies high-risk regions and provides real-time notifications to users, thereby enhancing situational awareness in urban environments. Hu et al. (Hu et al., 2025) developed a method to continuously update high-definition (HD) city maps with traffic signs detected from "driving record data" (dashboard camera feeds). Their approach matches detected signs from crowd-contributed videos against the existing map, flags changes, or new signs, and then fuses observations from multiple drivers to pinpoint the new sign locations.

Nourbakhshrezaei et al. (Nourbakhshrezaei et al., 2024) developed an Advanced Driver Assistance System (ADAS) within ITS to enhance safety at urban traffic intersections. The proposed framework integrates infrastructure-based sensing, such as roadside cameras, with computer vision techniques and location-based services to extend the driver's field of view and detect pedestrians or vehicles in occluded regions. By improving spatial awareness beyond onboard sensor limitations, the system reduces collision risk and enhances the safety of vulnerable road users. Yamazaki et al. (Yamazaki et al., 2025) propose a 3D mobile crowdsensing framework in which volunteers use their smartphones to collect point cloud data from the environment, which is then integrated into a city's digital twin model. Similarly, Nummenmaa et al. (Nummenmaa et al., 2025) explored close-range crowdsourced sensing using smartphone LiDAR and Augmented Reality (AR) gaming techniques to create digital replicas of real-world objects and scenes. They report that recent high-end phones can generate dense 3D scans of urban features, and by engaging users through interactive AR mini-games, large volumes of geospatial data can be gathered in a fun and scalable way. The proposed method builds on these insights and combines them in a novel way that enhances them with state-of-the-art vision–language models to automatically interpret the collected imagery. By situating our contributions in the context of these recent developments, we aim to advance the conversation on how urban digital twins can continuously self-update through the power of crowd sensing and artificial intelligence.

4. Contribution

The proposed system addresses critical limitations in the current generation of urban digital twins, particularly their reliance on static data sources or sparse fixed-sensor networks. Static datasets rapidly become outdated, failing to reflect the dynamic nature of urban infrastructure, while stationary sensors are expensive to deploy at scale and inherently limited in spatial coverage, often leaving areas such as sidewalks, bike lanes, or temporarily obstructed zones unsensed. Furthermore, many fixed sensors are not equipped to detect the wide range of temporary, context-specific events (e.g., construction activity, temporary signage, road closures) that frequently occur in urban environments. By leveraging mobile smartphones as roaming multi-modal sensing platforms, the proposed approach enables comprehensive data collection at street level, including regions not easily accessible by traditional infrastructure. When paired with advanced vision-language models, the system not only detects and localizes urban features and abnormalities, but also generates descriptive, human-readable metadata for each event along with its 3D geolocation. This marks a departure from conventional systems that only record categorical labels or sensor values. These descriptive insights are then integrated directly into the 3D digital twin, enriching its semantic depth and interpretability for city planners and engineers.

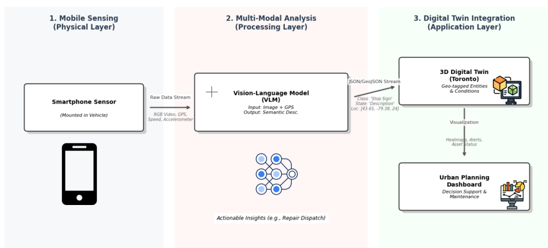


Figure 1. Architecture of the proposed Vision-Language Model integrating with Digital Twin.

5. Methodology

The proposed system integrates mobile crowd-sensing, vision-language modeling, and geospatial processing into a unified pipeline for continuously enriching urban digital twins with semantically rich, geo-referenced observations. The overall architecture of the system is illustrated in Fig. 1. The methodology consists of five main stages: (1) data acquisition, (2) preprocessing and geolocation refinement, (3) vision-language interpretation, (4) spatial structuring and storage, and (5) visualization and digital twin integration.

5.1 Data Acquisition

Data collection is performed using standard smartphones or dash-cam on vehicles, acting as mobile multi-modal sensing

Attribute	Description
Timestamp	2025-09-01T21:23:54+00:00
Device	dev-00000000
Coordinates	43.624540, -79.477565
Detections	3
Scene Type	Urban street / traffic queue
Detector Labels (Mobile App)	car, traffic light
VLM Summary	Urban roadway scene with multiple cars queued ahead, viewed from inside a vehicle on a tree-lined street.
Interpretation	This appears to be a city street segment with slow or stopped traffic, likely approaching an intersection or controlled roadway area.
Context	The image shows an urban corridor with traffic moving or waiting in multiple lanes, bordered by trees and mid-rise buildings.
Visible Objects	cars, streetlights, trees, buildings, roadway, vehicle dashboard/windshield edge
Observations	Multiple vehicles are ahead in the lane; a white car is visible on the right side; tree canopy lines the street; mid-rise buildings are visible in the background; the image is taken from inside a vehicle; streetlights are present along the road.
Mobility	Vehicle traffic is present; traffic appears slow or stopped; no pedestrians are clearly visible; lane markings are not clearly visible.
Infrastructure	Multi-lane paved roadway; street lighting; roadside trees; urban buildings adjacent to the street.
Safety Notes	Traffic congestion may require cautious driving; road conditions ahead are partially obscured by vehicles; possible limited visibility due to windshield reflections.

Figure 2. Case 1: Mobile-app detections and VLM enrichment for a congested urban street corridor.

platforms. As the vehicle moves through the urban environment, the device continuously captures image frames of the forward scene along with associated telemetry data, including timestamp, GPS coordinates, speed, and sensor accuracy. In addition to raw imagery, lightweight object detection is applied on-device to extract candidate objects (e.g., traffic signs, pedestrians, vehicles), represented as labeled bounding boxes with confidence scores. Each captured frame is therefore associated with both visual data and structured metadata, forming the basis for downstream analysis.

5.2 Preprocessing and Geolocation Refinement

Raw GPS measurements obtained from smartphones are often affected by noise, particularly in dense urban environments with signal obstruction. To improve spatial accuracy, a map-constrained geolocation refinement step is introduced. Specifically, each observation point is projected onto nearby pedestrian infrastructure using Open Street Map (OSM) data. Sidewalk geometries are retrieved through spatial queries, and the recorded GPS coordinate is snapped to the nearest valid sidewalk segment within a defined radius. This process reduces positional error and ensures that detected events are aligned with meaningful urban features (e.g., sidewalks or road edges), improving both spatial consistency and interpretability within the digital twin.

5.3 Vision-Language Interpretation

Each captured frame is processed using a vision-language model capable of jointly interpreting visual content and generating structured textual outputs. Unlike traditional computer vision pipelines that produce fixed class labels, the model generates rich, context-aware descriptions of the scene. For each image, the model generates a structured semantic representation consisting of a concise scene summary, a detailed environmental interpretation, a list of visible objects, contextual information about the urban setting, observations and inferred conditions, mobility-related insights, infrastructure elements, and safety-related notes. These outputs enable a higher-level semantic understanding of urban scenes, such as identifying degraded infrastructure, temporary obstructions, or complex contextual conditions that are not captured by simple object detection.

5.4 Spatial Structuring and Storage

Following semantic interpretation, each observation is transformed into a structured geospatial record that combines refined geographic coordinates, timestamp and device metadata, detected objects with associated confidence scores, and outputs generated by the vision-language model. The data is serialized into a standardized format, including GeoJSON representations for spatial indexing and visualization. Each observation is associated with its corresponding image and stored in a cloud-based storage system. Metadata and trajectory data are stored in compressed formats to enable efficient batch processing and retrieval.

5.5 Visualization and Digital Twin Integration

The processed observations are integrated into an interactive geospatial visualization interface, forming a dynamic layer within the digital twin. Each observation is represented as a geo-referenced point enriched with semantic attributes derived from the vision-language model. The interface enables real-time or batch visualization of collected observations, supports clustering of spatial events for scalable rendering, allows filtering based on detection confidence or semantic attributes, and provides interactive exploration of individual observations, including associated images and generated descriptions. By aggregating contributions from multiple users over time, the digital twin evolves into a continuously updated representation of the urban environment. The integration of semantic insights alongside spatial data enables planners and engineers to not only locate events, but also understand their context and potential impact, thereby enhancing decision-making processes.

6. Results and Discussion

The proposed framework was evaluated using geo-referenced observations captured by the mobile sensing application. In this workflow, the Android-based mobile application first detects potential objects or points of interest in the urban scene, while the vision-language model (VLM) subsequently enriches these detections with higher-level semantic interpretation, contextual understanding, and safety-relevant observations. Each case therefore illustrates the transition from raw mobile sensing outputs to semantically enriched urban knowledge suitable for digital twin integration.

6.1 Representative Case Studies

Case 1 (Fig 2) demonstrates the complementary roles of mobile sensing and semantic enrichment. The mobile application identifies basic traffic-related objects, including cars and a traffic light, indicating a potentially important roadway condition. The VLM then expands this limited detection set into a richer scene interpretation by identifying congestion, multi-lane traffic organization, surrounding buildings, vegetation, and visibility constraints. This case highlights how the proposed framework transforms sparse detections into a more informative urban traffic narrative that can support digital twin monitoring and traffic-aware interpretation.

Case 2 (Fig 3) shows a situation in which the mobile application produced a false positive object, identifying a traffic light as the main point of interest. Despite this false detected object, the VLM successfully reconstructs a correct understanding of the scene, including roadway structure, adjacent buildings, vegetation, visibility conditions, and overall mobility status. This case is important because it demonstrates that the proposed framework does not rely solely on dense detector outputs; rather, it uses VLM-based reasoning to recover urban context and infrastructure semantics even when the mobile detections are sparse.

Case 3 (Fig 4) illustrates the framework's usefulness for safety-sensitive urban interpretation. The mobile application detects a person and a traffic-related object, thereby flagging the location as potentially important. The VLM enriches this information by recognizing the stop sign, crosswalk, curbside vehicle, and the interaction context between pedestrians and the roadway. It also produces safety-aware statements regarding yielding behavior and restricted visibility caused by the interior framing of the vehicle. This demonstrates the value of combining low-level detection with semantic reasoning for pedestrian-safety and intersection analysis.

Case 4 (Fig 5) represents a more complex and dense urban setting involving multiple travel lanes, traffic signal infrastructure, pedestrians, and surrounding high-rise development. Here, the mobile application detects a person, traffic light, and car, which provides a useful initial indication of a mixed-use roadway environment. The VLM then builds on these detections by characterizing the scene as a signal-controlled urban junction with marked pedestrian infrastructure and high-rise surroundings. The semantic enrichment is especially valuable in this case because it connects individual detections into a coherent interpretation of urban activity, infrastructure, and safety conditions, which is highly relevant for digital twin representations of busy intersections.

Across all cases (2, 3, 4, 5, 6), a consistent pattern can be observed: the mobile sensing application serves as an efficient first-stage detector that identifies candidate objects or areas of interest, while the VLM performs the second-stage enrichment by generating contextual, environmental, mobility-related, and safety-related understanding. This two-step design is important because mobile detections alone are often too sparse or too low-level to support meaningful urban analysis. In contrast, the VLM outputs provide structured semantic descriptions that make the observations more useful for digital twin integration.

The results also show that the framework performs well across different urban contexts, including congested traffic corridors, light-traffic arterials, stop-controlled pedestrian crossings, and signalized intersections in dense built-up districts. In each case, the VLM adds information beyond the detector labels by identifying roadway type, surrounding infrastructure, pedestrian interaction, traffic condition, and visibility constraints. This supports the central objective of the proposed work: moving from raw mobile observations toward semantically enriched urban intelligence. At the same time, the results suggest that the quality of the final interpretation remains dependent on image clarity, viewing angle, and windshield-related artifacts. Some safety observations are inferred under partial visibility, and the

Attribute	**Description**
Timestamp	2025-09-01T21:21:51+00:00
Device	dev-00000000
Coordinates	43.623846, -79.479349
Detections	2
Scene Type	urban street
Detector Labels (Mobile App)	traffic light
VLM Summary	Urban roadway view from inside a vehicle on a multi-lane street near large buildings and trees.
Interpretation	The scene appears to be a daytime city street or arterial road with light traffic, viewed through a car windshield.
Context	An urban street corridor with adjacent buildings and roadside trees, likely part of a commercial or office district.
Visible Objects	roadway, cars, buildings, trees, streetlights, windshield reflection, vehicle dashboard
Observations	View is from inside a vehicle; road has multiple lanes; several cars are visible ahead and on the side; tall building edge is visible on the left; trees line the far side of the road; streetlights are present; sky is clear and bright.
Mobility	vehicle traffic is present; road appears drivable; no pedestrians are visible; no cyclists are visible.
Infrastructure	multi-lane asphalt road; street lighting; adjacent mid-rise building; roadside tree line.
Safety Notes	Windshield reflections may reduce visibility; traffic ahead requires normal driving caution.

Figure 3. Case 2: Mobile-app detections and VLM enrichment for an urban street with light traffic.

detector stage may omit important elements that are later recovered only by the VLM. Future work should therefore investigate stronger front-end detection, temporal fusion across consecutive frames, and broader quantitative evaluation across larger datasets.

7. Conclusion

This paper presents that combining multi-modal vision–language models with mobile crowdsensing can significantly enhance urban digital twins. The proposed framework enables automated extraction of geo-referenced, semantically rich features, improving spatial coverage and responsiveness beyond static datasets. This paper details the design and components of such a system and discusses its potential to improve urban planning workflows. The results confirm its potential for dynamic, self-updating digital twins that support efficient and informed planning and infrastructure management. Future efforts will focus on real-time processing, scalability, and predictive analytics.

References

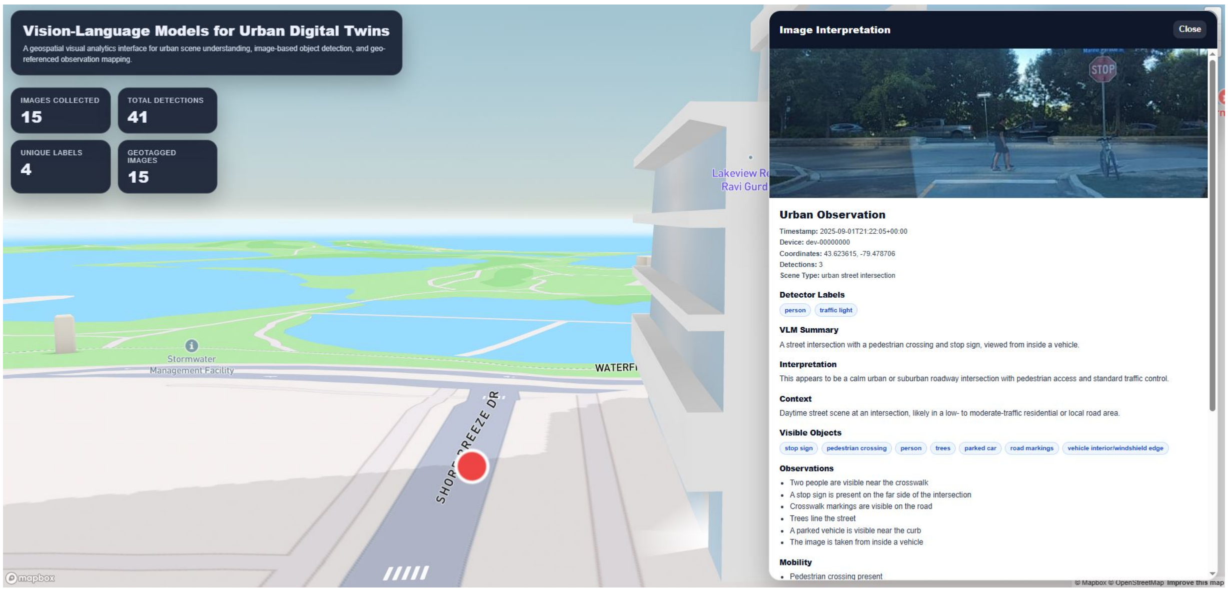
Hu, H., Wu, H., Huang, S., Huang, W., Liu, C., 2025. Incremental Crowd-Source Data Fusion and Map Update Method Based on Driving Data for Traffic Signs. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 641–647.

Jones, A., Smith, B., 2022. Smart cities and digital twins: A dynamic pairing. *Nature Urban Intelligence*, 1, 1–5.

Nourbakhshrezaei, A., Jadidi, M., Delavar, M. R., Moshiri, B., 2024. A novel context-aware system to improve driver’s field of view in urban traffic networks. *Journal of Intelligent Transportation Systems*, 28(3), 297–312.

Nourbakhshrezaei, A., Jadidi, M., Sohn, G., 2023. Improving cyclists’ safety using intelligent situational awareness system. *Sustainability*, 15(4), 2866.

Nummenmaa, T., Laato, S., Chambers, P., Yrttimaa, T., Vastaranta, M., Buruk, O. T., Hamari, J., 2025. Employing gamified crowdsourced close-range sensing in the pursuit of a digital



Attribute	**Description**
Timestamp	2025-09-01T21:22:05+00:00
Device	dev-00000000
Coordinates	43.623615, -79.478706
Detections	3
Scene Type	urban street intersection
Detector Labels (Mobile App)	person, traffic light
VLM Summary	A street intersection with a pedestrian crossing and stop sign, viewed from inside a vehicle.
Interpretation	This appears to be a calm urban or suburban roadway intersection with pedestrian access and standard traffic control.
Context	Daytime street scene at an intersection, likely in a low- to moderate-traffic residential or local road area.
Visible Objects	stop sign, pedestrian crossing, person, trees, parked car, road markings, vehicle interior/windshield edge
Observations	Two people are visible near the crosswalk; a stop sign is present on the far side of the intersection; crosswalk markings are visible on the road; trees line the street; a parked vehicle is visible near the curb; the image is taken from inside a vehicle.
Mobility	Pedestrian crossing present; vehicle traffic controlled by stop sign; road appears usable for cars; no active moving vehicles are clearly visible.
Infrastructure	Intersection; stop sign; marked crosswalk; curbed roadway; streetlight or utility pole present.
Safety Notes	Pedestrians are near the crossing area; driver should yield at the stop-controlled intersection; visibility of the intersection appears partially affected by the vehicle interior framing.

Figure 4. Case 3: Mobile-app detections and VLM enrichment for a stop-controlled intersection with pedestrian activity.

twin of the earth. *International Journal of Human–Computer Interaction*, 41(8), 4668–4684.

OpenAI, 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

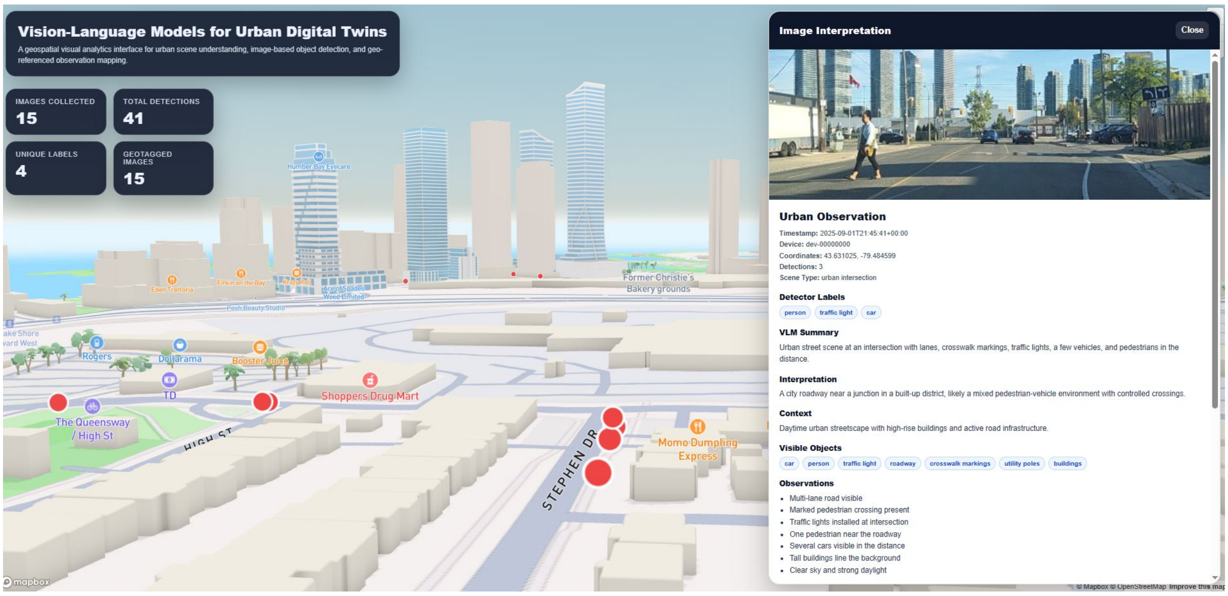
Radford, A., Kim, J., Hallacy, C. et al., 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Saidi, S., Idbraim, S., Karmoude, Y., Masse, A., Arbelo, M., 2024. Deep-learning for change detection using multi-modal fusion of remote sensing images: a review. *Remote Sensing*, 16(20), 3852.

Yamazaki, T., Watanabe, K., Kase, T., Hasegawa, K., Saida, K., Miyoshi, T., 2025. A 3D Mobile Crowdsensing Framework for Sustainable Urban Digital Twins. *IEEE Consumer Electronics Magazine*, 1–12 (early access).

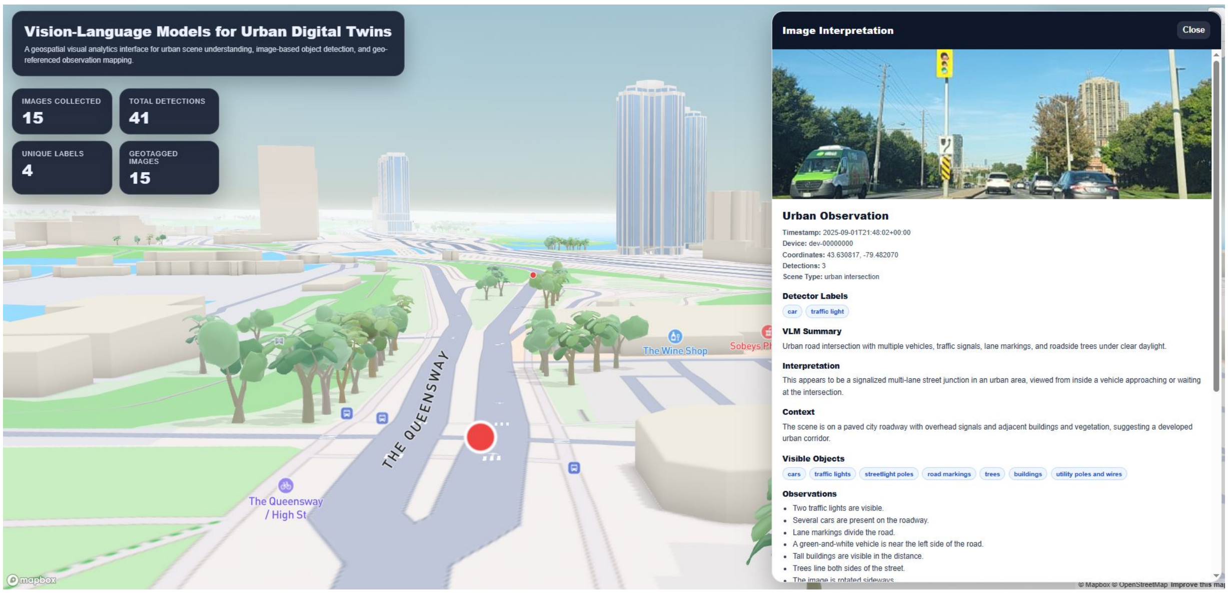
Zareei, M., Castañeda, C. A. L., Alanazi, F., Granda, F., Díaz, J. A. P., 2025. Machine learning model for road anomaly detection using smartphone accelerometer data. *IEEE Access*.

Zhang, Y., Yin, Y., Hu, Y., Sun, G., 2025. Fueling urban digital twins with mobile crowd data. *Nature Cities*.



Attribute	Description
Timestamp	2025-09-01T21:45:41+00:00
Device	dev-00000000
Coordinates	43.631025, -79.484599
Detections	3
Scene Type	urban intersection
Detector Labels (Mobile App)	person, traffic light, car
VLM Summary	Urban street scene at an intersection with lanes, crosswalk markings, traffic lights, a few vehicles, and pedestrians in the distance.
Interpretation	A city roadway near a junction in a built-up district, likely a mixed pedestrian-vehicle environment with controlled crossings.
Context	Daytime urban streetscape with high-rise buildings and active road infrastructure.
Visible Objects	car, person, traffic light, roadway, crosswalk markings, utility poles, buildings
Observations	Multi-lane road visible; marked pedestrian crossing present; traffic lights installed at intersection; one pedestrian near the roadway; several cars visible in the distance; tall buildings line the background; clear sky and strong daylight.
Mobility	Vehicular traffic on paved roadway; pedestrian crossing opportunity at marked crosswalk; signal-controlled intersection.
Infrastructure	Asphalt road surface; lane markings; crosswalk striping; traffic signal hardware; street lighting poles; sidewalk/road edge areas.
Safety Notes	Pedestrian is close to active traffic lanes; crossing should follow signal control; driver visibility appears generally good in daylight.

Figure 5. Case 4: Mobile-app detections and VLM enrichment for a signalized intersection in a high-density urban area.



Attribute	Description
Timestamp	2025-09-01T21:48:02+00:00
Device	dev-00000000
Coordinates	43.630817, -79.482070
Detections	3
Scene Type	urban intersection
Detector Labels (Mobile App)	car, traffic light
VLM Summary	Urban road intersection with multiple vehicles, traffic signals, lane markings, and roadside trees under clear daylight.
Interpretation	Signalized multi-lane street junction in an urban area, viewed from inside a vehicle approaching or waiting at the intersection.
Context	Paved city roadway with overhead signals, adjacent buildings, and vegetation, representing a developed urban corridor.
Visible Objects	cars, traffic lights, streetlight poles, road markings, trees, buildings, utility poles and wires
Observations	Two traffic lights are visible; several cars are present on the roadway; lane markings divide the road; a green-and-white vehicle is near the left side; tall buildings are visible in the distance; trees line both sides of the street; the image is rotated sideways.
Mobility	Vehicular traffic is present; intersection controlled by traffic signals; multiple lanes available; no pedestrians are visible.
Infrastructure	Traffic signal poles; street lighting; paved multi-lane roadway; lane divider markings; overhead utility lines; adjacent urban buildings.
Safety Notes	Traffic signal compliance is required; vehicles near the intersection require caution for turning and merging; no pedestrian activity is observed.

Figure 6. Case 5: Mobile-app detections and VLM enrichment for a signalized multi-lane urban intersection.