From Terrain to Space: A Survey on Multidomain Data Lifecycle for Urban Embodied Agents

Penglei Sun [†], Song Tang [†], Jiawen Wen, Runwei Guan, Yuxuan Liang [*], Weiping Ding, Yang Yang, Xiaowen Chu [*]

Posted Date: 11 March 2026

doi: 10.20944/preprints202601.2155.v2

Keywords: data lifecycle; smart cities; embodied agent; data analytics and data science

 doi:10.20944/preprints202601.2155.v2

Article

From Terrain to Space: A Survey on Multidomain Data Lifecycle for Urban Embodied Agents

Penglei Sun [1,†], Song Tang [1,†], Jiawen Wen [1], Runwei Guan [1], Yuxuan Liang [1], Weiping Ding [2], Yang Yang [1] and Xiaowen Chu [1]

[1] The authors are with Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China; psun012@connect.hkust-gz.edu.cn; stang428@connect.hkust-gz.edu.cn; jwen341@connect.hkust-gz.edu.cn; runwayrwguan@hkust-gz.edu.cn; yyiot@hkust-gz.edu.cn

[2] The author is with School of Artificial Intelligence and Computer Science of Nantong University, Nantong, China and City University of Macau, Macau, China; dwp9988@163.com

* Correspondence: yuxuanliang@hkust-gz.edu.cn (Y.N.); xwchu@hkust-gz.edu.cn (X.C.)

† Equal contribution.

Abstract

Urban Embodied Agents (UrbanEAs) are emerging to actively interact with complex, large-scale city environments and generate vast, heterogeneous data streams, moving beyond the single-vehicle of existing autonomous driving. However, urban environments present distinct challenges, including environmental variability, limited observability, and interaction complexity. These challenges hinder the effectiveness of existing embodied agents, which have focused on controlled indoor environments, and expose the inherent limitations of relying on single-domain data. Therefore, establishing a comprehensive data lifecycle to fuse multidomain data from terrain, aerial, and space is a strategy for developing actionable embodied capabilities from raw urban streams. Distinct from existing surveys that follow a model-centric paradigm for urban computing or autonomous driving, we systematically propose and review a comprehensive Data Lifecycle from a multidomain data perspective, which is critical for the UrbanEA. First, we propose a unified framework containing four key stages of this lifecycle: Data Perception, Data Management, Data Modeling, and Task Application. Next, we establish a taxonomy for each stage of the lifecycle. Specifically, we detail the evolution from static data storage to active agent memory, and analyze integration strategies designed to bridge multidomain gaps. We demonstrate how UrbanEAs empower downstream tasks, including Urban Scene Question-Answering (SQA), Vision-Language Navigation (VLN), and Human-Agent Collaboration (HAC). Finally, we outline the social impact of the data lifecycle of UrbanEA and open research problems with the future directions. Our survey provides a roadmap for designing the robust, high-performance data frameworks essential for these UrbanEAs.

Keywords: data lifecycle; smart cities; embodied agent;data analytics and data science

1. Introduction

Modern cities are complex systems, dynamically interwoven with physical elements (such as buildings and roads), social structures, economic activities, and environmental factors [1,2]. They are not only the cornerstones of modern civilization but are also continuously evolving, driven by cultural, economic, and technological advancements [3]. Contemporary cities have evolved beyond being collections of physical spaces to become hubs of diverse information sources, composed of Internet of Things (IoT) devices, geospatial data, and sensor data [4,5].

Against this backdrop, the concept of the **Urban Embodied Agent** (UrbanEA) [6] has emerged. It is an intelligent system (such as an autonomous vehicle, robot, or virtual avatar) that possesses a physical or virtual body to directly perceive, reason about, and execute actions within complex urban environments. By leveraging these core capabilities, such agents are envisioned to help address critical

 doi:10.20944/preprints202601.2155.v2

urban problems, including human-aware assistance and interactive services. UrbanEA represents an extension of digital urban computing agents and autonomous driving: moving beyond the digital analysis of the former (e.g., traffic flow models) and the point-to-point mobility of the latter, UrbanEA leverages diverse physical forms to treat the city as an interactive semantic space, utilizing common-sense reasoning to understand social intentions and execute complex tasks directly in the physical world [7].

To fulfill this vision, the UrbanEA's core capability lies in its data lifecycle, which defines the entire pipeline from data perception to embodied application. As shown in Figure 1, this pipeline contains the following stage:

1. **Data Perception**. The agent perceives multimodal data from multidomain (handheld, vehicle, drone, satellite) to comprehensively perceive the physical world, marking the starting point of the data lifecycle [8].
2. **Data Management**. Task-driven data architectures organize massive, heterogeneous urban perception data to the memory of the UrbanEAs.
3. **Data Modeling**. In this stage, the data modeling strategies address the data gaps and construct a unified urban cognition.
4. **Task Application**. The structured data, after being processed through modeling, supports advanced UrbanEA tasks such as Urban Scene Question-Answering (SQA), Vision-Language Navigation (VLN), and Human-Agent Collaboration (HAC), and generates a positive social impact.

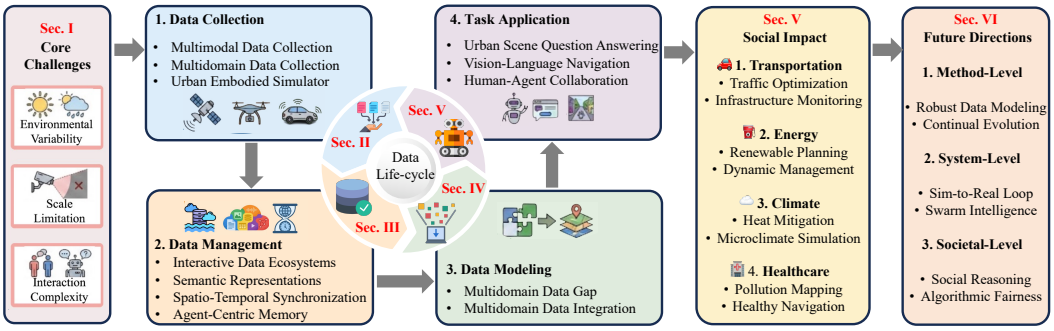


Figure 1. Overview of the proposed multidomain data lifecycle for Urban Embodied Agents, from perception to social impact.

Despite its importance, successfully implementing this end-to-end data lifecycle in urban settings remains insufficient. Existing embodied agent research focuses on indoor environments [9], which are relatively controlled, limited in scale, and structurally stable, while urban settings are different. Therefore, UrbanEA presents unique data-centric challenges that impact every stage of this lifecycle:

1. **Environmental Variability**. Urban environments are inherently dynamic and uncertain, unlike controlled indoor settings. **Data Perception** in outdoor scenes must handle dramatic variations in illumination (diurnal cycles, sunlight-shadow contrasts) and challenging weather conditions (rain, snow, fog), all of which degrade perception performance.
2. **Limited Observability**. Urban environments are vast, but individual sensors have limited coverage. Any single domain sensor (e.g., a vehicle-mounted camera or LiDAR) suffers from blind spots and occlusions due to its limited field of view and detection range. This results in spatially incomplete perceptual information, making it impossible to achieve a globally consistent scene understanding during **Data Modeling**.
3. **Interaction Complexity**. An urban environment is beyond a collection of physical spaces but a complex social environment composed of numerous intelligent agents. The behavior of these agents is not simple physical motion but is driven by a multi-layered set of rules. Their behavior follows both explicit rules (e.g., traffic laws) and implicit social norms (e.g., driving habits, pedes-

 doi:10.20944/preprints202601.2155.v2

trian etiquette, intentions signaled through body language). Understanding these interactions requires interpreting subtle cues like posture, gaze, and intent, which are far harder to capture and model during the **Data Modeling** and **Task Application** stages than simple physical motion.

These challenges highlight the complexity of the data lifecycle in urban environments, making it difficult for single-domain data to comprehensively capture the scene. To address these limitations, utilizing multidomain data for UrbanEA has emerged as a promising strategy [13,17,18]. By integrating perspectives from diverse domains—such as a vehicle's terrain perception, a drone's aerial view, and a satellite's global map—the system can mitigate the limitations of individual sensors and reduce blind spots to construct a more spatially complete scene understanding.

However, as shown in Table 1, existing reviews tend to adopt a task-centric or model-centric perspective, such as foundation models [12,19] or the graph neural network [10], which may not fully align with the requirements of this multidomain approach. Specifically, urban computing surveys primarily focus on digital agents designed for passive analytical tasks, such as traffic and weather prediction. In these frameworks, data is often treated as static inputs for offline reasoning, rather than the interactive streams required by physical agents to actively perceive and intervene in the world. Meanwhile, although existing embodied agent surveys discuss physical interaction, they generally focus on controlled indoor environments, leaving the dynamic gaps inherent in complex, city-scale environments largely unaddressed. Therefore, a systematic review covering the end-to-end lifecycle, from data perception to final application, for UrbanEA is still needed. To fill this gap, we present the first comprehensive survey on the data lifecycle for UrbanEA. Our research focuses on the entire pipeline, investigating how to efficiently store, query, and fuse this complex data to support downstream embodied agent applications. Our contributions are summarized as follows:

1) **Unified Data Lifecycle Framework**. We propose a systematic framework that organizes the existing UrbanEA research into an end-to-end pipeline. Unlike existing surveys that focus on isolated tasks or specific domain data, our framework integrates Data Perception, Management, Modeling, and Application, providing a holistic view of how urban data flows from raw sensors to an embodied agent.

2) **Fine-grained Multi-stage Taxonomy**. We establish a taxonomy for each stage of the lifecycle to clarify technical boundaries. Specifically, we categorize perception by modalities and domains, classify management by storage structure and capability, and taxonomy modeling strategies based on the specific domain gaps in the real world.

3) **Forward-looking Research Roadmap**. We identify cross-stage challenges and synthesize them into a strategic roadmap spanning Method-level, System-level, and Societal-level dimensions. By extending the discussion beyond technical metrics to broader social impacts, we aim to provide insights that may inspire future research in the UrbanEA community.

Table 1. Comparison with existing surveys for the model-centric urban computing survey and indoor embodied agent survey.

Survey	Year	Venue	City Platform	Multidomain Data	Embodied Agent	Data Life-cycle	Primary Perspective
Jin et al. [10]	2023	IEEE TKDE	✓				Model-centric
Rahmaniet al. [11]	2023	IEEE TITS	✓				Model-centric
Yang et al. [3]	2024	IEEE TKDE	✓	✓			Model-centric
Zhang et al. [12]	2024	ACM KDD	✓	✓			Model-centric
Cengiz et al. [4]	2025	Information Fusion	✓	✓			Model-centric
Zou et al. [13]	2025	Information Fusion	✓	✓			Model-centric
Song et al. [14]	2025	IEEE TKDE		✓			Model-centric
Yao et al. [15]	2025	IEEE TITS	✓				Model-centric
Mao et al. [16]	2025	IEEE TITS	✓				Data-centric
Our Work	-	-	✓	✓	✓	✓	Data-centric

 doi:10.20944/preprints202601.2155.v2

The rest of the survey is organized as follows: Sec. 2 reviews the urban sensing and simulation in data perception. Sec. 3 discusses the pipeline for data management. Sec. 4 surveys data modeling techniques to bridge key multidomain data gaps. Sec. 5 presents downstream task applications, and Sec. 6 explores the broader social impacts. Sec. 7 discusses future outlook and open challenges. Sec. 8 finally concludes the paper.

2. Data Perception

2.1. How to Perceive the City?

2.1.1. Vision Perception

For the visual sense ability, we divide it into the following main categories [20]. The multidomain data captures distinct aspects of the environment, often complementing each other. We visualize the vision perception in the terrain domain as the example, as shown in Figure 2.

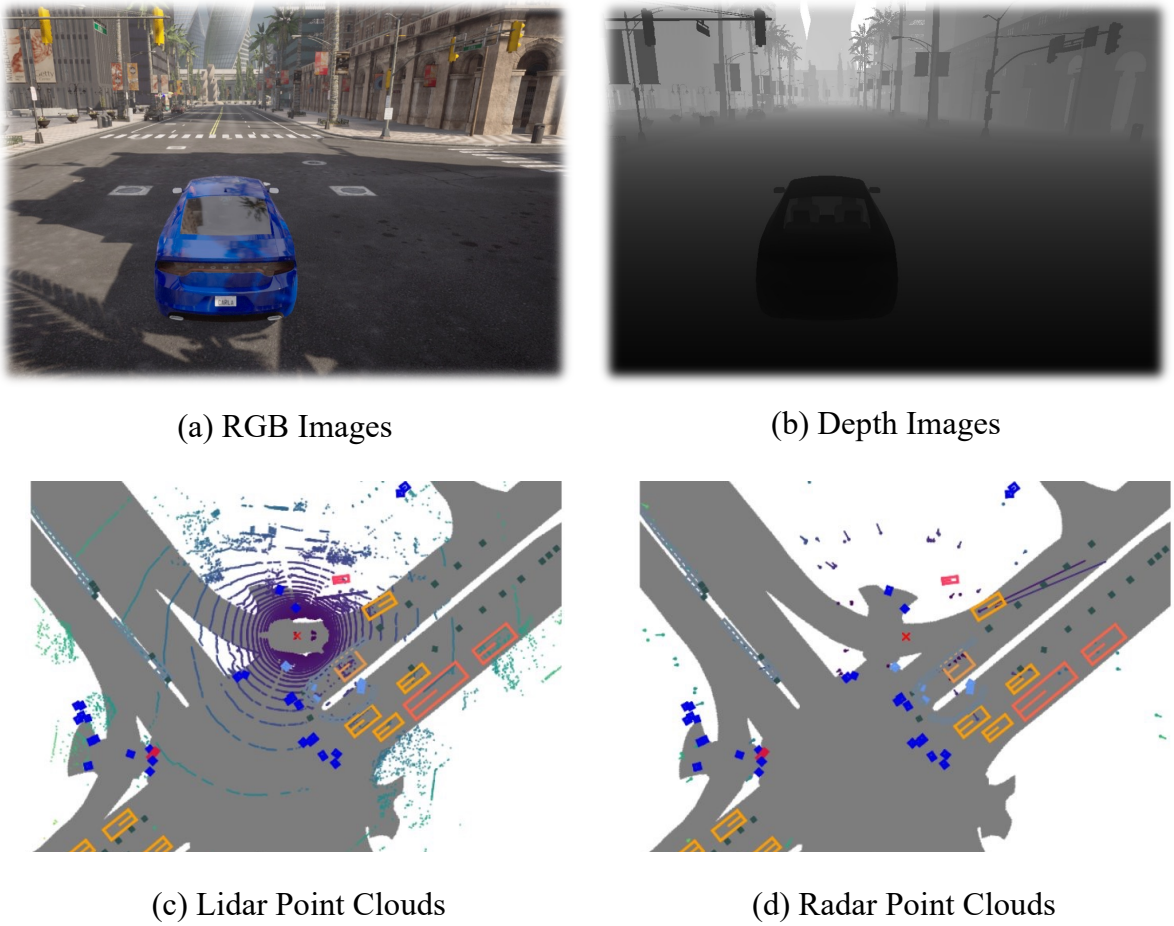
• **RGB Images**: These are standard color images, akin to what a human eye or a typical camera perceives. They are rich in texture, color, and semantic information, making them invaluable for tasks like object recognition, classification, and scene understanding (e.g., identifying road signs in autonomous driving, describing visual elements in spatial description) [21]. An RGB image can be represented as a three-dimensional tensor, denoted as $I_{RGB} \in \mathbb{Z}_{256}^{H\times W\times 3}$. Its dimensions are the image height H, width W, and 3 color channels (Red, Green, Blue). However, they are 2D projections and are sensitive to lighting conditions, lacking direct information about the 3D structure or distance to objects.

• **Depth Images**: Unlike RGB images that capture color, depth images encode distance information. Each pixel value typically represents the distance from the sensor to the corresponding point in the scene. A depth image can be represented as a 2D matrix, denoted as $I_{Depth} \in \mathbb{R}_{+}^{H\times W}$. Its dimensions are the image height H and width W. The value at each pixel (u, v) is a scalar representing the distance from the sensor to that point in the scene. This provides explicit geometric cues crucial for obstacle avoidance, navigation, and 3D reconstruction, often used in drone operation and robot tasks [22].

• **Lidar Point Clouds**: Light Detection and Ranging (Lidar) sensors actively emit laser beams and measure the reflected light to create a sparse but accurate 3D map of the surroundings. A Lidar point cloud is an unordered set of points, denoted as $\mathcal{P}_{LiDAR} = \{p_1, p_2, ..., p_N\}, p_i = (x_i, y_i, z_i, i_i) \in \mathbb{R}^4$. It consists of N points, where N is variable. Each point p_i contains at least its coordinates (x, y, z) in 3D space. It often includes reflection intensity, i, as well. These point clouds provide precise geometric structure and distance measurements over considerable ranges, largely independent of ambient light. While excellent for geometry, they typically lack the rich color and texture information found in RGB images.

• **Radar Point Clouds**: Radar sensors use radio waves instead of light. Radar data is also a set of points, denoted as $P_{Radar} = \{d_1, d_2, ..., d_M\}, d_j = (x_j, y_j, z_j, v_{x_j}, v_{y_j}, v_{z_j}) \in \mathbb{R}^6$. It consists of M detections, where M is typically much smaller than the number of Lidar points, N. (v_x, v_y, v_z) is the velocity vector of the detection. Similar to Lidar, they can generate point clouds representing detected objects. Radar's key advantage is its robustness in adverse weather conditions (rain, fog, snow), where Lidar and cameras might struggle. However, Radar typically provides lower resolution and less detailed shape information compared to Lidar or cameras [15,23].

 doi:10.20944/preprints202601.2155.v2



(a) RGB Images (b) Depth Images

(c) Lidar Point Clouds (d) Radar Point Clouds

Figure 2. Vision perception for Urban Embodied Agents. We visualize some cases in Carla Simulators [24].

2.1.2. Multi-Sensory Perception

In urban settings, multi-sensory technology, which integrates auditory (e.g., traffic sounds, water features), tactile (e.g., pavement textures, wind), even olfactory (e.g., floral scents, garbage odor), and thermal (e.g., temperature), is being applied to smart city fields like environmental monitoring [25] and public security [26,27]. Complementing these environmental modalities, kinematic and localization sensors, such as Inertial Measurement Units (IMU) and Global Positioning Systems (GPS), provide critical proprioceptive states and geospatial contexts, enabling agents to anchor their perceptions within the dynamic urban frame.

Sensing within environments is multi-sensory, extending well beyond the visual sensing [28]. To this end, efforts focused on combining audio and visual information, with various works aiming to train agents that can both see and hear by using integrated audio-visual simulations [29–33]. The domain of visual-tactile learning focuses on building realistic tactile simulation systems to allow agents to understand the world through physical interaction [34–38]. Multiply [39] proposes a multi-sensory sensing simulator. This platform incorporates a wide array of interactive data—including visual, audio, tactile, and thermal information—directly into large language models, thereby establishing a direct and powerful correlation among words, actions, and percepts.

2.2. Where to Perceive the City?

The urban perception requires a system capable of synergistically processing information from different observational dimensions [13,40,41]. We divide the urban perception based on the data domain, including handheld, vehicle, drone, plane, and satellite as shown in Figure 3. These platforms demonstrate spatio-temporal heterogeneity, resulting in a dynamic gap as discussed in Section 4.1.3. These platforms constitute the sensing infrastructure of urban computing systems. UrbanEA lever-

 doi:10.20944/preprints202601.2155.v2

ages these classic urban data sources (e.g., street view imagery, remote sensing) but requires higher frequency and lower latency to support embodied control.

• **Handheld.** The data from handheld devices is designed for close-quarters mapping and typically has a shorter range. For example, some professional handheld scanners have a flexible scanning range from 0.4 to 10 meters, making them ideal for detailed exterior facade work [42]. The terrain handheld scanners can achieve accuracies around 5-10 mm. The use of handheld devices for data acquisition suffers from low efficiency, limited coverage, and data inconsistencies caused by manual handling, as the sim-to-real fidelity gap discussed in Sec 4.1.2.

• **Vehicle.** These systems are designed for efficient corridor mapping and possess a range optimized for capturing roadside features from a moving vehicle [43,44]. The vehicle systems experience a slight reduction in accuracy compared to handheld static scanners, but they still deliver exceptional results suitable for most urban mapping tasks, capturing high-density data within 30 meters to 100 meters of the vehicle's path. In urban environments, tall buildings lead to GPS signal drift, while pedestrians and other vehicles create dynamic obstructions. These deficiencies may introduce the dynamic gap as discussed in Sec 4.1.3.

• **Drone.** Drones operate in a unique low-altitude domain, which allows them to achieve exceptionally high spatial resolutions with both photogrammetric and LiDAR sensors. For most professional urban mapping applications, drones can easily achieve a Ground Sample Distance (GSD) between 1 cm and 5 cm per pixel [45]. While its high flexibility is an advantage, it also causes variations in scale and perspective, posing a challenge for precise camera pose estimation.

• **Plane.** For urban mapping projects, the aerial plane typically delivers a GSD in the range of 5 cm to 30 cm. A GSD of 5-15 cm is sufficient for creating highly detailed and geometrically accurate city-wide 3D models at LOD2 (differentiated roof structures) and LOD3 (architectural models with major facade elements). At this resolution, it is possible to clearly identify individual buildings, roads, vegetation, and major infrastructure elements [46]. However, it is difficult to capture fine terrain-level details.

• **Satellite.** Satellites operate from low Earth orbit at altitudes that dwarf aerial platforms, yet technological advancements have enabled them to achieve remarkable spatial resolutions [47,48]. The commercial constellations offer panchromatic imagery with a native spatial resolution of approximately 30 cm. Although it provides wide coverage, it suffers from the spatio-temporal resolution with cloud-based occlusion, which may introduce the viewpoint gap.

	Sensing Range	Spatial Resolution	Temporal Resolution	Typical Tasks
Handheld	0.4m-10m	**Accuracy:** <1cm	Real-time to second-level	Detailed facade modeling, Crack detection
Vehicle	1m-100m	**Accuracy:**2cm–5cm	Real-time to minute-level	Street mapping, Road asset identification
Drone	50m-100m	**GSD:**1cm–5cm/px	Minutes to hours	Local 3D reconstruction, Infra inspection
Plane	1km-5km	**GSD:**5cm–30cm/px	Hours to days/years	City-level 3D modeling, road net mapping
Satellite	>10km	**GSD:**30cm–50cm/px	Seconds to weeks	Land use classification, Urban sprawl

Figure 3. Comparison between multidomain data in Urban Embodied Agents.

2.3. City Scene Simulators

The development of robust and reliable perception for outdoor environments relies on the simulation environments when the Internet agent is towards an embodied agent [49]. Existing indoor simulators [50–53] collect data from the handheld camera or scan sensors. Compared to indoor settings where controlled experiments can be collected from handheld cameras or scanner sensors, outdoor

 doi:10.20944/preprints202601.2155.v2

environments present challenges for real-world experimentation due to their complexity, dynamic nature, and safety concerns [54]. We classify the city simulators based on the perceptual capabilities an agent requires to move from observation to action:

Table 2. Comparison with existing Urban Embodied Agent simulators.

Environment	Year	Kinematics	Platform	Category	Modality				Data Source	Engine
					RGB	Depth	Radar	Lidar		
Cityscapes [55]	2016	✗	Terrain	Open-Loop	✓	✗	✗	✗	Street View	-
CARLA [24]	2017	✓	Terrain	Closed-Loop	✓	✓	✓	✓	Vehicle	UE 4
xView [56]	2018	✗	Aviation	Open-Loop	✓	✗	✗	✗	Satellite	-
TouchDown [57]	2019	✗	Terrain	Open-Loop	✓	✗	✗	✗	Street View	-
Nuscenes [58]	2020	✗	Terrain	Open-Loop	✓	✗	✓	✓	Vehicle	Nuscenes-Kit
Waymo [59]	2020	✗	Terrain	Open-Loop	✓	✗	✓	✓	Vehicle	Waymax
KITTI-360 [60]	2022	✗	Terrain	Open-Loop	✓	✗	✗	✓	Vehicle	-
STPLS3D [61]	2022	✗	Aviation	Open-Loop	✗	✗	✗	✓	Drone	-
SensatUrban [62]	2022	✗	Aviation	Open-Loop	✗	✗	✗	✓	Drone	-
UrbanBIS [63]	2023	✗	Aviation	Open-Loop	✓	✗	✗	✓	Drone	-
AerialVLN [64]	2023	✓	Aviation	Open-Loop	✓	✓	✗	✗	Drone	UE 4
GRUTopia [65]	2024	✓	Terrain	Closed-Loop	✓	✗	✗	✗	Virtuality	Isaac Sim
OpenUAV [66]	2024	✓	Aviation	Open-Loop	✓	✓	✗	✗	Drone	UE 4
UnrealZoo [67]	2024	✓	Terrain	Closed-Loop	✓	✗	✗	✗	Virtuality	UE 4/5
MetaUrban [68]	2025	✓	Terrain	Closed-Loop	✓	✓	✗	✓	Virtuality	Gym
OpenFly [69]	2025	✓	Aviation	Closed-Loop	✓	✓	✗	✓	Drone	UE 4, Google Earth, GTA V

2.3.1. Open-Loop Simulator

In this category, simulators function as replay platforms for real-world data logs. Their primary role is to evaluate the sensing system. These evaluations range in complexity, from semantic understanding that answers the question, *"What is it?"* Open-Loop Simulator can be divided into unimodal simulators and multimodal simulators. Unimodal simulators represent the foundational layer of virtual environment design, focusing on generating data for a single sensor modality. The goal is to train and validate specific sensing algorithms in a controlled manner for terrain and aerial simulation domains, such as StreetLearn [70], Cityscapes [55], xView [56], SensatUrban [62], UrbanBIS [63], and STPLS3D [61]. The multimodal simulator involves the integration of multiple, synchronized sensor streams. This is designed to replicate the comprehensive sensor suites of outdoor vision sensing, allowing for the development and testing of algorithms that create a more robust and reliable world model by combining the strengths of different modalities, such as Nuscenes [58], Waymo [59], KITTI-360 [60], AerialVLN [64].

2.3.2. Closed-Loop Simulator

This paradigm completes the sensing-action cycle. By enabling an agent's behaviors to interact with the simulation world, these simulators are equipped to evaluate the agent's capability: guiding

 doi:10.20944/preprints202601.2155.v2

action based on sensing. They move beyond passive observation to address the interaction problem for an autonomous agent: *"How should I react to it?"* This simulator focuses on the active, full-stack validation, testing the entire sensing-to-action loop and allowing for the evaluation of complex behaviors [71]. For terrain-based systems, including GRUTopia [65], UnrealZoo [67], and MetaUrban [68], leverage powerful physics and rendering engines like Isaac Sim and UE4/5, moving beyond sensing tasks. These environments simulate complex and dynamic weather phenomena (rain, fog, snow), realistic diurnal cycles with changing illumination, and intricate multi-agent interactions. In aviation-based simulation, like OpenUAV [66], and OpenFly [69] facilitate intricate interactions with the environment. A key evolution lies in the fidelity of control. This allows for the simulation of smooth, physically plausible flight dynamics, enabling agents to perform complex maneuvers and precise navigation that mimic real-world dexterity.

2.4. Discussion

Simulators are essential tools for showing the urban perception environment and developing UrbanEAs. However, existing simulators suffer from the sim-to-real gap. This gap manifests in two key areas: physical realism, such as accurately modeling sensor noise or complex weather and lighting effects, and behavioral realism, which involves simulating unpredictable human behaviors like varied driving habits or pedestrian movements. The core challenge is how to effectively quantify and reduce this gap. One of the future directions is to establish a Real-to-Sim-to-Real reinforcement loop [54]. This means using high-fidelity data from the real world to build and continuously refine the next generation of simulators, which in turn can train more capable agents.

3. Data Management

UrbanEAs operate in a continuous loop of sensing, reasoning, and acting. As detailed in Section 2, this process generates a data deluge characterized by extreme heterogeneity (structured logs vs. unstructured point clouds), high velocity (100Hz+ IMU streams), and semantic ambiguity. Existing raw data streams are ill-suited for direct consumption by downstream tasks: data modeling algorithms (Section 4) fail when timestamps are misaligned; safety-critical planning (Section 5) falters without explicit object relationships; and long-term learning requires historical replay capabilities that transient streams cannot provide [93,94].

Unlike traditional autonomous driving or drone vision systems that primarily operate as passive data collectors, UrbanEAs introduce three properties that reshape data management requirements [95, 96]. First, *interactivity*: agents continuously interact with the environment and with humans, creating closed-loop data streams where actions alter future observations, demanding data systems that support replay and counterfactual reasoning beyond static retrieval [96]. Second, *sociality*: agents operate within human social contexts, requiring data representations that capture not only spatial but also social relationships such as cooperative intentions and normative constraints [97,98]. Third, *action-awareness*: agents must tightly couple sensory records with decision logs and motor commands to support planning, experience replay, and continual learning [99]. These properties mean that UrbanEA data management is not merely a scaling problem, but a qualitative shift in what must be stored, how it must be organized, and how it must be queried. The following subsections examine each layer in detail (with representative systems summarized in Table 3). Moreover, we illustrate these challenges using the running example in Figure 4: an ego-agent navigating a rainy intersection with occlusions, involving sensor noise (rain), multi-source asynchronous data (vehicle sensors vs. RSU), and social reasoning about right-of-way.

 doi:10.20944/preprints202601.2155.v2

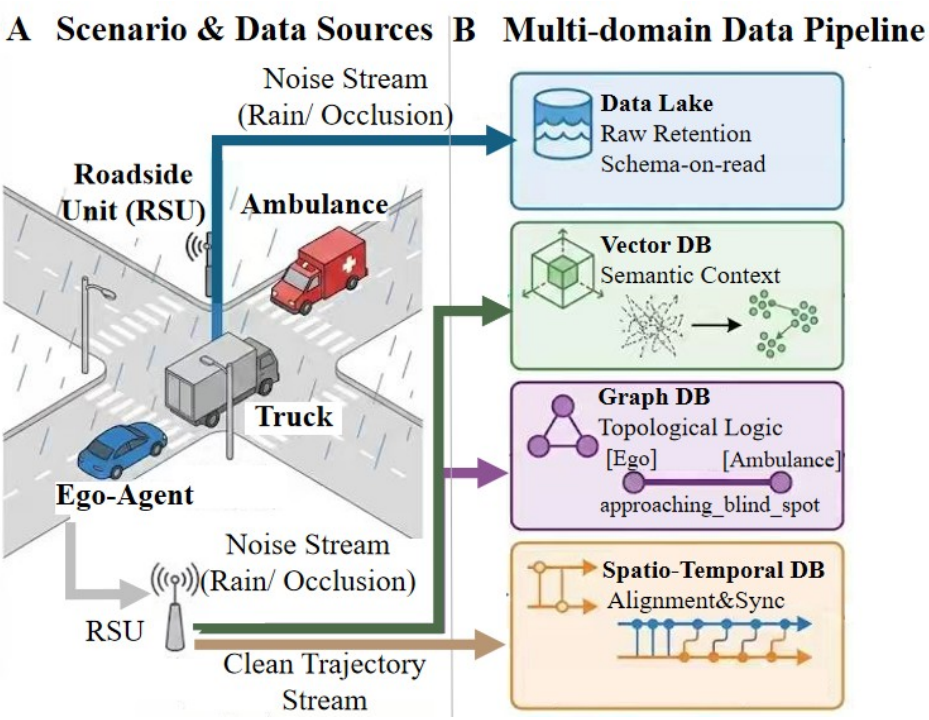


Figure 4. The example in the real-world scenario using DBs for the UrbanEAs.

doi:10.20944/preprints202601.2155.v2

Table 3. Comparative Evaluation of Storage Architectures for Urban Embodied Agents.

Architecture	Core Abstraction	Core Capabilities				Performance Profile			Typical Systems
		Heterogeneity	Relational Semantics	Semantic Search	Temporal Analysis	Read Latency	Write Throughput	Scalability	
Data Lakes	Raw Files	✓	✗	✗	✗	High	High	High	Lambda Arch. [72], Kappa Arch. [73], Lakehouse [74]
Multi-model DBs	Unified Model	✓	✓	✓	✓	Variable	Medium	Medium	Sinew [75], NoAM [76], UniBench [77]
Graph DBs	Nodes & Edges	✗	✓	✗	✗	Low	Low	High	Neo4j [78], JanusGraph [79], nSKG [80], Sg-CityU [81]
Vector DBs	High-dim Vectors	✗	✗	✓	✗	Low	Low	High	FAISS [82], Milvus [83], HNSW [84], PQ [85]
Time-Series DBs	Time/Value Pairs	✗	✗	✗	✓	Low	High	High	Gorilla [86], Apache IoTDB [87], TimescaleDB [88]
Spatio-Temporal DBs	Time/Spatial Info	✗	✓	✗	✓	Variable	Medium	High	MobilityDB [89], PostGIS [90], TrajMesa [91],TMan [92]

 doi:10.20944/preprints202601.2155.v2

3.1. From Static Datasets to Interactive Data Ecosystems

Traditional autonomous driving and drone systems rely on a publish-once paradigm: sensors record data during collection campaigns, which is then released as static datasets for offline training. UrbanEAs, by contrast, demand data ecosystems that continuously ingest, organize, and serve the heterogeneous data streams generated through ongoing agent-environment interaction. This subsection traces the evolution from static datasets to dynamic platforms capable of supporting interactive data lifecycles.

3.1.1. Data Lakes

The evolution of UrbanEA data management begins with existing autonomous simulators such as NuScenes [58] and Waymo Open Dataset [59], which establish the baseline practice of organizing multidomain streams (cameras, LiDAR, radar) via relational metadata with temporal indexing, global identifiers, and standardized coordinate transformations for cross-modal data integration.

However, while these datasets excel at providing synchronized data for offline training, they represent an *open-loop* paradigm: static, post-collection dissemination that cannot capture the interactive nature of UrbanEAs. When an embodied agent interacts with a pedestrian or coordinates with other agents, the resulting data is inherently closed-loop. The agent's actions alter the environment, which in turn shapes future observations. Static datasets cannot represent such bidirectional causality, nor can they store the social context (e.g., cooperative intentions, right-of-way negotiations) that governs these interactions.

This limitation motivated the development of Data Lake architectures [100,101], which adopt a schema-on-read philosophy: raw data is ingested in its native format without predefined schemas, with structure imposed only at query time [74,100]. This design shifts the paradigm from publish once to continuously sense and govern, accommodating the continuous, evolving data streams generated by interactive agents.

For UrbanEA, Data Lakes offer three strategic advantages: (1) *maximized raw data retention* with full lineage tracking, enabling reprocessing as perception models evolve; (2) *real-time streaming ingestion* supporting continuous closed-loop operation; (3) *multi-zone governance* enforcing quality standards from raw sensors through decision logs to social interaction records. In the intersection scenario, the data lake retains raw assets together with calibration metadata and the agent's decision history, preserving the complete interaction context for fusion and auditing.

Beyond batch-oriented data lakes, the shift toward live data ecosystems has spawned *dynamic map* platforms that continuously fuse crowdsourced observations into a shared, real-time world representation. LiveMap [102] demonstrates a crowdsourcing architecture where edge-computing nodes aggregate vehicle sensor uploads into a real-time dynamic map, achieving sub-second update latency for urban intersections. SIM-LDM [103] further standardizes this concept through a Local Dynamic Map (LDM) framework layered into static infrastructure, semi-static topology, transient objects, and highly dynamic agent states—a decomposition that naturally aligns with the multi-zone governance model of data lakes while supporting the real-time demands of interactive agents.

3.1.2. Multi-Model Databases

UrbanEAs generate data spanning fundamentally different structures, including structured metadata (relational), spatial and social relationships (graph), sensor streams (time-series), and increasingly, interaction logs and decision records. Multi-model databases [104,105] address this heterogeneity by supporting diverse data models within a unified system [106–108], enabling cross-model queries such as *"retrieve all agents (relational) that were spatially near (graph) the incident location during the last 5 minutes (time-series)"* as single atomic operations without cross-system coordination [109,110]. For UrbanEA, this unification is particularly valuable because interactive and socially situated scenarios inherently require simultaneous access to heterogeneous data: in the rainy intersection, a unified query can jointly

 doi:10.20944/preprints202601.2155.v2

express occlusion constraints, agent priority relations, and associated sensor slices within the relevant time window.

3.2. Representing Spatial and Social Semantics

A distinctive challenge for UrbanEA data management is the need to represent not only physical and spatial semantic, but also social semantics: the intentions, norms, and interactive dynamics that govern behavior in shared urban spaces [97,111]. Traditional autonomous driving systems largely ignore this social layer, representing scenes purely through spatial coordinates and object categories. UrbanEAs, by contrast, must reason about why a pedestrian hesitates at a crosswalk or why an ambulance has priority, demanding richer semantic representations in their underlying data substrates [95,112].

3.2.1. Graph Databases

Traditional data models struggle to natively express the complex relationships that govern urban systems [113–115], a limitation known as the semantic gap: raw data lacks the explicit relational context required for high-level reasoning [11,116]. Graph-based architectures bridge this gap by modeling entity relationships as graphs, where nodes represent entities and edges encode their relationships [98,117,118]. For UrbanEA, we argue that the standard scene graph formulation $G = (V, E)$ must be extended to a *socially augmented scene graph* that explicitly separates spatial and social relation types:

$$G_{\text{social}} = (V,\ E_{\text{spa}} \cup E_{\text{soc}},\ \mathcal{A}_V,\ \mathcal{A}_E) \tag{1}$$

where V is the set of entity nodes (vehicles, pedestrians, infrastructure elements), $E_{\text{spa}} \subseteq V \times V$ denotes spatial relation edges (e.g., *occludes*, *located_on_lane*, *in_front_of*), $E_{\text{soc}} \subseteq V \times V$ denotes social and normative relation edges (e.g., *yields_to*, *has_priority*, *cooperates_with*), and $\mathcal{A}_V : V \to \mathbb{R}^{d_v}$, $\mathcal{A}_E : E \to \mathbb{R}^{d_e}$ are attribute functions mapping nodes and edges to feature vectors encoding properties such as velocity, intent probability, and norm compliance. This formulation makes explicit a key distinction: E_{spa} edges are derivable from geometric perception, whereas E_{soc} edges require higher-level social reasoning—understanding that an ambulance has legal priority, or that a pedestrian's hesitation signals uncertainty about crossing intent. Figure 5 illustrates this distinction using our running intersection example.

 doi:10.20944/preprints202601.2155.v2

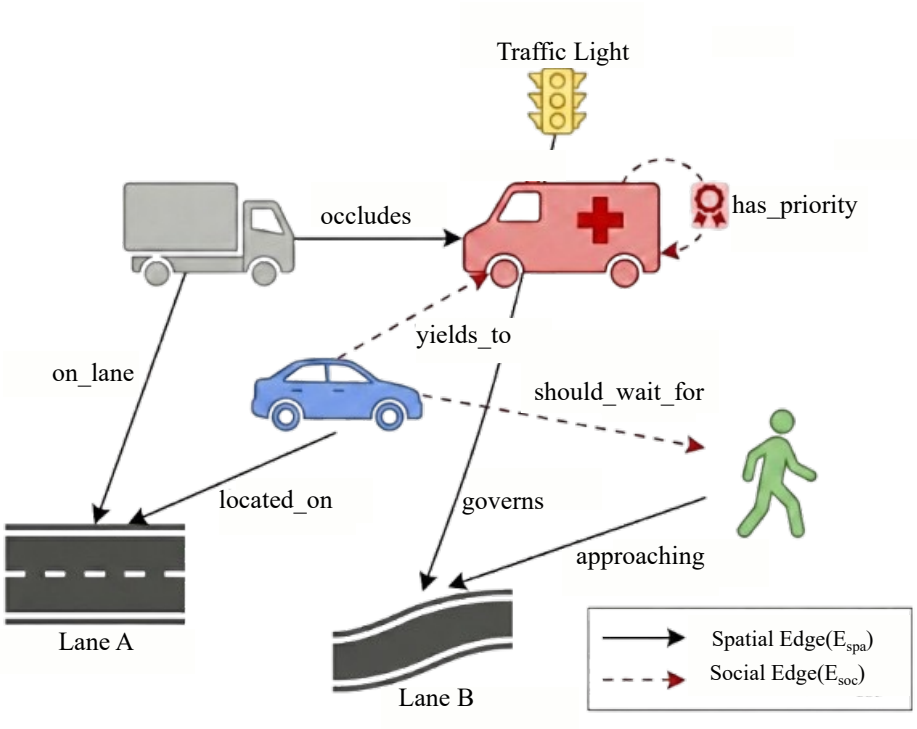


Figure 5. Socially augmented scene graph for the intersection scenario. Solid edges: spatial relations (E_{spa}); dashed red edges: social relations (E_{soc}).

Two complementary graph paradigms have emerged. Scene graphs follow a bottom-up approach, generated from sensor inputs to capture immediate context: Sg-CityU [81] decomposes 3D environments into object-centric representations, while T2SG [119] models road-level topology as lane connectivity graphs [81,119]. Knowledge graphs (KGs) adopt a top-down, ontology-driven approach [98,120,121]; nSKG [80] transforms nuScenes into ∼43 million RDF triples through formal ontology and systematic instance mapping. Importantly, recent KGs have begun encoding *social* semantics: the Connected Traffic Data Ontology (CTDO) [111] formalizes right-of-way rules, traffic norms, and vehicle connectivity as ontological relations, while Cities KG [112] maintains a city-scale dynamic geospatial KG with semantic 3D interfaces. This dichotomy reflects a trade-off: scene graphs prioritize speed for real-time tactical decisions, while KGs prioritize semantic richness for strategic reasoning [122].

Graph-based architectures also enable advanced reasoning through graph neural networks (GNNs) [123]; spatio-temporal GNNs jointly model spatial topology and temporal dynamics for tasks such as trajectory forecasting [10], as exemplified by SemanticFormer [124] which leverages nSKG's semantic *meta-paths* for multi-modal prediction. From a systems perspective, graph databases such as Neo4j [78] and JanusGraph [79] support multi-hop traversal queries via SPARQL and Cypher [98,113,118,120]. Recent work has further demonstrated that graph-structured world models can directly drive multi-agent planning. De Vos et al. [125] show that semantic graph world models can automatically configure multi-agent Model Predictive Controllers, where the graph encodes inter-agent constraints and coordination requirements. Holmberg et al. [126] propose a KG translation layer that converts high-level mission specifications into spatiotemporally grounded multi-agent path plans by reasoning over dynamic urban knowledge graphs. These results demonstrate that graph databases for UrbanEA are evolving from passive knowledge repositories into active computational substrates for planning and coordination.

3.2.2. Vector Databases

Relational models rely on exact, token-based queries, making them ill-suited for the vague, natural language queries that characterize human-agent interaction [92,127]. Vector databases address this

 doi:10.20944/preprints202601.2155.v2

by encoding multidomain data into high-dimensional vectors and leveraging Approximate Nearest Neighbor (ANN) search algorithms [82,84,85,128] for rapid, content-based semantic retrieval. The core operation measures semantic similarity between a query vector Q and data vectors D_i, often using cosine similarity [129]:

$$\text{sim}(Q, D_i) = \frac{Q \cdot D_i}{\|Q\|\|D_i\|} \tag{2}$$

The system returns items with the highest similarity scores [130]. In the intersection scenario, retrieving similar rainy intersection segments helps interpret degraded visuals when occlusion and glare lower confidence. This capability is essential for supporting Scene Question Answering (SQA) tasks (Section 5), where humans query agents in natural language, driving the adoption of vector databases for interactive visual scene retrieval [81,131–133]. Beyond standalone retrieval, vector databases have become a critical component of agent-centric memory systems through the Retrieval-Augmented Generation (RAG) paradigm. In this architecture, the agent encodes its accumulated experiences (observations, interaction episodes, spatial maps) into a vector store and retrieves relevant context at inference time to augment LLM-based reasoning [95,134]. This transforms vector databases from passive indexing systems into active cognitive components that shape how agents recall, reason, and plan—a theme we develop further in Section 3.4.

3.3. Synchronizing Perception, Decision, and Action

UrbanEAs operate in a tight perception-decision-action loop: sensory data informs decisions, decisions drive actions, and actions alter the environment to produce new sensory data [135]. Managing this temporal coupling requires data architectures that not only store high-frequency streams but also maintain the causal ordering between perception events, decision points, and action executions—a requirement that extends beyond the traditional time-series analytics inherited from IoT and monitoring applications.

3.3.1. Time-Series Databases

In UrbanEA, time-series databases serve two roles that reflect the distinct demands of embodied agents [136]. The first is managing high-frequency sensor streams. Vehicles, drones, and robots generate continuous data including IMU readings (100+ Hz), GPS trajectories (10 Hz), LiDAR scans (10–20 Hz), and vehicle dynamics (50+ Hz) [58,59]. Systems such as Gorilla [86], Apache IoTDB [87], and TimescaleDB [88] provide the write throughput, compression, and in-database analytics required for city-scale ingestion and real-time monitoring.

The second, and more distinctive for UrbanEA, is recording the complete action-perception loop of interactive agents. Unlike passive data collection in traditional autonomous driving, embodied agents continuously generate closed-loop interaction trajectories that must be logged as coherent time-series. We formalize such a trajectory as:

$$\tau = \left\{(s_t,\ a_t,\ o_{t+1},\ r_{t+1},\ c_t)\right\}_{t=0}^{T} \tag{3}$$

where s_t is the agent's internal state, a_t is the executed action, o_{t+1} is the resulting observation (causally influenced by a_t), r_{t+1} is the outcome reward or safety metric, and c_t encodes the social context at decision time (e.g., applicable traffic norms, inferred intentions of nearby agents). Traditional time-series databases store only observation sequences $\{o_t\}$; UrbanEA demands the full quintuple τ with causal links preserved, enabling action-conditioned queries such as *"retrieve all episodes where* a_t = *yield and* $r_{t+1} < \theta_{safe}$*"* for experience replay, debugging, and offline reinforcement learning. EmbodiedCity [137] exemplifies this by logging an agent's full trajectory, observations, actions, and rewards over entire interactive episodes. In multi-agent scenarios studied in DriveLM [133], synchronized time-series logs of all agents' states and decisions are crucial for understanding emergent social behaviors and training coordination strategies. In the intersection scenario, event-centric time windows around

 doi:10.20944/preprints202601.2155.v2

critical moments (e.g., urgent braking, yielding to the ambulance) organize ego-state streams together with the agent's decision history for fusion and audit.

3.3.2. Spatio-Temporal Databases

Spatio-temporal databases are critical for the simultaneous management of both spatial and temporal dimensions, a requirement that pure time-series databases do not fully address [138–141]. Hybrid systems integrating spatial and temporal extensions (e.g., PostgreSQL with PostGIS and Timescale) have demonstrated effective query performance for urban sensor data [142,143].

For UrbanEA, the primary value lies in managing the trajectories and mobility patterns of multiple interacting agents in shared urban spaces. Urban trajectories of vehicles, drones, and pedestrians are inherently spatio-temporal objects, and their interactions—yielding, following, avoiding—require joint spatial and temporal reasoning. Specialized systems like MobilityDB [89] provide native support for moving objects, enabling queries such as *"find all agents that passed through region R between time t_1 and t_2"* or *"compute the speed profile of agent A over the last hour"* [144]. More recently, the integration with foundation models for spatio-temporal data has enabled in-database prediction and real-time forecasting directly on data streams [19,145–147]. In the intersection scenario, alignment operators synchronize the RSU-detected ambulance trajectory with the ego timeline, producing per-timestamp bundles that couple sensor data with social context for downstream data modeling. Despite these capabilities, current spatio-temporal systems remain designed for passive analytics rather than active, closed-loop data management [96]; native support for action-conditioned queries and real-time cross-agent state synchronization remain open challenges [125,126].

3.4. Memory for UrbanEAs

The preceding subsections reviewed data architectures originally developed for passive data analytics that are now being adapted for UrbanEAs. However, a growing body of work argues that embodied agents require agent-centric memory architectures—integrated data substrates that unify the storage, retrieval, and reasoning capabilities that an agent needs across its entire lifecycle [96,99].

3.4.1. Multi-Layer Memory Stacks

Drawing inspiration from human cognitive architecture, several recent systems organize agent memory into hierarchically structured layers. Synthesizing these approaches, Figure 6 illustrates a unified multi-layer memory architecture for UrbanEA agents, organized along two primary dimensions: persistence and abstraction. This architecture facilitates a cognitive transition from raw, transient sensor data to persistent, high-level symbolic knowledge.

At the lowest level of abstraction and persistence is the sensory memory, which directly ingests raw data streams from multidomain urban sensors (e.g., cameras, depth sensors, LiDAR, radar, GPS, and IMUs). Because this sensory data experiences fast decay within milliseconds to seconds, the system employs attention filtering to extract only task-relevant features before passing the information upward. Following this is the working memory, acting as the real-time active processing hub. It maintains the current task context, tracks active interaction partners, and manages a short-term episodic buffer. This layer serves as a dynamic bridge between perception and action, continuously interacting with the agent's planner to execute immediate tasks.

Occupying the highest level of symbolic abstraction is the long-term memory, which is partitioned into three specialized sub-modules backed by dedicated database systems. First, Episodic memory records timestamped events and interaction traces, natively supported by time-series and spatio-temporal databases; Second, Semantic Memory encodes structural concepts and relational norms, managed by graph and multimodal databases; Third, Procedural memory archives learned action policies and skills, leveraging vector databases and data lakes. The information flow within this architecture operates in a dynamic loop: valuable interactions are consolidated upward into long-term memory, while relevant historical experiences and semantic rules are retrieved downward to augment the working memory, providing the necessary context for complex urban interactions.

 doi:10.20944/preprints202601.2155.v2

Several recent works validate the efficacy of this hierarchical design. For instance, Han et al. [95] apply this three-layer stack to LLM-powered urban agents, demonstrating how it allows agents to balance the speed of short-term reactive decisions with the depth of long-term strategic reasoning. Similarly, RoboMemory [99] instantiates this hierarchy for physical embodied systems. Crucially, it extends the episodic memory to record not only sensor observations but also the agent's decisions and their outcomes, enabling experience replay and lifelong learning—capabilities fundamentally absent from conventional databases. Finally, the Grounded Memory system [135] demonstrates that such multi-layer architectures can operate in real-time, maintaining coherent context across extended human-agent interaction sessions.

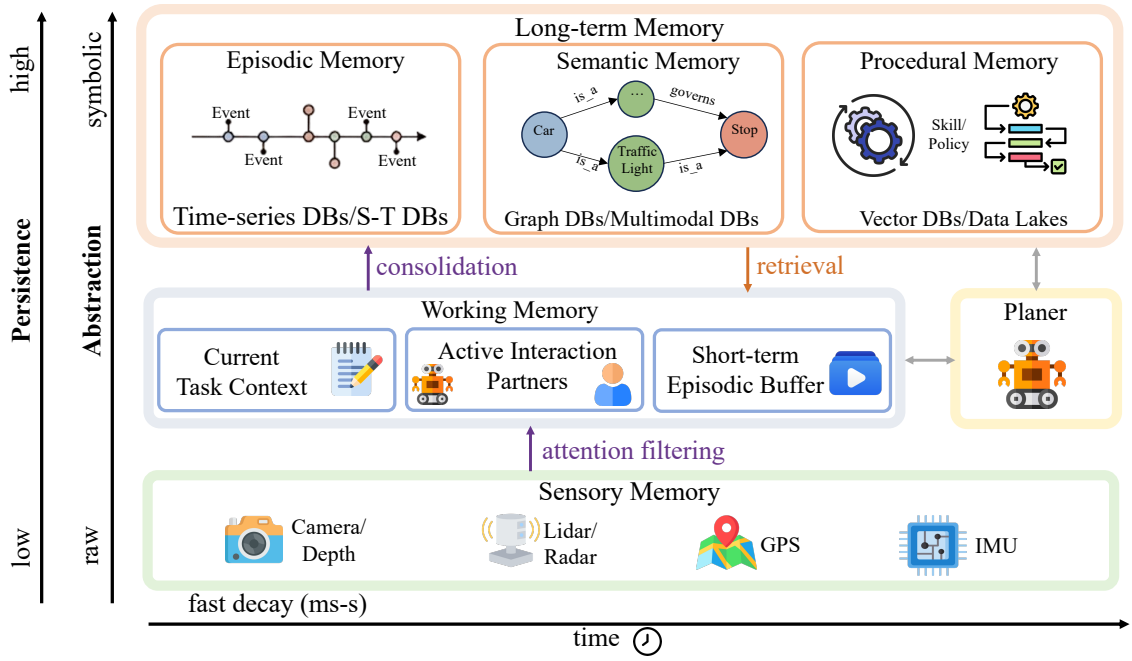


Figure 6. Multi-layer memory architecture for UrbanEA agents, with each long-term memory type backed by a specialized database system.

3.4.2. Embodied Retrieval-Augmented Generation

A complementary line of work extends Retrieval-Augmented Generation (RAG) from its origins in text-based QA to spatially grounded, embodied settings. Embodied-RAG [134] constructs a non-parametric memory as a hierarchical semantic forest—a multi-resolution spatial graph where nodes represent locations at varying granularities (rooms, corridors, specific objects) and are annotated with multimodal observations. When the agent receives a query, retrieval in such systems goes beyond the pure semantic similarity of conventional vector databases (Eq. 2) to incorporate spatial and temporal context. We characterize this as *context-aware embodied retrieval*:

$$m^* = \arg\max_{m \in \mathcal{M}} \Big[\underbrace{\alpha \cdot \mathrm{sim}(q,\, m)}_{\text{semantic}} + \underbrace{\beta \cdot \mathrm{prox}(\ell_{\text{agent}},\, \ell_m)}_{\text{spatial}} + \underbrace{\gamma \cdot \mathrm{rec}(t,\, t_m)}_{\text{temporal}} \Big] \tag{4}$$

where $\mathrm{sim}(q, m)$ measures semantic similarity between query q and memory node m, $\mathrm{prox}(\ell_{\text{agent}}, \ell_m)$ captures spatial proximity to the memory's grounded location, $\mathrm{rec}(t, t_m)$ favors recently observed memories, and α, β, γ are task-dependent weights. This formulation captures the shift from disembodied retrieval (only $\alpha \neq 0$) to embodied retrieval jointly conditioned on *where*, *when*, and *what*—effectively turning accumulated exploration into a queryable spatial database with explicit spatial structure.

 doi:10.20944/preprints202601.2155.v2

Mind Palace [148] extends this to long-horizon active question answering, maintaining a persistent spatial memory of explored environments and reasoning over it to plan where to explore next—demonstrating how agent memory can drive not just retrieval but also planning and action.

3.4.3. From Databases to Agent Memory

These agent-centric architectures share a common insight: the data substrates for embodied agents must go beyond the store-and-query paradigm of traditional databases to support *active memory management*—deciding what to remember, what to forget, and how to organize accumulated experience for efficient future use [96]. Table 4 highlights the key distinctions. For UrbanEA specifically, the convergence of these agent memory systems with the urban data architectures reviewed earlier suggests a natural integration path: data lakes and spatio-temporal databases provide the scalable storage backbone, graph databases encode the relational and social semantics, vector databases enable semantic retrieval, and agent memory frameworks orchestrate these components into a coherent cognitive substrate [95]. Realizing this integration remains an open challenge that we discuss further below.

Table 4. Comparison of traditional database approaches and agent-centric memory architectures for UrbanEA.

Dimension	Traditional DBs	Agent Memory
Data lifecycle	Store & query	Remember, forget, consolidate
Action coupling	Passive recording	Closed-loop logging
Retrieval mode	Exact / ANN search	Context-aware, goal-driven
Spatial grounding	Coordinate indexing	Hierarchical semantic maps
Social context	Absent	Norms, intentions, interaction history

3.5. Discussion

This section has surveyed the data management architectures underpinning UrbanEA, from storage systems inherited from traditional urban computing to emerging agent-centric memory substrates. Our central argument is that UrbanEA data management is shaped by three distinctive properties—interactivity, sociality, and action-awareness—that demand a qualitative shift beyond passive, analytics-oriented databases toward active, closed-loop data systems capable of logging actions as first-class data, encoding social semantics, and coordinating state across multiple agents [95,96,99]. Looking forward, we identify two converging trends. The first is an evolution from static datasets to dynamic data platforms that support continuous upload, annotation, querying, and simulation, moving toward living urban data ecosystems [50,97,102]. The second is the concept of data as a queryable world model: as agent memory architectures mature [134,148], the managed data asset itself becomes an implicit, inferable world model that blurs the boundaries between raw data, structured information, and actionable knowledge [40,149].

4. Data Modeling

Data in UrbanEA involves more than a passive feed from a single domain. It requires active perception to construct a coherent u urban environment model. This process is essential, not merely as the assembly of puzzle pieces for visual completeness, but to instantiate an actionable cognitive substrate that resolves inconsistencies between sources and fills the blind spots of individual sensors. Only through this process can a scattered collection of data become a logical and navigable urban tapestry capable of supporting agent interaction. Without effective modeling, the UrbanEA is restricted to disjointed and potentially contradictory perceptual cues, limiting its ability to plan and act safely [150].

However, constructing such a unified environmental representation presents challenges stemming from the heterogeneity of multidomain data [13,14]. While the previous section detailed the preparation of individual data streams, this section explores how to integrate them into a unified understanding [151]. As shown in Tab 5, we begin by analyzing the four gaps that influences this

 doi:10.20944/preprints202601.2155.v2

integration: at the data level, the Viewpoint and Scale Gap and the Sim-to-Real Fidelity Gap; at the spatio-temporal dimension, the Dynamics and Causality Gap; and at the cognitive level, the Geometric-Semantic Gap [152], as illustrated in Figure 7. After defining these challenges, we introduce the primary modeling strategies—Raw-data-level, Hierarchical-Feature, and Decision-level Integration—illustrated with state-of-the-art research examples [153].

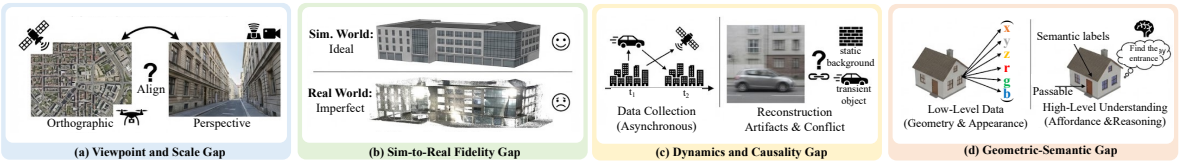


Figure 7. The domain gap between multidomain data for Urban Embodied Agents.

4.1. Multidomain Data Domain Gap for UrbanEA

Achieving comprehensive perception for UrbanEAs requires overcoming fundamental disparities between heterogeneous data sources [154]. Unlike general multi-modal fusion, urban environments present unique challenges rooted in the massive scale difference, complex dynamics, and the requirement for physical interaction. We categorize these challenges into four gaps: the Viewpoint and Scale Gap, the Sim-to-Real Fidelity Gap, the Dynamics and Causality Gap, and the Geometric-Semantic Gap [152], as illustrated in Figure 7.

Table 5. Comparison of different data integration methods. Cost, Semantic, Consist. and Robust. denote the computational cost, the semantic richness, spatial consistency, robustness to missing data, respectively.

Integration Strategy	Targeted Domain Gap	Cost	Semantic	Consist.	Robust.
Raw-data-level	Sim-to-Real Fidelity	High	Low	High	Low
Hierarchical-Feature	Dynamics & Causality Geometric-Semantic	Medium	High	Medium	Medium
Decision-level	Viewpoint & Scale	Low	Medium	Low	High

4.1.1. Viewpoint and Scale Gap

This gap arises from the disparity in perspective and resolution between data sources. Satellite and aerial data typically provide orthographic, top-down views covering large scales, whereas street-level sensors (cameras, LiDARs) capture perspective, ego-centric views with high local density. Aligning these domains is non-trivial due to the lack of overlapping visual features—for instance, a building's roof seen from space looks completely different from its facade seen from the street [155]. Furthermore, this creates a structural incompatibility: satellite imagery exists as regular 2D grids, while ego-centric data often comes as sparse, unordered 3D point clouds or continuous implicit fields [156]. Bridging this gap requires specialized algorithms to handle coordinate transformation and feature matching across vastly different scales [157].

4.1.2. Sim-to-Real Fidelity Gap

This gap describes the disparity between the idealized data often used in simulation and the imperfect, noisy data encountered in the physical world [158]. While UrbanEAs are frequently trained in clean synthetic environments, real-world deployment faces the Sim-to-Real challenge. For example, point clouds from real-world scans contain noise, holes due to occlusion, and uneven density, unlike the complete meshes in simulators [159,160]. Similarly, visual data degrades under complex urban illumination (e.g., overexposure, shadows, reflections on glass buildings), which are often absent or simplified in synthetic training data [161–163]. Overcoming this fidelity gap is crucial for ensuring that agents trained in simulation can robustly perceive and act in the wild [164–166].

 doi:10.20944/preprints202601.2155.v2

4.1.3. Dynamics and Causality Gap

The dynamics gap refers to the difficulty of maintaining a consistent world model in a constantly changing urban environment. Unlike static datasets, urban scenes are inherently dynamic. Temporally, data sources are rarely synchronized; a satellite image might be months old, while street views are real-time. This leads to inconsistencies where infrastructure present in the map might be under construction in the agent's view [167,168]. Spatially, the presence of moving agents (pedestrians, vehicles) creates ghosting artifacts in world modeling and conflicts in occupancy mapping. Data modeling algorithms must distinguish between the static city background and transient dynamic objects to prevent the agent from treating a moving car as a permanent wall [169–171].

4.1.4. Geometric-Semantic Gap

The semantic gap represents the cognitive disconnect between low-level sensory data (geometry and appearance) and high-level scene understanding required for decision-making [172]. While algorithms process pixel colors and 3D coordinates (x, y, z, r, g, b), UrbanEAs need to understand affordance not just identifying an object (e.g., *"a door"*), but understanding its function (e.g., *"passable"*) and state (e.g., *"open/closed"*). To bridge this gap, researchers integrate semantic priors (via segmentation or language models) and structured knowledge bases (like OpenStreetMap) into the integration pipeline [173,174]. This transforms a purely geometric map into a semantic representation, enabling the agent to perform complex reasoning tasks such as *"find the entrance"* [175].

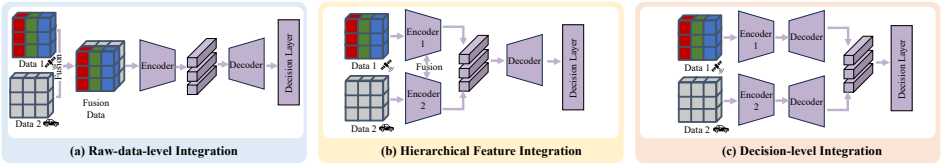


Figure 8. The different data integration strategy in the multidomain data for Urban Embodied Agents.

4.2. Multidomain Data Integration

To construct an actionable world model for UrbanEAs, integration strategies must bridge the domain gaps to ensure physical consistency and semantic understanding. These strategies can be classified into raw-data-level, feature-level, and decision-level Integration based on the stage in the pipeline [176–178].

4.2.1. Raw-Data-level Integration

In raw-data-level integration, heterogeneous data streams are merged at the input stage to construct a high-fidelity digital twin, primarily addressing the sim-to-real fidelity gap. For instance, combining RGB images from satellites with depth maps from drones creates a unified RGB-D stream that provides both appearance and geometry. This is crucial for sensor simulation, as depth information compensates for texture-less regions, ensuring the synthesized environment is geometrically consistent for agent training. Muturi et al. [179] utilize language prompts to align these heterogeneous representations at the outset, ensuring that the fused data preserves task-relevant features. Furthermore, Shang et al. [180] integrate RGB images with depth priors to enhance few-shot sensor simulation. By improving the fidelity of novel view synthesis, they allow agents to be trained in realistic simulations that closely mimic physical world observations, thereby reducing the domain shift during real-world deployment.

4.2.2. Hierarchical-Feature Integration

Hierarchical-feature integration combines representations at various abstraction levels to address the dynamics and causality gap and the scale gap. Urban environments are dynamic. Thus, integration must distinguish between the static background (the map) and transient objects (dynamic agents). The SUDS model [181] employs a joint optimization framework to fuse multidomain inputs, effectively

 doi:10.20944/preprints202601.2155.v2

decoupling dynamic agents from the static city. This separation is vital for navigation, preventing agents from treating moving vehicles as permanent obstacles. To bridge the geometric-semantic gap, GAANet [182] aligns cross-domain features in a graph space. By modeling the city as a topological graph rather than just pixels, it captures the spatial relationships required for path planning. For city-scale scalability, VastGaussian [183] and CityGaussian [184] adopt a divide-and-conquer strategy. Crucially, they introduce mechanisms to separate stable geometric colors from transient lighting effects (e.g., shadows), ensuring that the agent's perception system remains robust against illumination variations during long-term operation.

4.2.3. Decision-Level Integration

Decision-level integration processes domains independently to form high-level predictions, focusing on bridging the viewpoint gap and ensuring semantic consistency for decision-making. This strategy is particularly effective for aligning global planning (coarse aerial data) with local control (fine terrain data). Horizon-GS [185,186] employs a coarse-to-fine strategy, using aerial data to establish a global geometric prior before refining local details with street-view images. Similarly, City3D [187] utilizes terrain height maps to guide the completion of aerial LiDAR point clouds. For an embodied agent, this is not merely about visual completion but about constructing watertight physical collision bounds (e.g., filling missing walls) to support physics-based interaction. CrossView-GS [188] leverages multi-branch architectures to learn cross-view priors, enabling the agent to localize itself by matching limited onboard observations with global satellite maps, effectively bridging the disparity in observational perspectives.

5. Task Application

Having established the challenges of multidomain data management in Section III and the modeling strategies in Section IV, this section introduces how to apply this processed data to UrbanEA tasks. Traditional urban tasks, such as Traffic Flow Prediction and Point-of-Interest (POI) Recommendation, primarily operate as digital Agents [189,190]. These systems typically process aggregated data from a global perspective to mine patterns or forecast trends within a digital space. Their output is generally limited to information or suggestions, without direct physical intervention. In contrast, UrbanEA are defined by their physical or virtual body and their ability to interact with the environment. They perceive the environment, reason about physical constraints, and execute actions that actively change their state or the environment.

We introduce UrbanEA tasks, including Urban Scene QA (SQA), Vision-Language Navigation (VLN), and Human-Agent Collaboration (HAC). These tasks need to bridge the four gaps. For instance, an agent cannot answer a complex question (SQA) without first bridging the Semantic gap between raw sensor data and human-level concepts. Similarly, a navigation agent (VLN) inherently fails if it cannot resolve the dynamic gap to form a coherent world model. This section will explore how each of these applications leverages structured and fused data to achieve sophisticated cognitive capabilities in complex urban environments.

5.1. Urban SQA

5.1.1. Definition

Urban SQA enables intelligent systems to answer queries about their spatial context, making it a key task for environmental interpretation [191,192]. The goal is to develop a model ($\mathcal{F}$) that takes a scene representation (S) and a query (Q) as input to produce a textual answer (T). Optionally, the model can also output spatial grounding (B) via bounding boxes to localize entities. The scene representation S includes a point cloud ($S^{(p)}$) or multi-view images ($S^{(m)}$), while a query Q can be text ($Q^{(t)}$) or an egocentric image ($Q^{(e)}$). The task can thus be expressed as:

$$(T, B) = \mathcal{F}(S, Q). \tag{5}$$

 doi:10.20944/preprints202601.2155.v2

Effectively performing SQA requires the model to fuse spatial understanding with data processing. The question in SQA can be divided into Situation (such as "*Can the agent reach the warehouse from the current position?*") and Non-Situation (such as "*How many buildings are in the scene?*"), based on whether the query includes the agent's situation.

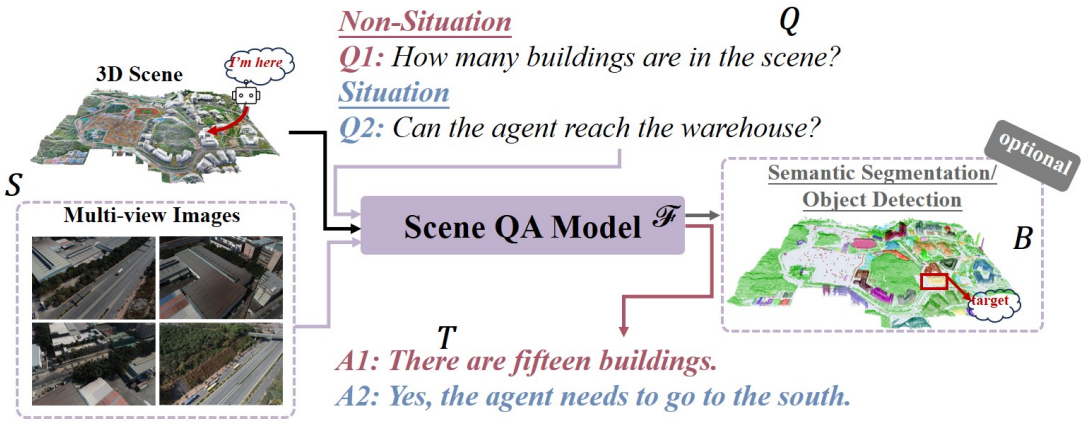


Figure 9. Definition of Urban Scene QA (SQA) for Urban Embodied Agents.

5.1.2. Classification

SQA can be categorized into two levels based on how the information required to answer a query is represented and acquired.

• **Passive SQA.** The first level, referred to as Passive SQA. In this task, the agent is a passive observer, relying on static information to answer questions without needing to explore or acquire new data through its own actions. It can be further divided into the following two sub-categories based on the difference in the scope of information: **Road-level QA** and **City-level QA**. The first is referred to as Road-level QA, which corresponds to road-level scene understanding, where the model analyzes a single, independent snapshot of the environment (e.g., a street-view image or a single frame of lidar data) to answer a question. Recent progress in autonomous driving research has stimulated the development of numerous SQA datasets designed to enhance road-level understanding capabilities, such as Nuscenes-QA [131], NuInstruct [193], NuPrompt [194], DriveLM [133], and VLAAD [132] and RoadSceneVQA [195]. After the data modeling, researchers employ cross-attention or the multi-layer perceptron mechanisms to achieve deep interaction between vision and language [131,133,194]. However, road-level QA focuses on instance-level queries and limited data in the roadside, which leads to an insufficient assessment of broader city scene comprehension and complex reasoning abilities.

To overcome the limitations of road-level Question Answering (QA), the field has progressed towards City-level QA, a paradigm that grants agents access to a comprehensive, prior model of an entire city for macroscopic spatial reasoning [81,196–199]. A challenge in this domain is maintaining information fidelity during the compression of vast urban data. To mitigate these challenges, researchers have primarily adopted two strategies. The first approach, hierarchical urban modeling, transforms vast and unstructured urban scenes into structured and queryable formats using a Relational Database or Graph Database, as discussed in Sec 3.2. For instance, GeoProg3D [199] and Sg-CityU [81] construct a structured tree or graph to abstract complex spatial information into objects and their interrelations, which simplifies spatial reasoning. SOBA [196] and EarthVQANet [200] involve using semantic segmentation to decompose large scenes into individual object units for analysis. The second strategy is to augment the compressed scene representation by integrating external information, such as using the geographic information based on the Spatio-Temporal Database, as discussed in Sec 3.2. This approach compensates for details lost during compression. OpenCity3D [198] and CityBench [197] align perception models with real Geographic Information Systems (like OpenStreetMap) to provide precise geographic coordinates and place names.

 doi:10.20944/preprints202601.2155.v2

• **Active Embodied QA.** To move beyond the cognitive bottlenecks of city-level QA, caused by the information loss and lack of dynamics in static models, Active Embodied QA (AEQA) shifts the agent from a passive analyst to an active explorer to gather high-fidelity, real-time information. Existing works includes EmbodiedCity [137] and CityEQA [201]. These datasets introduce tasks requiring an embodied agent to actively navigate and explore complex urban environments to answer open-vocabulary questions, assessing integrated navigation, sensing, and reasoning skills. The core advantage of AEQA is to transform an agent from a passive receiver of static information into an active explorer of the physical world, overcoming perceptual limitations through active action, and to locate and resolve ambiguity in complex environments. In road-level QA and city-level QA, a model is limited by the quality and viewpoint. For example, due to lighting effects, a black car appeared gray from a static viewpoint, leading to an incorrect judgment by agents. However, the agent in AEQA could actively adjust its observation pose by moving to the car's side, which reduced the impact of the lighting and allowed it to obtain the object's true attribute [201].

5.2. Vision-Language Navigation

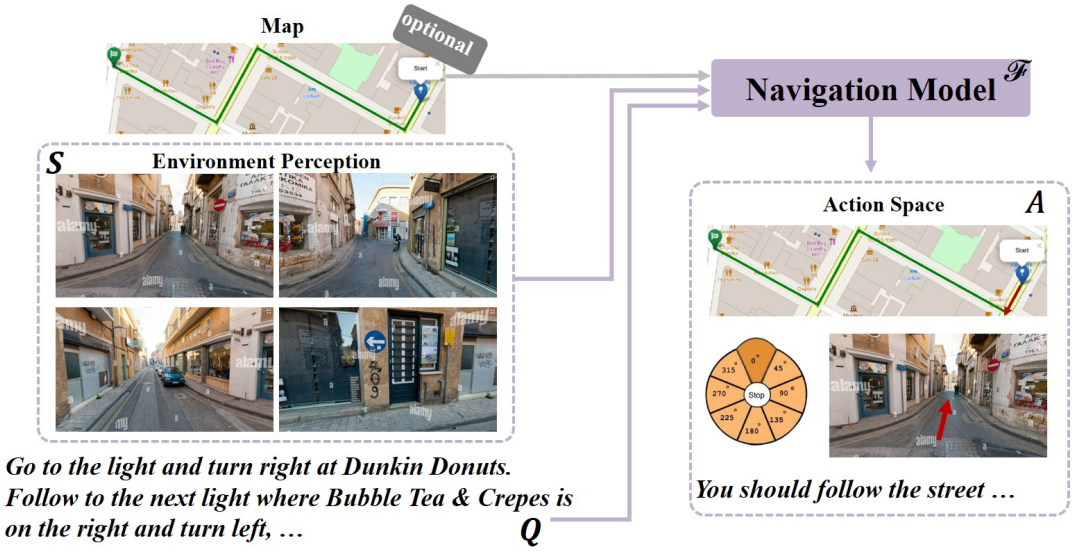


Figure 10. Definition of Vision-Language Navigation for Urban Embodied Agents.

5.2.1. Definition

Vision-and-Language Navigation (VLN) is a task requiring an embodied agent to navigate a realistic environment based on natural language instructions [202]. The core components involve the agent, the environment it perceives and acts within, and the language instruction guiding its movement. The VLN task necessitates a model or agent, denoted by $\mathcal{F}$, designed to process two inputs: the scene representation S perceived from the environment, and a series of natural language instructions $Q = \{q_1, q_2, ..., q_T\}$. The objective is to generate a sequence of actions $A = \{a_1, a_2, \ldots, a_T\}$ that directs the agent through the environment to fulfill the goal specified by the instruction Q. This process can be represented as the mapping:

$$A = \mathcal{F}(S, Q). \tag{6}$$

5.2.2. Classification

VLN tasks can be categorized based on their operational platform and observational perspective required of the agent $\mathcal{F}$. These tasks fall into two main paradigms: **Terrain-View Navigation** and **Aerial-View Navigation**.

• **Terrain-View Navigation.** It focuses on an agent's ability to perform navigation from a first-person, terrain-level perspective (such as that of a pedestrian or vehicle). They primarily utilize

 doi:10.20944/preprints202601.2155.v2

environments from static datasets, like street-view panoramas and predefined navigation graphs, to evaluate foundational skills like landmark recognition and path adherence. Datasets such as Touchdown [57] and map2seq [203] rely on simulated scenarios and fixed paths, which limit the agent's adaptive decision-making in complex, unpredictable real-world situations. Due to the semantic gap between instructions and first-person perception, researchers explore introducing auxiliary semantic data in vector databases and route and navigation data in time-series format to mitigate this gap. Research efforts expand training sets by automatically generating actions and instructions from unlabeled videos (e.g., VLN-video [204]), or by leveraging large language models to synthesize diverse auxiliary data, such as navigation rationales and landmark descriptions (e.g., FLAME [205], NavAgent [206]). Regarding multidomain fusion, strategies have evolved from early methods like direct feature vector concatenation (e.g., RconCAT [207]) and novel style-transfer fusion (e.g., VLN-Trans [208]), to utilizing large language models as a universal interface that converts all visual and historical information into natural language prompts for fusion and decision-making (e.g., Velma [206]).

• **Aerial-View Navigation.** It focuses on an agent (typically a drone) performing navigation using a third-person, top-down bird's-eye view. The core of this approach is leveraging external prior knowledge, such as geographic maps and top-down satellite imagery, to provide a global context for the agent's navigation plan. This Hybrid Spatio-Temporal Architecture directly tackles the pain point of navigating vast and unfamiliar environments, which is unavailable from a limited, first-person perspective. The primary challenge in this paradigm is aligning the provided language instructions to the specific spatial and visual features of the exo-viewpoint. These tasks includes AerialVLN [64], OpenUAV [209], CityNav [210], AutoFly [211] and AerialMind [212]. For instance, CityNav [210] incorporates an internal 2D spatial map representing landmarks mentioned in the instructions, which has been shown to markedly enhance navigation performance at a city scale. UAV-VLA [213] utilizes satellite imagery as a primary information domain for mission planning. In the image, the agent decomposes the high-level instruction and navigates to the destination. AVDN [214] is built within a simulator that uses top-down satellite images to represent the drone's visual observations. This gives both the agent and the human commander a bird's-eye view of the environment, simplifying the navigation challenge by providing inherent global context.

5.3. Human-Agent Collaboration

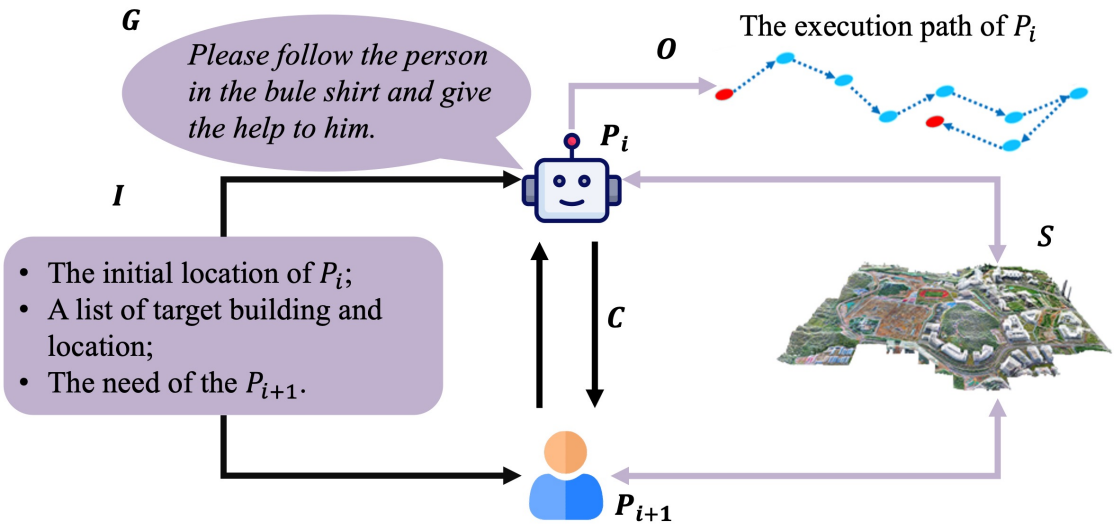


Figure 11. Definition of Human-Agent Collaboration for Urban Embodied Agents.

5.3.1. Definition

Human-Agent Collaboration (HAC) in urban environments enables a hybrid team of intelligent agents and human participants to work together towards a common purpose. The core task is to

 doi:10.20944/preprints202601.2155.v2

design a system function $\mathcal{F}$ that can process a collective goal (G), system-level input (I), and sensing information from the urban scene (S) to produce a coherent final output (O). This collaborative process is defined by the interaction of several key components. The system itself is composed of a heterogeneous set of n participants, $P = \{p_1, p_2, ..., p_n\}$, and the communication channels C. P includes both embodied agents and human participants. Each embodied agent operates based on its internal model m_i and goal g_i, while a human participant acts based on their expertise and assigned role r_i. The overall system task can be formally expressed as:

$$O = \mathcal{F}(G, S, I | P, C). \tag{7}$$

HAC necessitates the ability to decompose a high-level collective goal (G) into actionable sub-tasks and to coordinate the actions of both agent and human participants ($p_i \in P$) through structured communication channels (C) to achieve a unified result within the complex urban environment [215, 216]. In contrast to AI systems, Human-Agent Collaboration (HAC) leverages human strengths to build more efficient, robust, and trustworthy systems[217–219].

5.3.2. Classification

Urban Human-Agent Collaboration (HAC) tasks can be categorized based on their primary operational focus. There are two types of HAC tasks: Environment-centric HAC tasks, which aim to optimize the urban system, and Human-centric HAC tasks, which focus on assisting or interacting with individuals or groups within the city.

• **Environment-centric Human-Agent Collaboration.** In this category, the primary objective is to integrate AI agents into the urban fabric to enhance system-wide efficiency, planning, and management. Agents act as decision-support tools or autonomous managers for urban infrastructure distribution [68, 220–223]. These tasks often involve large-scale simulation, data analysis, and long-term optimization, with humans setting high-level goals and overseeing the system. For instance, some multi-agent frameworks try to employ a hierarchical architecture to process task data, effectively allocating responsibilities between robots and humans to balance safety and efficiency [224].

• **Human-centric Human-Agent Collaboration.** This category emphasizes the direct interaction and coordination between humans and agents to accomplish specific tasks within the urban environment and spatio-temporal data. Achieving this requires the data architecture capable of fusing inputs, ranging from human intent to real-time sensor streams, and leveraging specialized storage paradigms to support complex reasoning and action. The challenge in this task is bridging the gap between high-level, often ambiguous human intent and the low-level, concrete actions of agents. Try to address this challenges, the researchers try to build the agents with two capabilities: grounding human commands in the physical world and managing dynamic task execution over time. First, the agent can interpret the human's command by fusing the natural language instruction with both real-time environmental sensing and external world knowledge [225]. A command like *"find a safe place to land"* requires the system to semantically link the abstract concept of *'safety'* to visual and spatial features from its sensors. This process relies on a Vector Database to perform semantic search (connecting language to visual patterns) and Graph or Relational Databases to query structured world knowledge (e.g., GIS data identifying open areas, no-fly zones). This fusion of linguistic, perceptual, and knowledge-based data allows the agent to transform an abstract goal into a concrete, localized target [226–229].

5.4. Discussion

While urban computing excels at macroscopic optimization (e.g., traffic flow prediction), UrbanEA addresses microscopic execution (e.g., autonomous navigation through that traffic). The integration of these two fields promises a comprehensive solution for smart cities. The progression of tasks from passive scene understanding to active interaction reveals that future breakthroughs will depend on building more comprehensive world models. Future world models must learn universal principles, not just memorize patterns. This will allow a navigation agent trained on a US grid-style road

 doi:10.20944/preprints202601.2155.v2

network to successfully adapt to the unfamiliar roundabouts of Europe. Besides, the world models will be embedded with social intelligence. For tasks like HAC, future agents will learn to understand unwritten social norms (e.g., driving etiquette, personal space), moving beyond following traffic laws to become more intuitive and effective collaborators in society.

6. Social Impact

The urban environment is a socio-physical environment that integrates social behavior with physical scene [230]. This section aims to explore the profound potential of the UrbanEA to drive positive societal change, focusing on the social impacts it can deliver across the following domains.

6.0.1. Transportation

Modern urban transportation faces congestion and pollution due to the increasing number of vehicles, while complex road networks and unforeseen incidents continually threaten road safety [231–233]. Meanwhile, lagging public transportation planning leads to inefficient service, and the aging of critical infrastructure is difficult to maintain due to a lack of effective monitoring [234,235]. For instance, in the traffic flow management [236–238], the agent fuses real-time data from the terrain-level domain and the drone domain to perceive traffic flow dynamics, enabling adaptive traffic signal control to alleviate urban congestion, reduce carbon emissions, and improve the commuting experience [239].

6.0.2. Energy

In the context of the energy transition, despite the potential of clean energy sources like solar, planning for and maximizing their deployment in complex urban environments remains a challenge. Traditional energy management methods are macroscopic and static, incapable of performing fine-grained, dynamic management and optimization for energy-consuming units within a city [240–242]. For instance, the agent uses detailed 3D building models to calculate the solar conditions and shadow occlusions for every rooftop, generating a city-wide solar map [243–245].

6.0.3. Climate Change

The high-density buildings and surfaces of cities exacerbate problems like the urban heat island. The challenge for cities in adapting to climate change is the lack of tools capable of accurately simulating these risks. Macroscopic climate models are insufficient for guiding specific adaptations at the street and community levels. The agent fuses satellite, drone, and terrain data to accurately simulate the formation and distribution of the urban heat island effect [246]

6.0.4. Healthy Care

Factors within cities, such as air and noise pollution, the uneven distribution of green spaces, and pedestrian-unfriendly designs, invisibly harm public health and increase the risk of chronic diseases [247]. Through its multi-dimensional perception and analytical capabilities, the UrbanEA can quantify and visualize these health determinants, integrating public health goals into every aspect of urban planning. The agent integrates data from various urban sensors to generate dynamic, visual maps of air and noise pollution. This provides citizens with healthy route suggestions and helps environmental agencies pinpoint pollution sources, thereby improving the city's overall living environment quality [248,249].

7. Challenges and Future Directions

In this section, we summarize these challenges and suggest potentially feasible research directions, organizing them into methodological, systemic, and societal challenges.

 doi:10.20944/preprints202601.2155.v2

Table 6. Challenges and Future Directions for UrbanEA in Three Levels.

Level	Research Area	Challenge	Future Directions
Method	**Robust Modeling**	Imbalanced data (viewpoint, quality, temporal).	Fusion architectures for robust inference with missing data.
	Continual Evolution	One-shot understanding generating static snapshots.	Lifelong learning via memory consolidation.
System	**Sim-to-Real Loop**	Sim-to-Real Gaps in perceptual and action.	Real-to-Sim-to-Real loop via high-fidelity digital twins.
	Swarm Intelligence	Independent agents with only local sensing.	Scalable collaborative perception via state synchronization.
Societal	**Social Reasoning**	Physical sensing without abstract social knowledge.	Integrate social media, news, schedules, and weather alerts.
	Algorithmic Fairness	Data mirrors socio-economic disparities;	Transparent modeling to detect and mitigate data biases.

7.1. Method-Level Challenges

7.1.1. Robust Modeling

This is a challenge that directly extends the discussion of the sim-to-real fidelity gap in Section III. Future UrbanEA must operate reliably in the real world, where data domains, quality, and quantity are imbalanced. This imbalance manifests in several ways: viewpoint imbalance (e.g., massive volumes of terrain-level street views vs. limited aerial drone imagery), quality imbalance (e.g., high-precision professional LiDAR data vs. noisy crowd-sourced images), and temporal imbalance (e.g., static historical map data vs. sparse real-time sensor streams) [250].

Future Direction: Future research will explore new fusion architectures that not only fuse this heterogeneous data but also enable robust inference and decision-making when critical data is missing or of low quality, preventing the model from being overwhelmed by an abundance of poor data.

7.1.2. Continual Evolution

The majority of current UrbanEAs perform one-shot understanding on a static dataset, generating a static snapshot of a city. However, a city is a constantly evolving entity [3,251]. A future challenge is to evolve the agent's cognition from static snapshots to a dynamic world model capable of continual learning and incremental updates.

Future Direction: This requires solving critical problems such as catastrophic forgetting (forgetting old scenes when learning new ones), model drift, and efficiently managing never-ending data streams. Achieving this goal would elevate the solution to the dynamic gap from handling minute-long sequences to managing year-long urban evolution.

7.2. System-Level Challenges

7.2.1. Sim-to-Real Loop

While existing urban simulators like CARLA [24] and MetaUrban [68] are powerful, a sim-to-real gap persists, both in terms of perceptual realism (e.g., simulation of lighting, sensor noise) and behavioral realism (e.g., pedestrian and vehicle driving habits).

Future Direction: A future direction is to leverage UrbanEA to advance simulator development by constructing high-fidelity digital twins from real-world data. By extracting physical materials, lighting properties, and behavioral patterns, we can establish a Real-to-Sim-to-Real loop where real-world data improves the simulator, which in turn trains more capable agents. This cycle supports the convergence of UrbanEA and urban computing into a unified Cyber-Physical System (CPS). In this framework, the

 doi:10.20944/preprints202601.2155.v2

global intelligence from urban computing can guide the local policies of UrbanEAs, while UrbanEAs serve as mobile sensors to update the urban computing models in real-time.

7.2.2. Swarm Intelligence

The city of the future will be a massive distributed system composed of thousands of independent agents with only local sensing (e.g., autonomous vehicles, drones, infrastructure sensors). A challenge, and the key to the emergence of swarm intelligence, is enabling these agents to collaborate efficiently and safely [252,253].

Future Direction: Future research must address communication bottlenecks, decentralized decision-making consistency, and collaboration among heterogeneous agents [254]. In this vision, the unified world model built by the urban perception can serve as a digital bedrock, providing an authoritative and consistent environmental representation for all other micro-agents to query and interact with, thus enabling the leap from single-agent intelligence to large-scale swarm intelligence.

7.3. Societal-Level Challenges

7.3.1. Social Reasoning

This represents the leap across the semantic gap. Current data modeling operates at the level of physical sensing (geometry, appearance), whereas a future agent must be able to fuse non-physical, abstract social knowledge [255].

Future Direction: By integrating information like social media trends, news events, event schedules, and weather alerts, the agent can transition from answering *"what is happening"* to explaining *"why it is happening"*. This modeling of sensing with abstraction will transform the UrbanEAs from a powerful descriptive tool into an explanatory and predictive tool with rudimentary causal reasoning, providing the true intelligence required for the advanced tasks outlined in Section VI.

7.3.2. Algorithmic Fairness

The challenge of imbalanced multidomain data transcends a technical concern, posing an issue of social equity. Data perception often mirrors existing socio-economic disparities, leading to the over-representation of affluent areas and the under-representation of marginalized communities [256–258]. An agent trained on such data would develop a biased worldview; it will learn to equate data-richness with importance, effectively rendering data-poor areas invisible [259–261].

Future Direction: Future research will create data modeling methods that can be audited. It must focus on developing fairness-aware modeling algorithms. These methods must not only audit and quantify data-driven disparities but also actively correct for them. The goal is to ensure the agent's decisions promote allocative equity, ensuring that urban services and resources (discussed in Section VI) are distributed justly, even when the input data itself is unjust [262].

8. Conclusion

In this paper, we comprehensively survey the emerging field of data lifecycle for UrbanEA. We systematically review the existing work on this entire data lifecycle, including data perception, data management, data modeling, and downstream task applications. We compare mainstream simulators and data management architectures in terms of what they contain and how they are constructed. We analyze the mainstream multidomain data modeling strategies, highlighting the three core challenges in this field (environmental variability, scale limitation, and interaction complexity) and the four major gaps in multidomain data modeling. Finally, we point out potential research opportunities with existing tasks, and list several promising future directions for the field, such as continual learning, causal fusion, and large-scale multi-agent collaboration.

References

1. Bettencourt, L.M. The origins of scaling in cities. *science* **2013**, *340*, 1438–1441.

 doi:10.20944/preprints202601.2155.v2

2. Dong, L.; Duarte, F.; Duranton, G.; Santi, P.; Barthelemy, M.; Batty, M.; Bettencourt, L.; Goodchild, M.; Hack, G.; Liu, Y.; et al. Defining a city—delineating urban areas using cell-phone data. *Nature Cities* **2024**, *1*, 117–125.
3. Yang, L.; Luo, Z.; Zhang, S.; Teng, F.; Li, T. Continual learning for smart City: A survey. *IEEE Transactions on Knowledge and Data Engineering* **2024**.
4. Cengiz, B.; Adam, I.Y.; Ozdem, M.; Das, R. A survey on data fusion approaches in IoT-based smart cities: Smart applications, taxonomies, challenges, and future research directions. *Information Fusion* **2025**, *121*, 103102.
5. Bibri, S.E.; Huang, J. Artificial Intelligence of Things for Sustainable Smart City Brain and Digital Twin Systems: Environmental Synergies between Real-Time Management and Predictive Planning. *Environmental Science and Ecotechnology* **2025**, p. 100591.
6. Xu, F.; Zhang, J.; Gao, C.; Feng, J.; Li, Y. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813* **2023**.
7. Song, Y.; Sun, P.; Liu, H.; Li, Z.; Song, W.; Xiao, Y.; Zhou, X. Scene-driven multimodal knowledge graph construction for embodied AI. *IEEE Transactions on Knowledge and Data Engineering* **2024**, *36*, 6962–6976.
8. Yuan, Y.; Li, Z.; Zhao, B. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys* **2025**, *57*, 1–34.
9. Zhang, Y.; Ma, Z.; Li, J.; Qiao, Y.; Wang, Z.; Chai, J.; Wu, Q.; Bansal, M.; Kordjamshidi, P. Vision-and-Language Navigation Today and Tomorrow: A Survey in the Era of Foundation Models. *Transactions on Machine Learning Research*.
10. Jin, G.; Liang, Y.; Fang, Y.; Shao, Z.; Huang, J.; Zhang, J.; Zheng, Y. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering* **2023**, *36*, 5388–5408.
11. Rahmani, S.; Baghbani, A.; Bouguila, N.; Patterson, Z. Graph neural networks for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *24*, 8846–8885.
12. Zhang, W.; Han, J.; Xu, Z.; Ni, H.; Liu, H.; Xiong, H. Urban foundation models: A survey. In Proceedings of the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6633–6643.
13. Zou, X.; Yan, Y.; Hao, X.; Hu, Y.; Wen, H.; Liu, E.; Zhang, J.; Li, Y.; Li, T.; Zheng, Y.; et al. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion* **2025**, *113*, 102606.
14. Song, S.; Li, X.; Li, S.; Zhao, S.; Yu, J.; Ma, J.; Mao, X.; Zhang, W.; Wang, M. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering* **2025**.
15. Yao, S.; Guan, R.; Peng, Z.; Xu, C.; Shi, Y.; Ding, W.; Lim, E.G.; Yue, Y.; Seo, H.; Man, K.L.; et al. Exploring radar data representations in autonomous driving: A comprehensive review. *IEEE Transactions on Intelligent Transportation Systems* **2025**.
16. Mao, R.; Li, Y.; Li, G.; Hildre, H.P.; Zhang, H. A systematic survey of digital twin applications: Transferring knowledge from automotive and aviation to maritime industry. *IEEE Transactions on Intelligent Transportation Systems* **2025**.
17. Liu, H.; Tong, Y.; Han, J.; Zhang, P.; Lu, X.; Xiong, H. Incorporating multi-source urban data for personalized and context-aware multi-modal transportation recommendation. *IEEE Transactions on Knowledge and Data Engineering* **2020**, *34*, 723–735.
18. Ruan, W.; Wang, W.; Zhong, S.; Chen, W.; Liu, L.; Liang, Y. Cross space and time: A spatio-temporal unitized model for traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems* **2025**.
19. Liang, Y.; Wen, H.; Xia, Y.; Jin, M.; Yang, B.; Salim, F.; Wen, Q.; Pan, S.; Cong, G. Foundation models for spatio-temporal data science: A tutorial and survey. In Proceedings of the Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, 2025, pp. 6063–6073.
20. Li, Z.; Zhang, T.; Zhou, M.; Tang, D.; Zhang, P.; Liu, W.; Yang, Q.; Shen, T.; Wang, K.; Liu, H. Mipd: A multi-sensory interactive perception dataset for embodied intelligent driving. *IEEE Transactions on Intelligent Transportation Systems* **2025**.
21. Qu, J.; Liu, R.W.; Gao, Y.; Guo, Y.; Zhu, F.; Wang, F.Y. Double domain guided real-time low-light image enhancement for ultra-high-definition transportation surveillance. *IEEE Transactions on Intelligent Transportation Systems* **2024**, *25*, 9550–9562.

 doi:10.20944/preprints202601.2155.v2

22. Testolina, P.; Barbato, F.; Michieli, U.; Giordani, M.; Zanuttigh, P.; Zorzi, M. Selma: Semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *24*, 7012–7024.
23. Guan, R.; Jia, L.; Yao, S.; Yang, F.; Xu, S.; Purwanto, E.; Zhu, X.; Man, K.L.; Lim, E.G.; Smith, J.; et al. Watervg: Waterway visual grounding based on text-guided vision and mmwave radar. *IEEE Transactions on Intelligent Transportation Systems* **2025**, *26*, 7275–7291.
24. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An open urban driving simulator. In Proceedings of the Conference on robot learning. PMLR, 2017, pp. 1–16.
25. Bisio, I.; Delfino, A.; Grattarola, A.; Lavagetto, F.; Sciarrone, A. Ultrasounds-based context sensing method and applications over the Internet of Things. *IEEE Internet of Things Journal* **2018**, *5*, 3876–3890.
26. Phipps, A.; Ouazzane, K.; Vassilev, V. Enhancing cyber security using audio techniques: a public key infrastructure for sound. In Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2020, pp. 1428–1436.
27. Li, K.; Liu, M. Combined influence of multi-sensory comfort in winter open spaces and its association with environmental factors: Wuhan as a case study. *Building and Environment* **2024**, *248*, 111037.
28. Yin, C.; Chen, P.Y.; Yao, B.; Wang, D.; Caterino, J.; Zhang, P. SepsisLab: early sepsis prediction with uncertainty quantification and active sensing. In Proceedings of the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6158–6168.
29. Chen, C.; Jain, U.; Schissler, C.; Gari, S.V.A.; Al-Halah, Z.; Ithapu, V.K.; Robinson, P.; Grauman, K. Soundspaces: Audio-visual navigation in 3d environments. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer, 2020, pp. 17–36.
30. Chen, C.; Schissler, C.; Garg, S.; Kobernik, P.; Clegg, A.; Calamia, P.; Batra, D.; Robinson, P.; Grauman, K. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems* **2022**, *35*, 8896–8911.
31. Clarke, S.; Gao, R.; Wang, M.; Rau, M.; Xu, J.; Wang, J.H.; James, D.L.; Wu, J. Realimpact: A dataset of impact sound fields for real objects. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1516–1525.
32. Gan, C.; Gu, Y.; Zhou, S.; Schwartz, J.; Alter, S.; Traer, J.; Gutfreund, D.; Tenenbaum, J.B.; McDermott, J.H.; Torralba, A. Finding fallen objects via asynchronous audio-visual integration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10523–10533.
33. Gao, R.; Li, H.; Dharan, G.; Wang, Z.; Li, C.; Xia, F.; Savarese, S.; Fei-Fei, L.; Wu, J. Sonicverse: A multisensory simulation platform for embodied household agents that see and hear. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 704–711.
34. Narang, Y.; Sundaralingam, B.; Macklin, M.; Mousavian, A.; Fox, D. Sim-to-real for robotic tactile sensing via physics-based simulation and learned latent projections. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021, pp. 6444–6451.
35. Gao, R.; Si, Z.; Chang, Y.Y.; Clarke, S.; Bohg, J.; Fei-Fei, L.; Yuan, W.; Wu, J. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10598–10608.
36. Gao, R.; Dou, Y.; Li, H.; Agarwal, T.; Bohg, J.; Li, Y.; Fei-Fei, L.; Wu, J. The objectfolder benchmark: Multisensory learning with neural and real objects. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17276–17286.
37. Gao, R.; Chang, Y.Y.; Mall, S.; Fei-Fei, L.; Wu, J. ObjectFolder: A Dataset of Objects with Implicit Visual, Auditory, and Tactile Representations. In Proceedings of the Conference on Robot Learning, 2021.
38. Calandra, R.; Owens, A.; Jayaraman, D.; Lin, J.; Yuan, W.; Malik, J.; Adelson, E.H.; Levine, S. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters* **2018**, *3*, 3300–3307.
39. Hong, Y.; Zheng, Z.; Chen, P.; Wang, Y.; Li, J.; Gan, C. Multiply: A multisensory object-centric embodied large language model in 3d world. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26406–26416.
40. Zhang, W.; Han, J.; Xu, Z.; Ni, H.; Liu, H.; Xiong, H. Towards urban general intelligence: A review and outlook of urban foundation models. *arXiv preprint arXiv:2402.01749* **2024**.

 doi:10.20944/preprints202601.2155.v2

41. Fadhel, M.A.; Duhaim, A.M.; Saihood, A.; Sewify, A.; Al-Hamadani, M.N.; Albahri, A.; Alzubaidi, L.; Gupta, A.; Mirjalili, S.; Gu, Y. Comprehensive systematic review of information fusion methods in smart cities and urban environments. *Information Fusion* **2024**, *107*, 102317.
42. El-Omari, S.; Moselhi, O. Integrating 3D laser scanning and photogrammetry for progress measurement of construction work. *Automation in construction* **2008**, *18*, 1–9.
43. Navares-Vázquez, J.C.; Qiu, Z.; Arias, P.; Balado, J. HoloLens 2 performance analysis for indoor/outdoor 3D mapping. *Journal of Building Engineering* **2025**, p. 112826.
44. Rashdi, R.; Garrido, I.; Balado, J.; Del Río-Barral, P.; Rodríguez-Somoza, J.L.; Martínez-Sánchez, J. Comparative Evaluation of LiDAR systems for transport infrastructure: case studies and performance analysis. *European Journal of Remote Sensing* **2024**, *57*, 2316304.
45. Seifert, E.; Seifert, S.; Vogt, H.; Drew, D.; Van Aardt, J.; Kunneke, A.; Seifert, T. Influence of drone altitude, image overlap, and optical sensor resolution on multi-view reconstruction of forest images. *Remote sensing* **2019**, *11*, 1252.
46. Girindran, R.; Boyd, D.S.; Rosser, J.; Vijayan, D.; Long, G.; Robinson, D. On the reliable generation of 3D city models from open data. *Urban Science* **2020**, *4*, 47.
47. Zhang, H.K.; Roy, D.P.; Yan, L.; Li, Z.; Huang, H.; Vermote, E.; Skakun, S.; Roger, J.C. Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences. *Remote sensing of environment* **2018**, *215*, 482–494.
48. Xiao, C.; Zhou, J.; Xiao, Y.; Huang, J.; Xiong, H. Refound: Crafting a foundation model for urban region understanding upon language and visual foundations. In Proceedings of the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 3527–3538.
49. Duan, J.; Yu, S.; Tan, H.L.; Zhu, H.; Tan, C. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence* **2022**, *6*, 230–244.
50. Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. Habitat: A platform for embodied ai research. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9339–9347.
51. Puig, X.; Undersander, E.; Szot, A.; Cote, M.D.; Yang, T.Y.; Partsey, R.; Desai, R.; Clegg, A.; Hlavac, M.; Min, S.Y.; et al. Habitat 3.0: A Co-Habitat for Humans, Avatars, and Robots. In Proceedings of the The Twelfth International Conference on Learning Representations.
52. Szot, A.; Clegg, A.; Undersander, E.; Wijmans, E.; Zhao, Y.; Turner, J.; Maestre, N.; Mukadam, M.; Chaplot, D.S.; Maksymets, O.; et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems* **2021**, *34*, 251–266.
53. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5828–5839.
54. Zhao, C.; Wang, X.; Lv, Y.; Tian, Y.; Lin, Y.; Wang, F.Y. Parallel transportation in TransVerse: From foundation models to DeCAST. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *24*, 15310–15327.
55. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.
56. Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M.; Bulatov, Y.; McCord, B. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856* **2018**.
57. Chen, H.; Suhr, A.; Misra, D.; Snavely, N.; Artzi, Y. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12538–12547.
58. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11621–11631.
59. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.
60. Liao, Y.; Xie, J.; Geiger, A. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *45*, 3292–3310.

 doi:10.20944/preprints202601.2155.v2

61. Chen, M.; Hu, Q.; Yu, Z.; Thomas, H.; Feng, A.; Hou, Y.; McCullough, K.; Ren, F.; Soibelman, L. STPLS3D: A Large-Scale Synthetic and Real Aerial Photogrammetry 3D Point Cloud Dataset. In Proceedings of the 33rd British Machine Vision Conference Proceedings, BMVC 2022, 2022.
62. Hu, Q.; Yang, B.; Khalid, S.; Xiao, W.; Trigoni, N.; Markham, A. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision* **2022**, *130*, 316–343.
63. Yang, G.; Xue, F.; Zhang, Q.; Xie, K.; Fu, C.W.; Huang, H. UrbanBIS: a large-scale benchmark for fine-grained urban building instance segmentation. In Proceedings of the ACM SIGGRAPH 2023 Conference Proceedings, 2023, pp. 1–11.
64. Liu, S.; Zhang, H.; Qi, Y.; Wang, P.; Zhang, Y.; Wu, Q. Aerialvln: Vision-and-language navigation for uavs. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15384–15394.
65. Wang, H.; Chen, J.; Huang, W.; Ben, Q.; Wang, T.; Mi, B.; Huang, T.; Zhao, S.; Chen, Y.; Yang, S.; et al. Grutopia: Dream general robots in a city at scale. *arXiv preprint arXiv:2407.10943* **2024**.
66. Wang, X.; Yang, D.; Wang, Z.; Kwan, H.; Chen, J.; Wu, W.; Li, H.; Liao, Y.; Liu, S. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *arXiv preprint arXiv:2410.07087* **2024**.
67. Zhong, F.; Wu, K.; Wang, C.; Chen, H.; Ci, H.; Li, Z.; Wang, Y. Unrealzoo: Enriching photo-realistic virtual worlds for embodied ai. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 5769–5779.
68. Wu, W.; He, H.; He, J.; Wang, Y.; Duan, C.; Liu, Z.; Li, Q.; Zhou, B. MetaUrban: An Embodied AI Simulation Platform for Urban Micromobility. *International Conference on Learning Representation* **2025**.
69. Gao, Y.; Li, C.; You, Z.; Liu, J.; Li, Z.; Chen, P.; Chen, Q.; Tang, Z.; Wang, L.; Yang, P.; et al. OpenFly: A Versatile Toolchain and Large-scale Benchmark for Aerial Vision-Language Navigation. *arXiv preprint arXiv:2502.18041* **2025**.
70. Mirowski, P.; Banki-Horvath, A.; Anderson, K.; Teplyashin, D.; Hermann, K.M.; Malinowski, M.; Grimes, M.K.; Simonyan, K.; Kavukcuoglu, K.; Zisserman, A.; et al. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292* **2019**.
71. Liu, Y.; Liu, S.; Chen, B.; Yang, Z.X.; Xu, S. Fusion-Perception-to-Action Transformer: Enhancing Robotic Manipulation with 3D Visual Fusion Attention and Proprioception. *IEEE Transactions on Robotics* **2025**.
72. Warren, J.; Marz, N. *Big Data: Principles and best practices of scalable realtime data systems*; Simon and Schuster, 2015.
73. Kreps, J. Questioning the Lambda Architecture. O'Reilly Radar, 2014.
74. Azzabi, S.; Alfughi, Z.; Ouda, A. Data lakes: A survey of concepts and architectures. *Computers* **2024**, *13*, 183.
75. Tahara, D.; Diamond, T.; Abadi, D.J. Sinew: a SQL system for multi-structured data. In Proceedings of the Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, 2014, pp. 815–826.
76. Bugiotti, F.; Cabibbo, L.; Atzeni, P.; Torlone, R. Database design for NoSQL systems. In Proceedings of the International Conference on Conceptual Modeling. Springer, 2014, pp. 223–231.
77. Zhang, C.; Lu, J.; Xu, P.; Chen, Y. UniBench: a benchmark for multi-model database management systems. In Proceedings of the Technology conference on performance evaluation and benchmarking. Springer, 2018, pp. 7–23.
78. Neo4j, Inc.. Neo4j, 2010.
79. The JanusGraph Project. JanusGraph, 2017.
80. Mlodzian, L.; Sun, Z.; Berkemeyer, H.; Monka, S.; Wang, Z.; Dietze, S.; Halilaj, L.; Luettin, J. nuScenes Knowledge Graph-A comprehensive semantic representation of traffic scenes for trajectory prediction. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 42–52.
81. Sun, P.; Song, Y.; Liu, X.; Yang, X.; Wang, Q.; Li, T.; Yang, Y.; Chu, X. 3d question answering for city scene understanding. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 2156–2165.
82. Meta AI Research. FAISS, 2017.
83. Wang, J.; Yi, X.; Guo, R.; Jin, H.; Xu, P.; Li, S.; Wang, X.; Guo, X.; Li, C.; Xu, X.; et al. Milvus: A purpose-built vector data management system. In Proceedings of the Proceedings of the 2021 international conference on management of data, 2021, pp. 2614–2627.
84. Malkov, Y.A.; Yashunin, D.A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* **2018**, *42*, 824–836.

 doi:10.20944/preprints202601.2155.v2

85. Jegou, H.; Douze, M.; Schmid, C. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* **2010**, *33*, 117–128.
86. Pelkonen, T.; Franklin, S.; Teller, J.; Cavallaro, P.; Huang, Q.; Meza, J.; Veeraraghavan, K. Gorilla: A fast, scalable, in-memory time series database. *Proceedings of the VLDB Endowment* **2015**, *8*, 1816–1827.
87. Wang, C.; Qiao, J.; Huang, X.; Song, S.; Hou, H.; Jiang, T.; Rui, L.; Wang, J.; Sun, J. Apache iotdb: A time series database for iot applications. *Proceedings of the ACM on Management of Data* **2023**, *1*, 1–27.
88. Timescale. TimescaleDB, 2018.
89. Zimányi, E.; Sakr, M.; Lesuisse, A. MobilityDB: A mobility database based on PostgreSQL and PostGIS. *ACM Transactions on Database Systems (TODS)* **2020**, *45*, 1–42.
90. OSGeo, P.P.. PostGIS. https://postgis.net/, 2025.
91. Li, R.; He, H.; Wang, R.; Ruan, S.; He, T.; Bao, J.; Zhang, J.; Hong, L.; Zheng, Y. TrajMesa: A distributed NoSQL-based trajectory data management system. *IEEE Transactions on Knowledge and Data Engineering* **2021**, *35*, 1013–1027.
92. He, H.; Xu, Z.; Li, R.; Bao, J.; Li, T.; Zheng, Y. TMan: a high-performance trajectory data management system based on key-value stores. In Proceedings of the 2024 IEEE 40th International Conference on Data Engineering (ICDE). IEEE, 2024, pp. 4951–4964.
93. Kumar, R.; Bhanu, M.; Mendes-Moreira, J.; Chandra, J. Spatio-temporal predictive modeling techniques for different domains: a survey. *ACM Computing Surveys* **2024**, *57*, 1–42.
94. Wang, S.; Bao, Z.; Culpepper, J.S.; Cong, G. A survey on trajectory data management, analytics, and learning. *ACM Computing Surveys (CSUR)* **2021**, *54*, 1–36.
95. Han, J.; Ning, Y.; Yuan, Z.; Ni, H.; Liu, F.; Lyu, T.; Liu, H. Large Language Model Powered Intelligent Urban Agents: Concepts, Capabilities, and Applications. *arXiv preprint arXiv:2507.00914* **2025**.
96. Lu, Y.; Tang, H. Multimodal data storage and retrieval for embodied ai: A survey. *arXiv preprint arXiv:2508.13901* **2025**.
97. Xu, F.; Zhang, J.; Gao, C.; Feng, J.; Li, Y. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813* **2023**.
98. Liu, Y.; Ding, J.; Fu, Y.; Li, Y. Urbankg: An urban knowledge graph system. *ACM Transactions on Intelligent Systems and Technology* **2023**, *14*, 1–25.
99. Lei, M.; Cai, H.; Cui, Z.; Tan, L.; Hong, J.; Hu, G.; Zhu, S.; Wu, Y.; Jiang, S.; Wang, G.; et al. RoboMemory: A Brain-inspired Multi-memory Agentic Framework for Lifelong Learning in Physical Embodied Systems **2025**.
100. Sawadogo, P.; Darmont, J. On data lake architectures and metadata management. *J. Intell. Inf. Syst.* **2021**, *56*, 97–120.
101. Hai, R.; Koutras, C.; Quix, C.; Jarke, M. Data lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering* **2023**, *35*, 12571–12590.
102. Liu, Q.; Han, T.; Xie, J.; Kim, B. Real-time dynamic map with crowdsourcing vehicles in edge computing. *IEEE Transactions on Intelligent Vehicles* **2022**, *8*, 2810–2820.
103. Ito, S.; Okamura, R.; Zhao, C.; Azumi, T. SIM-LDM: Local Dynamic Map Generation Framework using Autonomous Driving Simulator. In Proceedings of the Proceedings of the 4th International Workshop on Real-time and IntelliGent Edge computing, 2025, pp. 1–6.
104. Lu, J.; Holubová, I. Multi-model Data Management: What's New and What's Next? In Proceedings of the EDBT, 2017.
105. Yeo, J.; Cho, H.; Park, J.W.; Hwang, S.W. Multimodal KB harvesting for emerging spatial entities. *IEEE Transactions on Knowledge and Data Engineering* **2017**, *29*, 1073–1086.
106. Košmerl, I.; Rabuzin, K.; Šestak, M. Multi-model databases-Introducing polyglot persistence in the big data world. In Proceedings of the 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE, 2020, pp. 1724–1729.
107. Khine, P.P.; Wang, Z. A review of polyglot persistence in the big data world. *Information* **2019**, *10*, 141.
108. Bimonte, S.; Gallinucci, E.; Marcel, P.; Rizzi, S. Data variety, come as you are in multi-model data warehouses. *Information Systems* **2022**, *104*, 101734.
109. Kolev, B.; Valduriez, P.; Bondiombouy, C.; Jiménez-Peris, R.; Pau, R.; Pereira, J. CloudMdsQL: querying heterogeneous cloud data stores with a common language. *Distrib. Parallel Databases* **2016**, *34*, 463–503.
110. Mihai, G. Multi-model database systems: The state of affairs. *Economics and Applied Informatics* **2020**, pp. 211–215.

 doi:10.20944/preprints202601.2155.v2

111. Viktorović, M.; Yang, D.; Vries, B.d. Connected Traffic Data Ontology (CTDO) for intelligent urban traffic systems focused on connected (Semi) autonomous vehicles. *Sensors* **2020**, *20*, 2961.
112. Chadzynski, A.; Li, S.; Grišiūtė, A.; Chua, J.; Hofmeister, M.; Yan, J.; Tai, H.Y.; Lloyd, E.; Tsai, Y.K.; Agarwal, M.; et al. Semantic 3D city interfaces—Intelligent interactions on dynamic geospatial knowledge graphs. *Data-Centric Engineering* **2023**, *4*, e20.
113. Xiao, C.; Zhou, J.; Huang, J.; Zhu, H.; Xu, T.; Dou, D.; Xiong, H. A contextual master-slave framework on urban region graph for urban village detection. In Proceedings of the 2023 IEEE 39th International Conference on Data Engineering (ICDE). IEEE, 2023, pp. 736–748.
114. Fang, Z.; Long, Q.; Song, G.; Xie, K. Spatial-temporal graph ode networks for traffic flow forecasting. In Proceedings of the Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021, pp. 364–373.
115. Guo, S.; Lin, Y.; Wan, H.; Li, X.; Cong, G. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering* **2021**, *34*, 5415–5428.
116. Wang, Z.; Han, F.; Zhao, S. A Survey on Knowledge Graph Related Research in Smart City Domain. *ACM Transactions on Knowledge Discovery from Data* **2024**, *18*, 1–31.
117. Kumar Kaliyar, R. Graph databases: A survey. In Proceedings of the International Conference on Computing, Communication & Automation. IEEE, 2015, pp. 785–790.
118. Desai, M.; G Mehta, R.; P Rana, D. Issues and challenges in big graph modelling for smart city: an extensive survey. *International Journal of Computational Intelligence & IoT* **2018**, *1*.
119. Lv, C.; Qi, M.; Liu, L.; Ma, H. T2sg: Traffic topology scene graph for topology reasoning in autonomous driving. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 17197–17206.
120. Liu, Y.; Ding, J.; Li, Y. KnowSite: Leveraging urban knowledge graph for site selection. In Proceedings of the Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, 2023, pp. 1–12.
121. Liu, J.; Li, T.; Ji, S.; Xie, P.; Du, S.; Teng, F.; Zhang, J. Urban flow pattern mining based on multi-source heterogeneous data fusion and knowledge graph embedding. *IEEE Transactions on Knowledge and Data Engineering* **2021**, *35*, 2133–2146.
122. Zareian, A.; Karaman, S.; Chang, S.F. Bridging knowledge graphs to generate scene graphs. In Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 606–623.
123. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Transactions on Neural Networks* **2008**, *20*, 61–80.
124. Sun, Z.; Wang, Z.; Halilaj, L.; Luettin, J. Semanticformer: Holistic and semantic traffic scene representation for trajectory prediction using knowledge graphs. *IEEE Robotics and Automation Letters* **2024**.
125. de Vos, K.; Torta, E.; Bruyninckx, H.; Martínez, C.L.; van de Molengraft, M. Automatic configuration of multi-agent model predictive controllers based on semantic graph world models. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 7194–7200.
126. Holmberg, E.; Ioup, E.; Abdelguerfi, M. A Knowledge-Graph Translation Layer for Mission-Aware Multi-Agent Path Planning in Spatiotemporal Dynamics. *arXiv preprint arXiv:2510.21695* **2025**.
127. He, H.; Li, R.; Ruan, S.; He, T.; Bao, J.; Li, T.; Zheng, Y. Trass: Efficient trajectory similarity search based on key-value data stores. In Proceedings of the 2022 IEEE 38th International conference on Data Engineering (ICDE). IEEE, 2022, pp. 2306–2318.
128. Sun, F.; Qi, J.; Chang, Y.; Fan, X.; Karunasekera, S.; Tanin, E. Urban region representation learning with attentive fusion. In Proceedings of the 2024 IEEE 40th International Conference on Data Engineering (ICDE). IEEE, 2024, pp. 4409–4421.
129. Lim, J.H.; Kang, W.J.; Singh, S.; Narasimhalu, D. Learning similarity matching in multimedia content-based retrieval. *IEEE Transactions on Knowledge and Data Engineering* **2001**, *13*, 846–850.
130. Chen, Y.; Sampathkumar, H.; Luo, B.; Chen, X.w. ilike: Bridging the semantic gap in vertical image search by integrating text and visual features. *IEEE Transactions on Knowledge and Data Engineering* **2012**, *25*, 2257–2270.
131. Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; Jiang, Y.G. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 4542–4550.
132. Park, S.; Lee, M.; Kang, J.; Choi, H.; Park, Y.; Cho, J.; Lee, A.; Kim, D. Vlaad: Vision and language assistant for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 980–987.

 doi:10.20944/preprints202601.2155.v2

133. Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; Li, H. Drivelm: Driving with graph visual question answering. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 256–274.
134. Xie, Q.; Min, S.Y.; Ji, P.; Yang, Y.; Zhang, T.; Xu, K.; Bajaj, A.; Salakhutdinov, R.; Johnson-Roberson, M.; Bisk, Y. Embodied-rag: General non-parametric embodied memory for retrieval and generation. *arXiv preprint arXiv:2409.18313* **2024**.
135. Ocker, F.; Deigmöller, J.; Smirnov, P.; Eggert, J. A grounded memory system for smart personal assistants. *arXiv preprint arXiv:2505.06328* **2025**.
136. Ragab, M.; Gong, P.; Eldele, E.; Zhang, W.; Wu, M.; Foo, C.S.; Zhang, D.; Li, X.; Chen, Z. Evidentially calibrated source-free time-series domain adaptation with temporal imputation. *IEEE Transactions on Knowledge and Data Engineering* **2025**.
137. Gao, C.; Zhao, B.; Zhang, W.; Mao, J.; Zhang, J.; Zheng, Z.; Man, F.; Fang, J.; Zhou, Z.; Cui, J.; et al. EmbodiedCity: A Benchmark Platform for Embodied Agent in Real-world City Environment. *arXiv preprint arXiv:2410.09604* **2024**.
138. Li, R.; He, H.; Wang, R.; Huang, Y.; Liu, J.; Ruan, S.; He, T.; Bao, J.; Zheng, Y. Just: Jd urban spatio-temporal data engine. In Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020, pp. 1558–1569.
139. Guo, Y.; Wang, T.; Chen, Z.; Shao, Z. A Storage Model with Fine-Grained In-Storage Query Processing for Spatio-Temporal Data. In Proceedings of the 2025 IEEE 41st International Conference on Data Engineering (ICDE). IEEE, 2025, pp. 669–682.
140. Shi, H.; Du, S.; Yang, Y.; Zhang, J.; Li, T.; Zheng, Y. A Knowledge-Guided Pre-Training Temporal Data Analysis Foundation Model for Urban Computing. *IEEE Transactions on Knowledge and Data Engineering* **2025**.
141. Chen, J.; Zhang, A. On hierarchical disentanglement of interactive behaviors for multimodal spatiotemporal data with incompleteness. In Proceedings of the Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 213–225.
142. Vitale, V.N.; Martino, S.D.; Peron, A.; Russo, M.; Battista, E. How to manage massive spatiotemporal dataset from stationary and non-stationary sensors in commercial DBMS? *Knowledge and Information Systems* **2024**, *66*, 2063–2088.
143. Chen, M.; Li, Z.; Huang, W.; Gong, Y.; Yin, Y. Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology. In Proceedings of the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 319–328.
144. Vasudevan, A.B.; Dai, D.; Van Gool, L. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision* **2021**, *129*, 246–266.
145. Mao, Y.; Zhou, H.; Chen, L.; Qi, R.; Sun, Z.; Rong, Y.; He, X.; Chen, M.; Mumtaz, S.; Frascolla, V.; et al. A Survey on Spatio-Temporal Prediction: From Transformers to Foundation Models. *ACM Computing Surveys* **2025**.
146. Xie, P.; Ma, M.; Li, T.; Ji, S.; Du, S.; Yu, Z.; Zhang, J. Spatio-temporal dynamic graph relation learning for urban metro flow prediction. *IEEE Transactions on Knowledge and Data Engineering* **2023**, *35*, 9973–9984.
147. Li, Z.; Xia, L.; Tang, J.; Xu, Y.; Shi, L.; Xia, L.; Yin, D.; Huang, C. Urbangpt: Spatio-temporal large language models. In Proceedings of the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 5351–5362.
148. Ginting, M.F.; Kim, D.K.; Meng, X.; Reinke, A.; Krishna, B.J.; Kayhani, N.; Peltzer, O.; Fan, D.D.; Shaban, A.; Kim, S.K.; et al. Enter the mind palace: Reasoning and planning for long-term active embodied question answering. *arXiv preprint arXiv:2507.12846* **2025**.
149. Liu, Y.; Chen, W.; Bai, Y.; Liang, X.; Li, G.; Gao, W.; Lin, L. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics* **2025**.
150. Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems* **2019**, *20*, 3782–3795.
151. Huang, J.; Yong, S.; Ma, X.; Linghu, X.; Li, P.; Wang, Y.; Li, Q.; Zhu, S.C.; Jia, B.; Huang, S. An embodied generalist agent in 3D world. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 20413–20451.

 doi:10.20944/preprints202601.2155.v2

152. Christodoulides, A.; Tam, G.K.; Clarke, J.; Smith, R.; Horgan, J.; Micallef, N.; Morley, J.; Villamizar, N.; Walton, S. Survey on 3D Reconstruction Techniques: Large-Scale Urban City Reconstruction and Requirements. *IEEE Transactions on Visualization and Computer Graphics* **2025**.
153. Li, S.; Tang, H. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040* **2024**.
154. Chen, X.; Xu, W.; Zhang, S.; Cai, Y. Pedestrian Crossing Intention Prediction via Progressive Multimodal Token Fusion for Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems* **2025**.
155. Özyeşil, O.; Voroninski, V.; Basri, R.; Singer, A. A survey of structure from motion*. *Acta Numerica* **2017**, *26*, 305–364.
156. Lorensen, W.E.; Cline, H.E. Marching cubes: A high resolution 3D surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*; 1998; pp. 347–353.
157. Xu, L.; Xiangli, Y.; Peng, S.; Pan, X.; Zhao, N.; Theobalt, C.; Dai, B.; Lin, D. Grid-guided neural radiance fields for large urban scenes. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8296–8306.
158. Zhang, Q.; Wei, Y.; Han, Z.; Fu, H.; Peng, X.; Deng, C.; Hu, Q.; Xu, C.; Wen, J.; Hu, D.; et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947* **2024**.
159. Wolff, K.; Kim, C.; Zimmer, H.; Schroers, C.; Botsch, M.; Sorkine-Hornung, O.; Sorkine-Hornung, A. Point cloud noise and outlier removal for image-based 3D reconstruction. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 118–127.
160. Melas-Kyriazi, L.; Rupprecht, C.; Vedaldi, A. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 12923–12932.
161. Henderson, P.; Ferrari, V. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision* **2020**, *128*, 835–854.
162. Wu, C.; Liu, Y.; Dai, Q.; Wilburn, B. Fusing multiview and photometric stereo for 3d reconstruction under uncalibrated illumination. *IEEE transactions on visualization and computer graphics* **2010**, *17*, 1082–1095.
163. Kerl, C.; Souiai, M.; Sturm, J.; Cremers, D. Towards illumination-invariant 3D reconstruction using ToF RGB-D cameras. In Proceedings of the 2014 2nd International Conference on 3D Vision. IEEE, 2014, Vol. 1, pp. 39–46.
164. Bai, K.; Zhang, L.; Chen, Z.; Wan, F.; Zhang, J. Close the sim2real gap via physically-based structured light synthetic data simulation. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 17035–17041.
165. Stoffregen, T.; Scheerlinck, C.; Scaramuzza, D.; Drummond, T.; Barnes, N.; Kleeman, L.; Mahony, R. Reducing the sim-to-real gap for event cameras. In Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 534–549.
166. Köhler, T.; Bätz, M.; Naderi, F.; Kaup, A.; Maier, A.; Riess, C. Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data. *IEEE transactions on pattern analysis and machine intelligence* **2019**, *42*, 2944–2959.
167. Rematas, K.; Liu, A.; Srinivasan, P.P.; Barron, J.T.; Tagliasacchi, A.; Funkhouser, T.; Ferrari, V. Urban radiance fields. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12932–12942.
168. Deng, T.; Jiang, L.; Shi, Y.; Wu, J.; Wu, Z.; Yan, S.; Zhang, X.; Yan, H. Driving visual saliency prediction of dynamic night scenes via a spatio-temporal dual-encoder network. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *25*, 2413–2423.
169. Fink, L.; Rückert, D.; Franke, L.; Keinert, J.; Stamminger, M. Livenvs: Neural view synthesis on live rgb-d streams. In Proceedings of the SIGGRAPH Asia 2023 Conference Papers, 2023, pp. 1–11.
170. Stier, N.; Ranjan, A.; Colburn, A.; Yan, Y.; Yang, L.; Ma, F.; Angles, B. Finerecon: Depth-aware feed-forward network for detailed 3d reconstruction. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 18423–18432.
171. Huang, Y.; Du, J.; Yang, Z.; Zhou, Z.; Zhang, L.; Chen, H. A survey on trajectory-prediction methods for autonomous driving. *IEEE transactions on intelligent vehicles* **2022**, *7*, 652–674.
172. Chu, T.; Zhang, P.; Liu, Q.; Wang, J. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4937–4946.

 doi:10.20944/preprints202601.2155.v2

173. Huang, Z.; Jampani, V.; Thai, A.; Li, Y.; Stojanov, S.; Rehg, J.M. Shapeclipper: Scalable 3d shape learning from single-view images via geometric and clip-based consistency. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 12912–12922.
174. Wang, C.; Jiang, R.; Chai, M.; He, M.; Chen, D.; Liao, J. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics* **2023**, *30*, 4983–4996.
175. Mittal, P.; Cheng, Y.C.; Singh, M.; Tulsiani, S. Autosdf: Shape priors for 3d completion, reconstruction and generation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 306–315.
176. Zhao, F.; Zhang, C.; Geng, B. Deep multimodal data fusion. *ACM computing surveys* **2024**, *56*, 1–36.
177. Meng, L.; Tan, A.H.; Xu, D. Semi-supervised heterogeneous fusion for multimedia data co-clustering. *IEEE Transactions on Knowledge and Data Engineering* **2013**, *26*, 2293–2306.
178. Wang, C.; Zuo, K.; Zhang, S.; Lei, H.; Hu, P.; Shen, Z.; Wang, R.; Zhao, P. PFNet: Large-scale traffic forecasting with progressive spatio-temporal fusion. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *24*, 14580–14597.
179. Muturi, T.W.; Kyem, B.A.; Asamoah, J.K.; Owor, N.J.; Dzinyela, R.; Danyo, A.; Adu-Gyamfi, Y.; Aboah, A. Prompt-guided spatial understanding with rgb-d transformers for fine-grained object relation reasoning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 5280–5288.
180. Shang, Y.; Lin, Y.; Zheng, Y.; Fan, H.; Ding, J.; Feng, J.; Chen, J.; Tian, L.; Li, Y. Urbanworld: An urban world model for 3d city generation. *arXiv preprint arXiv:2407.11965* **2024**.
181. Turki, H.; Zhang, J.Y.; Ferroni, F.; Ramanan, D. Suds: Scalable urban dynamic scenes. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12375–12385.
182. Zheng, Z.; Zhou, M.; Shang, Z.; Wei, X.; Pu, H.; Luo, J.; Jia, W. GAANet: Graph Aggregation Alignment Feature Fusion for Multispectral Object Detection. *IEEE Transactions on Industrial Informatics* **2025**.
183. Lin, J.; Li, Z.; Tang, X.; Liu, J.; Liu, S.; Liu, J.; Lu, Y.; Wu, X.; Xu, S.; Yan, Y.; et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5166–5175.
184. Liu, Y.; Luo, C.; Fan, L.; Wang, N.; Peng, J.; Zhang, Z. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 265–282.
185. Jiang, L.; Ren, K.; Yu, M.; Xu, L.; Dong, J.; Lu, T.; Zhao, F.; Lin, D.; Dai, B. Horizon-GS: Unified 3D Gaussian Splatting for Large-Scale Aerial-to-Ground Scenes. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 26789–26799.
186. Vuong, K.; Ghosh, A.; Ramanan, D.; Narasimhan, S.; Tulsiani, S. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 21674–21684.
187. Huang, J.; Stoter, J.; Peters, R.; Nan, L. City3D: Large-scale building reconstruction from airborne LiDAR point clouds. *Remote Sensing* **2022**, *14*, 2254.
188. Zhang, C.; Cao, Y.; Zhang, L. CrossView-GS: Cross-view Gaussian Splatting For Large-scale Scene Reconstruction. *arXiv preprint arXiv:2501.01695* **2025**.
189. Feng, J.; Liu, T.; Du, Y.; Guo, S.; Lin, Y.; Li, Y. Citygpt: Empowering urban spatial cognition of large language models. In Proceedings of the Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, 2025, pp. 591–602.
190. Li, Y.; Pan, Y.; Zhu, G.; He, S.; Xu, M.; Xu, J. Charging-Aware Task Assignment for Urban Logistics with Electric Vehicles. *IEEE Transactions on Knowledge and Data Engineering* **2025**.
191. Lee, L.H.; Braud, T.; Hosio, S.; Hui, P. Towards augmented reality driven human-city interaction: Current research on mobile headsets and future challenges. *ACM Computing Surveys (CSUR)* **2021**, *54*, 1–38.
192. Kim, B.S.; Kim, J.; Lee, D.; Jang, B. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Computing Surveys* **2025**, *57*, 1–35.
193. Ding, X.; Han, J.; Xu, H.; Liang, X.; Zhang, W.; Li, X. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13668–13677.

 doi:10.20944/preprints202601.2155.v2

194. Wu, D.; Han, W.; Liu, Y.; Wang, T.; Xu, C.z.; Zhang, X.; Shen, J. Language prompt for autonomous driving. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 8359–8367.
195. Guan, R.; Hu, R.; Chen, S.; Xiao, N.; Xia, X.; Liu, J.; Chen, B.; Tang, Z.; Ouyang, N.; Liang, S.; et al. RoadSceneVQA: Benchmarking Visual Question Answering in Roadside Perception Systems for Intelligent Transportation System. *arXiv preprint arXiv:2511.18286* **2025**.
196. Wang, J.; Zheng, Z.; Chen, Z.; Ma, A.; Zhong, Y. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2024, Vol. 38, pp. 5481–5489.
197. Feng, J.; Zhang, J.; Liu, T.; Zhang, X.; Ouyang, T.; Yan, J.; Du, Y.; Guo, S.; Li, Y. Citybench: Evaluating the capabilities of large language models for urban tasks. *arXiv preprint arXiv:2406.13945* **2024**.
198. Bieri, V.; Zamboni, M.; Blumer, N.S.; Chen, Q.; Engelmann, F. Opencity3d: What do vision-language models know about urban environments? In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025, pp. 5147–5155.
199. Yasuki, S.; Miyanishi, T.; Inoue, N.; Kurita, S.; Sakamoto, K.; Azuma, D.; Taki, M.; Matsuo, Y. GeoProg3D: Compositional Visual Reasoning for City-Scale 3D Language Fields. *arXiv preprint arXiv:2506.23352* **2025**.
200. Wang, J.; Ma, A.; Chen, Z.; Zheng, Z.; Wan, Y.; Zhang, L.; Zhong, Y. EarthVQANet: Multi-task visual question answering for remote sensing image understanding. *ISPRS Journal of Photogrammetry and Remote Sensing* **2024**, *212*, 422–439.
201. Zhao, Y.; Xu, K.; Zhu, Z.; Hu, Y.; Zheng, Z.; Chen, Y.; Ji, Y.; Gao, C.; Li, Y.; Huang, J. Cityeqa: A hierarchical llm agent on embodied question answering benchmark in city space. *arXiv preprint arXiv:2502.12532* **2025**.
202. Gu, J.; Stefani, E.; Wu, Q.; Thomason, J.; Wang, X. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 7606–7623.
203. Schumann, R.; Riezler, S. Generating Landmark Navigation Instructions from Maps as a Graph-to-Text Problem. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 489–502.
204. Li, J.; Padmakumar, A.; Sukhatme, G.; Bansal, M. Vln-video: Utilizing driving videos for outdoor vision-and-language navigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 18517–18526.
205. Xu, Y.; Pan, Y.; Liu, Z.; Wang, H. Flame: Learning to navigate with multimodal llm in urban environments. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 9005–9013.
206. Schumann, R.; Zhu, W.; Feng, W.; Fu, T.J.; Riezler, S.; Wang, W.Y. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 18924–18933.
207. Xiang, J.; Wang, X.; Wang, W.Y. Learning to Stop: A Simple yet Effective Approach to Urban Vision-Language Navigation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 699–707.
208. Zhu, W.; Wang, X.; Fu, T.J.; Yan, A.; Narayana, P.; Sone, K.; Basu, S.; Wang, W.Y. Multimodal Text Style Transfer for Outdoor Vision-and-Language Navigation. In Proceedings of the Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 1207–1221.
209. Wang, X.; Yang, D.; Wang, Z.; Kwan, H.; Chen, J.; Wu, W.; Li, H.; Liao, Y.; Liu, S. Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Methodology. In Proceedings of the The Thirteenth International Conference on Learning Representations.
210. Lee, J.; Miyanishi, T.; Kurita, S.; Sakamoto, K.; Azuma, D.; Matsuo, Y.; Inoue, N. CityNav: A Large-Scale Dataset for Real-World Aerial Navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 5912–5922.
211. Sun, X.; Si, W.; Ni, W.; Li, Y.; Wu, D.; Xie, F.; Guan, R.; Xu, H.Y.; Ding, H.; Wu, Y.; et al. AutoFly: Vision-Language-Action Model for UAV Autonomous Navigation in the Wild. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
212. Chen, C.; Liang, S.; Guan, R.; Sun, X.; Zhao, H.; Jiang, H.; Huang, T.; Ding, H.; Han, Q.L. AerialMind: Towards Referring Multi-Object Tracking in UAV Scenarios. *arXiv preprint arXiv:2511.21053* **2025**.

 doi:10.20944/preprints202601.2155.v2

213. Sautenkov, O.; Yaqoot, Y.; Lykov, A.; Mustafa, M.A.; Tadevosyan, G.; Akhmetkazy, A.; Cabrera, M.A.; Martynov, M.; Karaf, S.; Tsetserukou, D. UAV-VLA: Vision-language-action system for large scale aerial mission generation. In Proceedings of the 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2025, pp. 1588–1592.
214. Fan, Y.; Chen, W.; Jiang, T.; Zhou, C.; Zhang, Y.; Wang, X. Aerial Vision-and-Dialog Navigation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 3043–3061.
215. Tran, K.T.; Dao, D.; Nguyen, M.D.; Pham, Q.V.; O'Sullivan, B.; Nguyen, H.D. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322* **2025**.
216. Feng, X.; Chen, Z.Y.; Qin, Y.; Lin, Y.; Chen, X.; Liu, Z.; Wen, J.R. Large Language Model-based Human-Agent Collaboration for Complex Task Solving. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 1336–1357.
217. Zou, H.P.; Huang, W.C.; Wu, Y.; Miao, C.; Li, D.; Liu, A.; Zhou, Y.; Chen, Y.; Zhang, W.; Li, Y.; et al. A Call for Collaborative Intelligence: Why Human-Agent Systems Should Precede AI Autonomy. *arXiv preprint arXiv:2506.09420* **2025**.
218. Fu, J.; Han, H.; Su, X.; Fan, C. Towards human-AI collaborative urban science research enabled by pre-trained large language models. *Urban Informatics* **2024**, *3*, 8.
219. Han, J.; Ning, Y.; Yuan, Z.; Ni, H.; Liu, F.; Lyu, T.; Liu, H. Large Language Model Powered Intelligent Urban Agents: Concepts, Capabilities, and Applications. *arXiv preprint arXiv:2507.00914* **2025**.
220. Wu, W.; He, H.; Zhang, C.; He, J.; Zhao, S.Z.; Gong, R.; Li, Q.; Zhou, B. Towards autonomous micromobility through scalable urban simulation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 27553–27563.
221. Zheng, Y.; Lin, Y.; Zhao, L.; Wu, T.; Jin, D.; Li, Y. Spatial planning of urban communities via deep reinforcement learning. *Nature Computational Science* **2023**, *3*, 748–762.
222. Ali, M.I.; Gao, F.; Mileo, A. Citybench: A configurable benchmark to evaluate rsp engines using smart city datasets. In Proceedings of the International semantic web conference. Springer, 2015, pp. 374–389.
223. Romeu-Guallart, P.; Zamora-Martinez, F. SML2010. *UCI Machine Learning Repository* **2014**.
224. Xu, H.; Yuan, J.; Zhou, A.; Xu, G.; Li, W.; Ban, X.; Ye, X. Genai-powered multi-agent paradigm for smart urban mobility: Opportunities and challenges for integrating large language models (llms) and retrieval-augmented generation (rag) with intelligent transportation systems. *arXiv preprint arXiv:2409.00494* **2024**.
225. Li, A.; Wang, Z.; Zhang, J.; Li, M.; Qi, Y.; Chen, Z.; Zhang, Z.; Wang, H. UrbanVLA: A Vision-Language-Action Model for Urban Micromobility. *arXiv preprint arXiv:2510.23576* **2025**.
226. Zhang, Z.; Chen, M.; Zhu, S.; Han, T.; Yu, Z. MMCNav: MLLM-empowered Multi-agent Collaboration for Outdoor Visual Language Navigation. In Proceedings of the Proceedings of the 2025 International Conference on Multimedia Retrieval, 2025, pp. 1767–1776.
227. Chen, W.; Yu, X.; Shang, L.; Xi, J.; Jin, B.; Zhao, S. Urban Emergency Rescue Based on Multi-Agent Collaborative Learning: Coordination Between Fire Engines and Traffic Lights. *arXiv preprint arXiv:2502.16131* **2025**.
228. Wang, X.; Yang, D.; Liao, Y.; Zheng, W.; Dai, B.; Li, H.; Liu, S.; et al. UAV-Flow Colosseo: A Real-World Benchmark for Flying-on-a-Word UAV Imitation Learning. *arXiv preprint arXiv:2505.15725* **2025**.
229. Jiang, K.; Cai, X.; Cui, Z.; Li, A.; Ren, Y.; Yu, H.; Yang, H.; Fu, D.; Wen, L.; Cai, P. Koma: Knowledge-driven multi-agent framework for autonomous driving with large language models. *IEEE Transactions on Intelligent Vehicles* **2024**.
230. Zheng, Y.; Xu, F.; Lin, Y.; Santi, P.; Ratti, C.; Wang, Q.R.; Li, Y. Urban planning in the era of large language models. *Nature Computational Science* **2025**, pp. 1–10.
231. Lin, Z.; Gao, K.; Wu, N.; Suganthan, P.N. Scheduling eight-phase urban traffic light problems via ensemble meta-heuristics and Q-learning based local search. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *24*, 14415–14426.
232. Ouyang, K.; Liang, Y.; Liu, Y.; Tong, Z.; Ruan, S.; Zheng, Y.; Rosenblum, D.S. Fine-grained urban flow inference. *IEEE Transactions on Knowledge and Data Engineering* **2020**, *34*, 2755–2770.
233. Chu, T.; Wang, J.; Codecà, L.; Li, Z. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE transactions on intelligent transportation systems* **2019**, *21*, 1086–1095.
234. Mouratidis, K. Time to challenge the 15-minute city: Seven pitfalls for sustainability, equity, livability, and spatial analysis. *Cities* **2024**, *153*, 105274.

 doi:10.20944/preprints202601.2155.v2

235. Creß, C.; Bing, Z.; Knoll, A.C. Intelligent transportation systems using roadside infrastructure: A literature survey. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *25*, 6309–6327.
236. Hamissi, A.; Dhraief, A. A survey on the unmanned aircraft system traffic management. *ACM Computing Surveys* **2023**, *56*, 1–37.
237. Ahmed, A.; Outay, F.; Farooq, M.U.; Saeed, S.; Adnan, M.; Ismail, M.A.; Qadir, A. Real-time road occupancy and traffic measurements using unmanned aerial vehicle and fundamental traffic flow diagrams. *Personal and Ubiquitous Computing* **2023**, *27*, 1669–1680.
238. Yu, X.; Wang, J.; Yang, Y.; Huang, Q.; Qu, K. BIGCity: A universal spatiotemporal model for unified trajectory and traffic state data analysis. In Proceedings of the 2025 IEEE 41st International Conference on Data Engineering (ICDE). IEEE, 2025, pp. 4455–4469.
239. Liu, A.; Zhang, Y. CrossST: An Efficient Pre-Training Framework for Cross-District Pattern Generalization in Urban Spatio-Temporal Forecasting. In Proceedings of the 2025 IEEE 41st International Conference on Data Engineering (ICDE). IEEE, 2025, pp. 2935–2948.
240. Perera, A.T.D.; Javanroodi, K.; Mauree, D.; Nik, V.M.; Florio, P.; Hong, T.; Chen, D. Challenges resulting from urban density and climate change for the EU energy transition. *Nature Energy* **2023**, *8*, 397–412.
241. Jin, X.; Zhang, C.; Xiao, F.; Li, A.; Miller, C. A review and reflection on open datasets of city-level building energy use and their applications. *Energy and Buildings* **2023**, *285*, 112911.
242. Wang, L.; Shao, J.; Ma, Y. Does China's low-carbon city pilot policy improve energy efficiency? *Energy* **2023**, *283*, 129048.
243. Lindahl, J.; Johansson, R.; Lingfors, D. Mapping of decentralised photovoltaic and solar thermal systems by remote sensing aerial imagery and deep machine learning for statistic generation. *Energy and AI* **2023**, *14*, 100300.
244. Gasparyan, H.A.; Davtyan, T.A.; Agaian, S.S. A novel framework for solar panel segmentation from remote sensing images: Utilizing Chebyshev transformer and hyperspectral decomposition. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–11.
245. Lodhi, M.K.; Tan, Y.; Wang, X.; Masum, S.M.; Nouman, K.M.; Ullah, N. Harnessing rooftop solar photovoltaic potential in Islamabad, Pakistan: A remote sensing and deep learning approach. *Energy* **2024**, *304*, 132256.
246. Golestani, Z.; Borna, R.; Khaliji, M.A.; Mohammadi, H.; Jafarpour Ghalehteimouri, K.; Asadian, F. Impact of urban expansion on the formation of urban heat islands in Isfahan, Iran: a satellite base analysis (1990–2019). *Journal of Geovisualization and Spatial Analysis* **2024**, *8*, 32.
247. Fan, X.; Ji, T.; Jiang, C.; Li, S.; Jin, S.; Song, S.; Wang, J.; Hong, B.; Chen, L.; Zheng, G.; et al. Mousi: Poly-visual-expert vision-language models. *arXiv preprint arXiv:2401.17221* **2024**.
248. Elgendy, H.; Sharshar, A.; Aboeitta, A.; Ashraf, Y.; Guizani, M. Geollava: Efficient fine-tuned vision-language models for temporal change detection in remote sensing. *arXiv preprint arXiv:2410.19552* **2024**.
249. Zhuo, L.; ZHANG, E.; Shuo, P.; Sichun, L.; Ying, L.; WITLOX, F. Assessing urban emergency medical services accessibility for older adults considering ambulance trafficability using a deep learning approach. *Sustainable Cities and Society* **2025**, p. 106804.
250. Li, J.; Wang, S.; Zhang, J.; Miao, H.; Zhang, J.; Yu, P.S. Fine-grained urban flow inference with incomplete data. *IEEE Transactions on Knowledge and Data Engineering* **2022**, *35*, 5851–5864.
251. Yang, M.; Li, X.; Xu, B.; Nie, X.; Zhao, M.; Zhang, C.; Zheng, Y.; Gong, Y. STDA: Spatio-Temporal Deviation Alignment Learning for Cross-city Fine-grained Urban Flow Inference. *IEEE Transactions on Knowledge and Data Engineering* **2025**.
252. Kennedy, J. Swarm intelligence. In *Handbook of nature-inspired and innovative computing: integrating classical models with emerging technologies*; Springer, 2006; pp. 187–219.
253. Han, X.; Zhu, C.; Zhu, H.; Zhao, X. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In Proceedings of the Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, 2025, pp. 814–825.
254. Chen, S.; Chen, Y.; Pan, C.; Ali, I.; Pan, J.; He, W. Distributed adaptive platoon secure control on unmanned vehicles system for lane change under compound attacks. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *24*, 12637–12647.
255. Gao, C.; Xu, F.; Chen, X.; Wang, X.; He, X.; Li, Y. Simulating human society with large language model agents: City, social media, and economic system. In Proceedings of the Companion Proceedings of the ACM Web Conference 2024, 2024, pp. 1290–1293.

 doi:10.20944/preprints202601.2155.v2

256. Liu, Y.; Zhang, X.; Ding, J.; Xi, Y.; Li, Y. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In Proceedings of the Proceedings of the ACM web conference 2023, 2023, pp. 4150–4160.
257. Akhtar, Z.; Qazi, U.; Sadiq, R.; El-Sakka, A.; Sajjad, M.; Ofli, F.; Imran, M. Mapping Flood exposure, damage, and Population needs using remote and social sensing: a case study of 2022 Pakistan floods. In Proceedings of the Proceedings of the ACM Web Conference 2023, 2023, pp. 4120–4128.
258. Zhang, S.; Li, J.; Shi, L.; Ding, M.; Nguyen, D.C.; Tan, W.; Weng, J.; Han, Z. Federated learning in intelligent transportation systems: Recent applications and open problems. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *25*, 3259–3285.
259. Yan, A.; Howe, B. Fairness-aware demand prediction for new mobility. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, pp. 1079–1087.
260. Wang, G.; Zhang, Y.; Fang, Z.; Wang, S.; Zhang, F.; Zhang, D. FairCharge: A data-driven fairness-aware charging recommendation system for large-scale electric taxi fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **2020**, *4*, 1–25.
261. Rong, C.; Feng, J.; Ding, J. Goddag: Generating origin-destination flow for new cities via domain adversarial training. *IEEE Transactions on Knowledge and Data Engineering* **2023**, *35*, 10048–10057.
262. Zhou, Q.; Wu, J.; Zhu, M.; Zhou, Y.; Xiao, F.; Zhang, Y. LLM-QL: a LLM-Enhanced Q-Learning Approach for Scheduling Multiple Parallel Drones. *IEEE Transactions on Knowledge and Data Engineering* **2025**.