From Terrain to Space: A Survey on Multi-Domain Data Lifecycle for Urban Embodied Agents

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

Posted Date: 28 January 2026

doi: 10.20944/preprints202601.2155.v1

Keywords: data lifecycle; urban computing; embodied agent; multimodal data; data fusion

Article

From Terrain to Space: A Survey on Multi-Domain Data Lifecycle for Urban Embodied Agents

Penglei Sun †, Song Tang †, Jiawen Wen, Yuxuan Liang *, Yang Yang, Xiaowen Chu *

The authors are with Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
* Correspondence: yuxuanliang@hkust-gz.edu.cn (Y.L.); xwchu@hkust-gz.edu.cn (X.C.)
† Equal contribution.

Abstract

Urban Embodied Agents (UrbanEAs) are emerging to interact with complex, large-scale city environments, generating vast, heterogeneous data streams. While embodied agent research has focused on controlled indoor environments, these settings lack the complexity of the physical world. In contrast, urban environments present distinct challenges, including environmental variability, limited observability, and interaction complexity. These challenges hinder the effectiveness of conventional agents. Therefore, establishing a comprehensive data lifecycle to fuse multi-domain data from terrain, aerial, and space is an essential strategy for developing actionable embodied capabilities from raw urban streams. Distinct from existing surveys that follow a model-centric paradigm for urban computing, we systematically propose and review a comprehensive Data Lifecycle from a multi-domain data perspective, which is essential for the UrbanEA. First, we propose a unified framework containing four key stages of this lifecycle: Data Perception, Data Management, Data Fusion, and Task Application. Next, we establish a taxonomy for each stage of the lifecycle. Finally, we outline the social impact of the data lifecycle of UrbanEA and open research problems. Our survey provides a rigorous roadmap for designing the robust, high-performance data frameworks essential for these UrbanEAs.

Keywords: data lifecycle; urban computing; embodied agent; multimodal data; data fusion

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 urban problems, including delivery and transportation. Compared to conventional urban agents, the core distinction lies in this embodiment: conventional agents may only analyze and predict within digital space (e.g., traffic flow models), whereas an UrbanEA uses its physical form to perceive and execute actions, interacting with the world directly [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 multiple domains (handheld, vehicle, drone, satellite) to comprehensively perceive the physical world, marking the starting point of the data lifecycle.

 doi:10.20944/preprints202601.2155.v1

2. **Data Management**. Task-driven storage architectures (such as graph, vector, and spatio-temporal databases) organize and query massive, heterogeneous urban perception data, laying a solid foundation for subsequent fusion and interaction.
3. **Data Fusion**. In this stage, the fusion strategies address the data gaps and construct a unified urban cognition.
4. **Task Application**. The structured data, after being processed through fusion, 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.

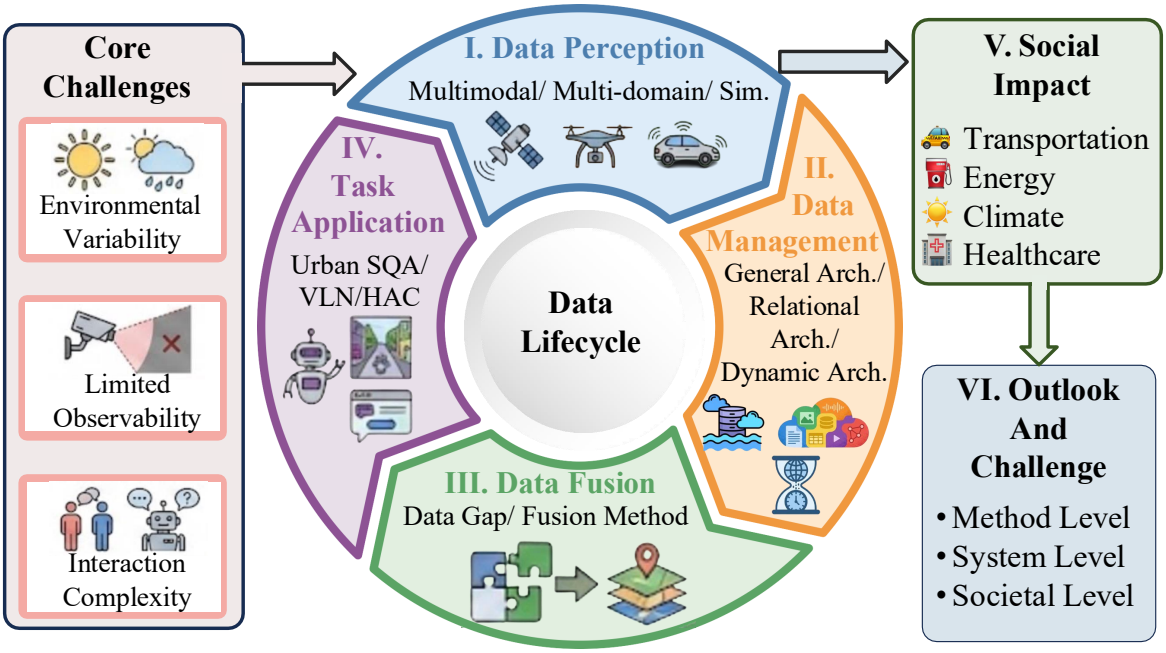

Figure 1. Overview of the proposed multi-domain data lifecycle for Urban Embodied Agents, from perception to social impact, where Sim. denotes the simulator and Arch. denotes the Architecture.

Despite its importance, successfully implementing this end-to-end data lifecycle in urban settings remains insufficient. Existing embodied agent research focuses on indoor environments [8], 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 sensor (e.g., a vehicle-mounted camera or LiDAR) suffers from blind spots and occlusions due to its limited Field of View (FoV) and detection range. This results in spatially incomplete perceptual information, making it impossible to achieve a globally consistent scene understanding during **Data Fusion**.
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, pedestrian 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 Fusion** and **Task Application** stages than simple physical motion.

 doi:10.20944/preprints202601.2155.v1

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 multi-domain data for UrbanEA has emerged as a promising strategy [13,15]. 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 [10,12] or the graph neural network [9], which may not fully align with the requirements of this multi-domain approach.

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

Survey	Year	Venue	City Platform	Multi-domain Data	Embodied Agent	Data Life-cycle	Primary Perspective
Xu et al. [6]	2023	arXiv	✓				Model-centric
Jin et al. [9]	2023	IEEE TKDE	✓				Model-centric
Yang et al. [3]	2024	IEEE TKDE	✓	✓			Model-centric
Zhang et al. [10]	2024	ACM KDD	✓	✓			Model-centric
Cengiz et al. [4]	2025	Information Fusion	✓	✓			Model-centric
Lu et al. [11]	2025	arxiv		✓	✓		Data-centric
Liang et al. [12]	2025	ACM KDD	✓			✓	Model-centric
Zou et al. [13]	2025	Information Fusion	✓	✓			Model-centric
Song et al. [14]	2025	IEEE TKDE		✓			Model-centric
Our Work	-	-	✓	✓	✓	✓	Data-centric

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 spatio-temporal 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, Fusion, 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 fusion strategies based on the specific Domain Gaps (Representation, Quality, Spatio-temporal, and Semantic) 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.

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 fusion techniques to bridge key multi-domain 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.

 doi:10.20944/preprints202601.2155.v1

2. Data Perception: Sensing and Simulation

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. The multimodal 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). 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.

• **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.

 doi:10.20944/preprints202601.2155.v1

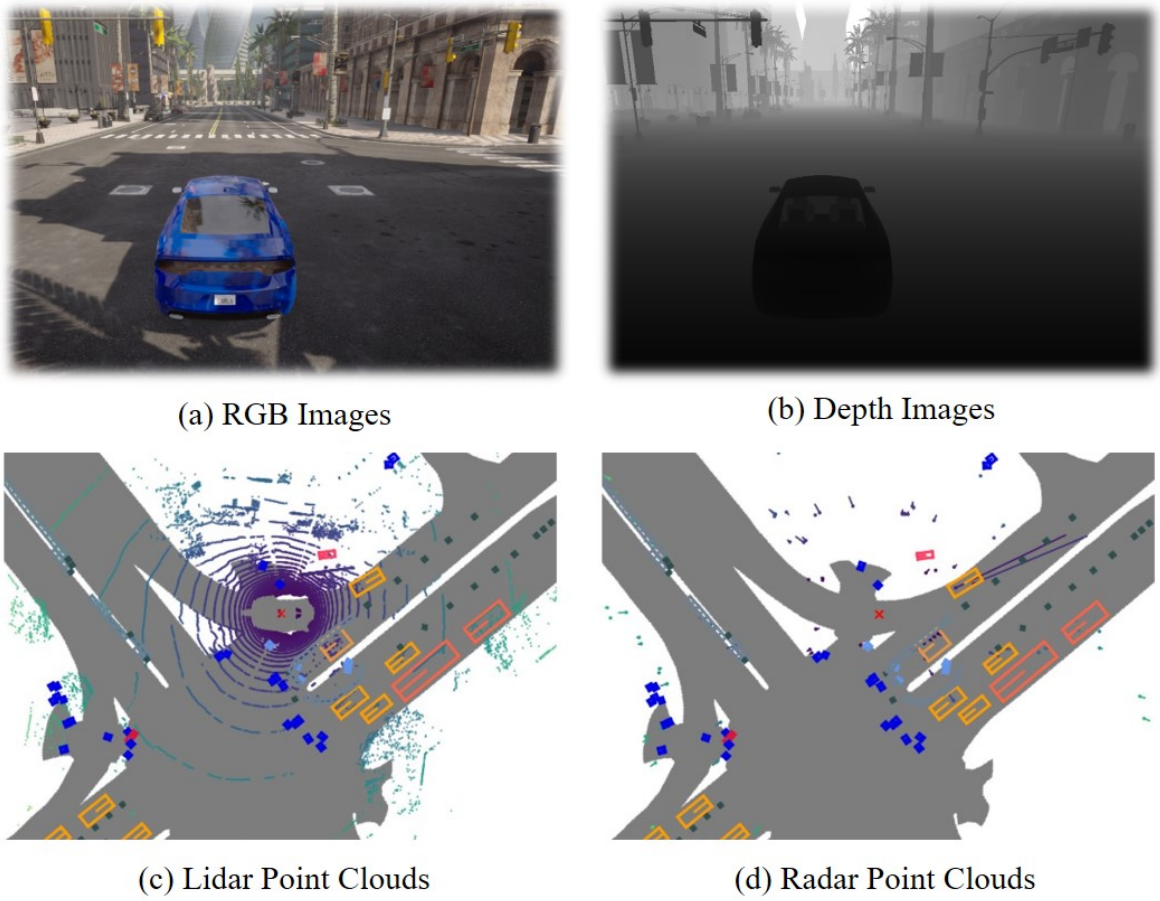

Figure 2. Vision perception for Urban Embodied Agents. We visualize some cases in Carla Simulators [16].

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 [17] and public security [18,19]. Sensing within environments is multi-sensory, extending well beyond the visual sensing [20]. 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 [21–25]. The domain of visual-tactile learning focuses on building realistic tactile simulation systems to allow agents to understand the world through physical interaction [26–30]. Multiply [31] 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 [13,32,33]. 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 spatio-temporal gap as discussed in Section 4.1.

 doi:10.20944/preprints202601.2155.v1

	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 multi-domain data in Urban Embodied Agents.

• **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 [34]. 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 Quality Gap discussed in Sec 4.1.

• **Vehicle.** These systems are designed for efficient corridor mapping and possess a range optimized for capturing roadside features from a moving vehicle [35,36]. 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 Spatio-Temporal Gap as discussed in Sec 4.1.

• **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 [37]. 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 [38]. 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 [39,40]. 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 Density gap.

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 [41]. Existing indoor simulators [42–45] 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 environments present challenges for real-world experimentation due to their complexity, dynamic

 doi:10.20944/preprints202601.2155.v1

nature, and safety concerns. Therefore, as shown in Figure 4, 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 [46]	2016	✗	Terrain	Open-Loop	✓	✗	✗	✗	Street View	-
CARLA [16]	2017	✓	Terrain	Closed-Loop	✓	✓	✓	✓	Vehicle	UE 4
xView [47]	2018	✗	Aviation	Open-Loop	✓	✗	✗	✗	Satellite	-
TouchDown [48]	2019	✗	Terrain	Open-Loop	✓	✗	✗	✗	Street View	-
Nuscenes [49]	2020	✗	Terrain	Open-Loop	✓	✗	✓	✓	Vehicle	Nuscenes-Kit
Waymo [50]	2020	✗	Terrain	Open-Loop	✓	✗	✓	✓	Vehicle	Waymax
KITTI-360 [51]	2022	✗	Terrain	Open-Loop	✓	✗	✗	✓	Vehicle	-
STPLS3D [52]	2022	✗	Aviation	Open-Loop	✗	✗	✗	✓	Drone	-
SensatUrban [53]	2022	✗	Aviation	Open-Loop	✗	✗	✗	✓	Drone	-
UrbanBIS [54]	2023	✗	Aviation	Open-Loop	✓	✗	✗	✓	Drone	-
AerialVLN [55]	2023	✓	Aviation	Open-Loop	✓	✓	✗	✗	Drone	UE 4
GRUTopia [56]	2024	✓	Terrain	Closed-Loop	✓	✗	✗	✗	Virtuality	Isaac Sim
OpenUAV [57]	2024	✓	Aviation	Open-Loop	✓	✓	✗	✗	Drone	UE 4
UnrealZoo [58]	2024	✓	Terrain	Closed-Loop	✓	✗	✗	✗	Virtuality	UE 4/5
MetaUrban [59]	2025	✓	Terrain	Closed-Loop	✓	✓	✗	✓	Virtuality	Gym
OpenFly [60]	2025	✓	Aviation	Closed-Loop	✓	✓	✗	✓	Drone	UE 4, Google Earth, GTA V

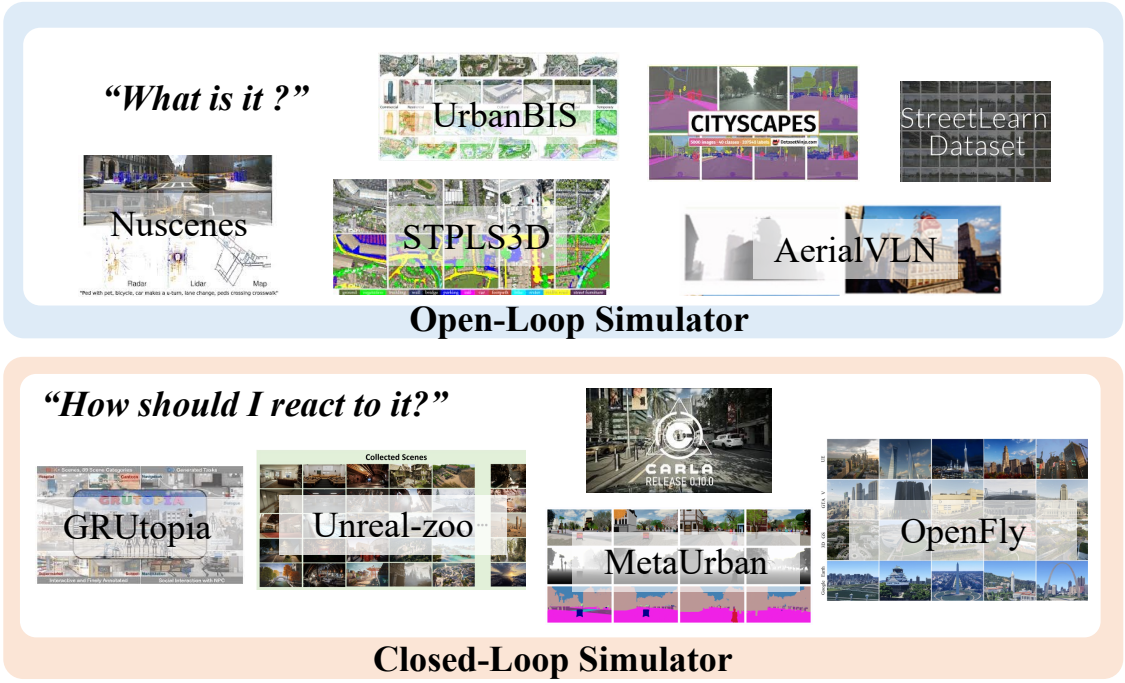

Figure 4. Existing simulators for Urban Embodied Agents. They are classified into open-loop and closed-loop simulators.

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

 doi:10.20944/preprints202601.2155.v1

domains, such as StreetLearn [61], Cityscapes [46], xView [47], SensatUrban [53], UrbanBIS [54], and STPLS3D [52]. 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 [49], Waymo [50], KITTI-360 [51], AerialVLN [55].

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 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 [62]. For terrain-based systems, including GRUTopia [56], UnrealZoo [58], and MetaUrban [59], 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 [57], and OpenFly [60] 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. 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: Storage and Querying

UrbanEAs operate in a data-intensive environment as detailed in Section 2. However, this raw data deluge is difficult to use directly for the downstream embodied agent task. For instance, data fusion can fail with misaligned timestamps, semantic relationships between objects remain implicit, and safety-critical tasks may lack the necessary data freshness and consistency. The fundamental challenge is thus to transform this unstructured data torrent into a well-organized and queryable knowledge base that supports both real-time perception and long-term learning [63,64].

Effective data management provides the structured backbone required for robust fusion (Section 4), real-time task execution (Section 5), and auditable systems that respect social considerations (Section 6). Navigating the inherent trade-offs between latency, scale, and query complexity under the data lifecycle challenges (Section 1) requires specialized solutions. We therefore examine six complementary storage architectures that form the basis of modern urban data systems (summarized in Table 3): Data Lakes, Multi-model and Graph Databases, Vector Databases, Time-Series Databases, and Spatio-temporal Databases.

We illustrate their respective roles using a running example in this section, as shown in Figure 5: an autonomous vehicle (ego-agent) approaching a busy, rainy intersection. Its view is partially occluded by a large truck, while a roadside unit (RSU) detects an ambulance coming from the blind spot. This scenario, involving sensor noise (rain), occlusions (truck), and multi-source asynchronous data (vehicle

 doi:10.20944/preprints202601.2155.v1

sensors vs. RSU), requires a sophisticated data management strategy to ensure safety. The following subsections will examine how various database paradigms address these specific issues.

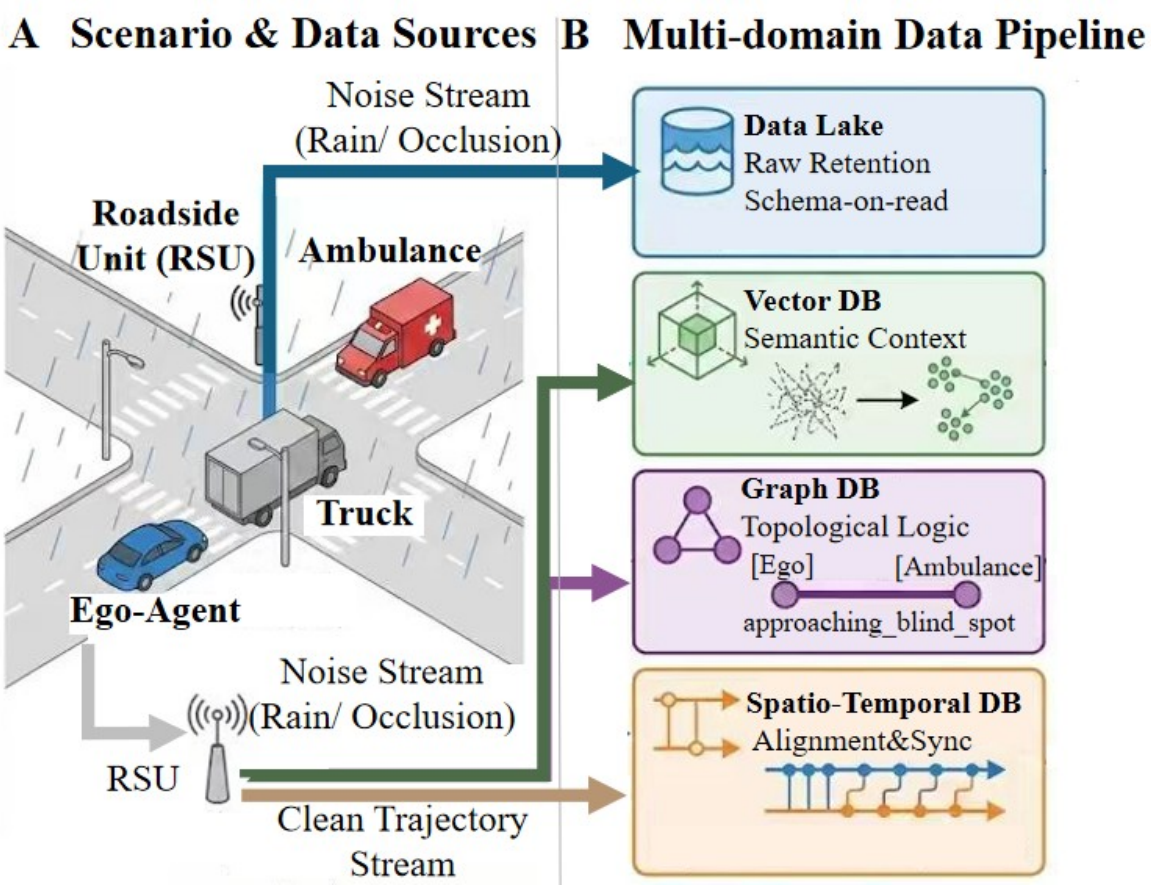

Figure 5. The example using DBs for the Urban Embodied Agents.

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. [65], Kappa Arch. [66], Lakehouse [67]
Multi-model DBs	Unified Model	✓	✓	✓	✓	Variable	Medium	Medium	Sinew [68], NoAM [69], UniBench [70]
Graph DBs	Nodes & Edges	✗	✓	✗	✗	Low	Low	High	Neo4j [71], JanusGraph [72], nSKG [73], Sg-CityU [74]
Vector DBs	High-dim Vectors	✗	✗	✓	✗	Low	Low	High	FAISS [75], Milvus [76], HNSW [77], PQ [78]
Time-Series DBs	Time/Value Pairs	✗	✗	✗	✓	Low	High	High	Gorilla [79], Apache IoTDB [80], TimescaleDB [81]
Spatio-Temporal DBs	Time/Spatial Info	✗	✓	✗	✓	Variable	Medium	High	MobilityDB [82], PostGIS [83], TrajMesa [84],TMan [85]

3.1. General & Unified Architectures

• **Data Lakes**. The evolution of UrbanEA data management began with large-scale autonomous driving datasets such as NuScenes [49] and Waymo Open Dataset [50]. These datasets established a baseline practice: organizing multimodal streams (cameras, LiDAR, radar) via relational metadata while storing high-volume unstructured data (images, point clouds) as separate files linked by unique identifiers. NuScenes exemplifies this paradigm by providing temporal indexing for synchronized sensor data, global identifiers enabling cross-modal annotation ("annotate 3D once, project to 2D everywhere"), and standardized coordinate transformation APIs for multimodal fusion. For example, the complete transformation chain from LiDAR to camera frame is formalized as:

$$T_{\text{cam}\leftarrow\text{lidar}} = (T_{\text{ego}\leftarrow\text{cam}})^{-1} \cdot T_{\text{ego}\leftarrow\text{lidar}} \tag{1}$$

 doi:10.20944/preprints202601.2155.v1

Then, the projection of a LiDAR point P_{lidar} onto the image plane can be expressed as:

$$p_{\text{homo}} = K \cdot \text{proj}(T_{\text{cam}\leftarrow\text{lidar}} \cdot P_{\text{lidar}}) \tag{2}$$

where K is the camera intrinsic matrix, T represents the rigid body transformation matrices, and $\text{proj}(\cdot)$ denotes taking the first three dimensions of the homogeneous coordinate. Every matrix in these equations is directly supplied by the dataset management system [49].

However, while these datasets excel at providing synchronized data for offline training, they are fundamentally limited to static, post-collection dissemination and cannot support the real-time, streaming requirements of production UrbanEAs systems. This limitation motivated the development of Data Lake architectures [86,87], which provide a principled framework for managing continuous, heterogeneous data streams from city-scale platforms. Unlike the static datasets above, Data Lakes adopt a "schema-on-read" philosophy: raw data is ingested in its native format without predefined schemas, with structure imposed only at query time [67,86]. This design fundamentally shifts the paradigm from "publish once" to "continuously sensing and govern."

The architectural evolution of Data Lakes reflects urban escalating demands for real-time capabilities. The Lambda Architecture [65,88] was designed with parallel batch and speed layers to concurrently support both historical analytics and real-time updates, which in turn allows agents to utilize archival data for training while also accessing live streams for immediate decision-making. Kappa Architecture [66] simplified this by focusing solely on stream processing, optimizing for the low-latency responsiveness critical to reactive UrbanEA. More advanced Zone-Based Governance models, including Data-Pond [89] and Functional Data Lake [90,91], organize data into logical zones (Raw $\rightarrow$ Curated $\rightarrow$ Analytics) with progressive quality control, ensuring that agents access appropriately validated data for different tasks.

The Data Lakehouse [67] represents the latest development, integrating the flexibility of data lakes with the robust data management and transactional guarantees of data warehouses, thus providing a promising solution for UrbanEA systems that require both exploratory analytics and reliable operational queries. For city setting, Data Lakes offer three strategic advantages: (1) *maximized raw data retention* with full lineage tracking, enabling reprocessing as models evolve; (2) *real-time streaming ingestion* via architectures like Kappa, supporting continuous UrbanEAs; (3) *multi-zone governance* that enforces data quality standards across the pipeline from raw sensors to agent decision-making. In a rainy urban intersection, multimodal streams typically include multi-camera images, LiDAR point clouds, ego-state signals, and roadside trajectories. A data lake retains these raw assets, while an accompanying catalog captures calibration, timestamps, and lineage. This schema-on-read design preserves weather artefacts and occlusion evidence for later fusion and auditing without imposing premature structure.

• **Multi-model Databases**. UrbanEAs generate data with fundamentally different structures: structured metadata (relational), spatial relationships (graph), and sensor streams (time-series). Multi-model databases [92,93] offer a unified approach to the heterogeneity problem by natively supporting diverse data models within one system, a capability that data lakes and other single-model architectures only partially provide.

Based on their architectural design, multi-model systems are generally classified into two main categories [94,95]. The first, Polyglot Persistence, involves dedicating separate, specialized database systems to each data type (such as PostgreSQL for metadata, Neo4j for scene graphs, and InfluxDB for sensor streams). While this approach optimizes performance for individual modalities, it introduces significant complexity for holistic tasks that depend on cross-database queries and data integration [96]. Conversely, the second category, Unified Multi-Model Databases [97], integrates various data models within a single instance. This design enables seamless cross-model queries to be executed within a unified transactional context. Such a paradigm is highly suitable for the consolidated data management needs of urban applications; a query like "retrieve all vehicles (relational) that were spatially near (graph) the incident location during the last 5 minutes (time-series)" can be processed as a single atomic

 doi:10.20944/preprints202601.2155.v1

operation, eliminating the need for cross-system coordination. In the rainy intersection scenario, such unified queries can simultaneously express occlusion and priority constraints while returning the associated sensor slices for the relevant time window.

For UrbanEAs, the most significant advantage of multi-model databases is their capacity to *unify heterogeneous data management*. This unification, in turn, simplifies data pipelines and reduces the overhead associated with schema conversion. These systems can also strike a balance between low-latency responses for real-time sensing and high-throughput processing for offline analysis by employing diverse indexing strategies tailored to each specific data model [98]. Notable examples include Sinew [68], which merges document and graph models to achieve flexible schema evolution; NoAM [69], which is tailored for aggregation-oriented NoSQL workloads; and UniBench [70], a comprehensive benchmark used to evaluate the performance of multi-model databases across varied query patterns.

3.2. Semantic & Relational Architectures

• **Graph Databases**. Beyond data lakes and multi-model paradigms, traditional data models struggle to natively express the complex relationships that govern urban systems [99–101]. This limitation manifests as the Semantic Gap, where raw data lacks explicit relational context for high-level reasoning [102]. To bridge this gap, graph-based architectures leverage graph databases [103–106] and graph neural networks to model entity relationships as graphs $G = (V, E)$, where nodes represent entities and edges encode their relationships [107], such as modeling the 'occludes' relationship between the vehicle and the truck. This provides the explicit relational context needed for downstream reasoning tasks like scene question answering. In the intersection scenario, perception outputs are materialized into a temporally indexed scene graph where nodes represent the ego vehicle, the truck, the ambulance, lanes and signals, and edges encode relations such as occludes, located_on_lane, has_priority, and approaching_from_blind_spot. Each relation carries the originating timestamp and pointers to the source frames in the lake, grounding semantic reasoning back to raw data without ambiguity. In UrbanEA, two distinct yet complementary graph representation paradigms have emerged, each serving different scales and purposes. The first paradigm, **scene graphs**, follows a bottom-up, task-specific approach. Scene graphs are typically generated in an end-to-end manner directly from sensor inputs (e.g., multi-view images or LiDAR scans) to capture the immediate perceptual context of a scene [74,108]. For example, Sg-CityU [74] constructs urban scene graphs that decompose complex 3D environments into structured object-centric representations with spatial relationships, simplifying downstream tasks like spatial reasoning and navigation. T2SG (Traffic Topology Scene Graph) [108] specializes in modeling road-level topology for autonomous driving: it represents lanes as nodes and their connectivity (e.g., predecessor, successor) and control relationships (e.g., governed by a specific traffic light) as edges. The second paradigm, knowledge graphs (KGs), adopts a top-down, ontology-driven approach aimed at building comprehensive, reusable knowledge bases for entire urban domains [107,109,110]. A landmark example is nuScenes Knowledge Graph (nSKG) [73], which transforms the nuScenes autonomous driving dataset into a structured KG comprising approximately 43 million RDF triples. nSKG is constructed by first defining formal ontologies, and then systematically extracting and mapping instance data from the dataset's annotations into this ontological framework [73]. This dichotomy reflects a fundamental trade-off: scene graphs prioritize speed, task-relevance, and tight integration with perception models, making them suitable for real-time, tactical decision-making; KGs prioritize completeness, semantic richness, and long-term knowledge persistence, making them ideal for strategic reasoning, offline simulation, and cross-domain knowledge integration [111].

Beyond static storage, graph-based architectures enable advanced reasoning through graph neural networks (GNNs) [112]. GNNs operate by iteratively propagating and aggregating information across graph edges, where each node updates its representation by combining information from its neighbors through learnable transformations, enabling the model to reason about both local interactions and global context [9]. In urban applications, (spatio-temporal graph neural networks) STGNNs extend this by jointly modeling spatial topology (e.g., road network connectivity) and temporal dynamics (e.g.,

 doi:10.20944/preprints202601.2155.v1

traffic flow evolution), making them indispensable for predictive tasks such as trajectory forecasting and traffic prediction [9]. A representative application is SemanticFormer [113], which leverages the rich semantic context provided by nSKG to perform multi-modal trajectory prediction. SemanticFormer employs a hierarchical heterogeneous graph encoder that captures interactions between agents, road elements, and traffic rules by applying attention mechanisms over *meta-paths*—semantically meaningful sequences of relations (e.g., vehicle $\rightarrow$ *on_lane* $\rightarrow$ lane $\rightarrow$ *governed_by* $\rightarrow$ traffic_light).

From a systems perspective, graph databases such as Neo4j [71] and JanusGraph [72] provide the infrastructure for storing and querying large-scale urban KGs using declarative languages like SPARQL and Cypher. These systems support multi-hop traversal queries essential for tasks like multi-step path planning (e.g., "*Find all lanes connected to the current lane that are not controlled by a red light*") [105]. In practice, hybrid architectures are increasingly common: structured metadata and time-series data are stored in relational or time-series databases (as discussed previously), while complex relational and semantic information is offloaded to graph databases, with both systems queried in a federated manner [107]. Real-world applications demonstrate the power of this paradigm: Urban Region Graph [99] uses graph-based spatial representations for urban village detection; knowledge-driven site selection systems like KnowSite [109] leverage urban KGs to recommend optimal commercial locations by reasoning over multi-hop relationships between venues, demographics, and transportation networks, outperforming purely data-driven approaches [107].

• **Vector Databases**. Another limitation of relational models is their reliance on exact, token-based queries, making them ill-suited for vague, natural language queries [85,114]. Vector databases address this by encoding multimodal data into high-dimensional vectors and leveraging Approximate Nearest Neighbor (ANN) search algorithms [75,77,78,115] for rapid, content-based semantic retrieval. The core operation is to measure the semantic similarity between a query vector Q and data vectors D_i in the database, often using cosine similarity [116]:

$$\text{sim}(Q, D_i) = \frac{Q \cdot D_i}{\|Q\|\|D_i\|} \tag{3}$$

The system returns items with the highest similarity scores [117]. In the intersection, retrieving similar rainy intersection segments helps interpret degraded visuals when occlusion and glare lower confidence. This capability is crucial for advanced human-agent interaction, particularly for supporting the Scene Question Answering (SQA) tasks in section V. The demand for such advanced interactive tasks, exemplified by work like NuScenes-QA [118], has spurred the need for vector databases, making it possible to query vast visual scenes using natural language [74,118–120].

3.3. Spatio-temporal & Dynamic Architectures

• **Time-Series Databases**. In data of UrbanEAs, time-series databases serve two primary roles [121]. First, they manage high-frequency sensor data streams from embodied agents. For instance, in autonomous driving research, vehicles generate continuous streams of IMU data (at 100+ Hz), GPS trajectories (at 10 Hz), LiDAR scans (at 10-20 Hz), and vehicle dynamics (velocity, acceleration, steering angle at 50+ Hz). These streams are ingested into time-series databases for real-time monitoring, offline analysis, and model training [49,50]. Systems like Gorilla [79], originally developed by Facebook for operational monitoring, demonstrate the capability to handle billions of data points per day with high write throughput and efficient compression, making them suitable for city-scale deployments. Similarly, Apache IoTDB [80] provides a unified time-series database specifically designed for IoT applications, with native support for irregular sampling intervals and multi-tenant isolation—features particularly valuable for managing heterogeneous sensor deployments across different urban districts or infrastructure operators. CrocodileDB [122] introduces resource-efficient query execution by exploiting temporal slackness in query deadlines, enabling graceful performance degradation under resource constraints—a critical capability for edge deployment scenarios in UrbanEA systems where computational resources may be limited. Second, in the context of interactive and reinforcement

 doi:10.20944/preprints202601.2155.v1

learning environments, time-series databases store the complete decision-making history of agents. For example, in EmbodiedCity [123], an agent's trajectory, observations, actions, and rewards over an entire episode are logged as time-series data, enabling replay, debugging, and offline reinforcement learning. Similarly, in multi-agent driving scenarios studied in DriveLM [120], the synchronized time-series logs of all agents' states and decisions are crucial for understanding emergent behaviors and training coordination strategies.

Beyond basic storage and retrieval, time-series databases increasingly support advanced analytics directly within the database engine. For instance, TimescaleDB [81] provides built-in functions for continuous aggregation (e.g., computing moving averages in real-time as data arrives), time-bucketing (grouping data into regular intervals for analysis), and hyperfunctions (domain-specific aggregations like percentile estimation over time). Around urgent braking events at the intersection, event-centric windows organize ego-state streams for fusion and audit.

• **Spatio-Temporal Databases**. In data of UrbanEAs, 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 [124–127]. A primary role is enabling hybrid systems. Recent empirical studies validate this approach by integrating PostgreSQL with extensions like PostGIS (for spatial capabilities) and Timescale (for temporal optimization) [128,129]. This combination demonstrates substantial query time reductions for both stationary (e.g., parking monitoring) and non-stationary (e.g., railway track monitoring) sensor data. Crucially, these studies reveal that for trajectory data, lightweight BRIN (Block Range Index) indexes can outperform traditional R-trees in both query performance and storage overhead, challenging conventional wisdom in spatial indexing. A second role is trajectory and mobility data management. Urban trajectories (vehicles, drones, pedestrians) are spatio-temporal objects. Specialized systems like MobilityDB [82] extend databases with native support for moving objects, enabling queries such as *"find all vehicles that passed through region R between time t_1 and t_2"* or *"compute the speed profile of agent A over the last hour,"* which is foundational for traffic analysis and behavior mining [130]. More recently, the integration with machine learning frameworks has enabled in-database prediction [12,131–133]. The emergence of foundation models for Spatio-Temporal Data represents a paradigm shift, allowing pre-trained models to be deployed directly on spatio-temporal data streams for real-time forecasting (e.g., traffic prediction, demand estimation) without moving data to external systems. In the intersection scenario, alignment operators synchronize the RSU ambulance trajectory with the ego timeline and unify coordinate frames, producing per-timestamp bundles ready for fusion.

3.4. Discussion

Effective data management in urban embodied systems transforms raw, heterogeneous sensor streams into actionable knowledge through layered architectures. It begins with Data Lakes that retain raw assets and lineage, continues with Graph Databases that render critical relations explicit and queryable, and relies on Spatio-temporal Databases to synchronize heterogeneous streams in time and space while unifying coordinate frames. Together, these components produce coherent, analysis-ready data for fusion and downstream tasks.

We observe two trends. The first is an evolution from static datasets to dynamic data platforms. The future paradigm of data management will shift from publishing one-off, static datasets to building dynamic, open data ecosystems where the community can continuously upload, annotate, query, and simulate, thereby accelerating iteration across the entire field [42]. The second is the concept of data as a queryable world model. When data management and querying technologies reach their zenith, the managed, structured data asset itself constitutes an implicit, inferable world model, blurring the lines between raw data, information, and knowledge. This offers a novel and exciting avenue for building more general and powerful UrbanEA [32,134].

 doi:10.20944/preprints202601.2155.v1

4. Data Fusion: Bridging Domain Gaps

Data in UrbanEA is beyond the passive feed from a single domain, but an active perception combining multi-domain information. This combination is essential, acting like the assembly of “puzzle pieces" which resolves inconsistencies between sources and fills the blind spots of individual sensors. Only through this process can a scattered collection of data become a coherent and complete urban “tapestry". Without effective fusion, the UrbanEA is left with a collection of scattered, independent, and potentially contradictory perceptual puzzle pieces.

However, this fusion process presents challenges, stemming from the heterogeneity of multi-domain data [13,14]. While the previous section detailed the preparation of each “puzzle piece," including the preprocessing and management of individual data streams, this section explores how to stitch them together into a unified understanding [135]. We will begin by analyzing the four core “gaps" that impede data fusion: at the data level, the Representation Gap and the Quality & Density Gap; at the spatio-temporal dimension, the Spatio-Temporal Gap; and at the cognitive level, the Semantic Gap. After defining these challenges, we will introduce the primary fusion strategies developed to bridge them. These include Early Fusion, Feature-Level Fusion, Late Fusion, and Prior-Based Fusion, which will be illustrated with state-of-the-art research examples [136].

4.1. The Domain Gap between Multi-Domain Data

The challenges in fusing multi-domain data can be analyzed from four perspectives: data structure, data quality, spatio-temporal basis, and information level. These perspectives reveal four obstacles, or “gaps," that must be addressed: the Representation Gap, the Quality & Density Gap, the Spatio-temporal Gap, and the Semantic Gap [137], as shown in Figure 6.

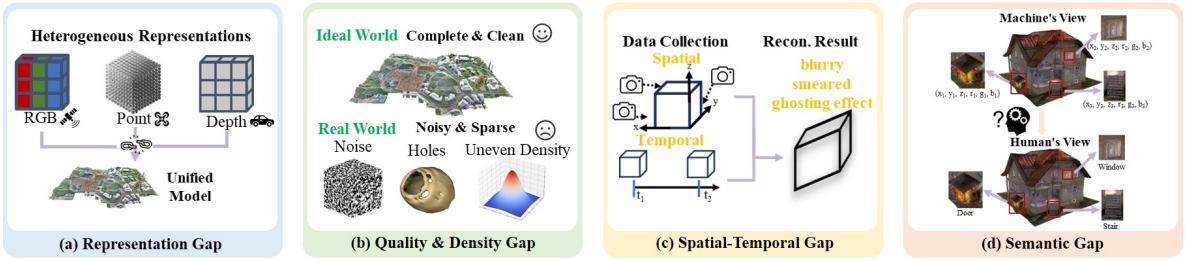

Figure 6. The domain gap between multi-domain data for Urban Embodied Agents.

• **Representation Gap**. The representation gap refers to the fundamental differences in the most basic data structures, mathematical representations, and organizational methods across various data domains, as shown in Figure 6 (a). For example, satellite images are regular 2D grids composed of pixels, while point clouds are sparse, unordered sets of 3D points in a Cartesian coordinate system [138]. Meanwhile, 3D meshes are frequent in vehicles. They consist of vertices, edges, and faces, and implicit representations are continuous field functions [139]. These data structures are not compatible, requiring algorithms to incorporate specialized modules to process and convert these heterogeneous representations for effective fusion [140].

• **Quality & Density Gap**. The quality & density gap, as shown in Figure 6 (b), describes the inconsistencies across different data domains in terms of their precision, completeness, noise levels, and data point density, as well as the disparity between real-world collected data and idealized (e.g., synthetic) data [141]. This gap is manifested in several ways. For instance, point clouds from low-altitude domain offer high geometric precision, and the reconstruction in the ideal world needs complete and clean data. However, in the real-world data collection, they face challenges such as noise, holes, and uneven density among domains [142,143]. While image data from vehicle is dense, its quality degrades under complex illumination conditions (such as overexposure, shadows, and reflections), and it cannot directly provide geometric information [144–146]. Furthermore, the artifacts and imperfections common in real-world data create a “sim-to-real" gap when compared to clean, idealized synthetic training data [147–149].

 doi:10.20944/preprints202601.2155.v1

• **Spatio-Temporal Gap**. The spatio-temporal gap refers to the difficulty of aligning data from different sensors and different points in time within a unified spatial coordinate system and temporal sequence, as shown in Figure 6 (c). This gap has two aspects. Spatially, the coordinate systems of different domains are difficult to unify perfectly. The perception algorithms are sensitive to accurate sensor parameters, and even minor pose errors can severely impact quality [150,151]. Temporally, because urban scenes contain moving objects (like vehicles), the content of data collected at different moments will be inconsistent, leading to deformative and elongated artifacts in the final projection [152].
• **Semantic Gap**. The semantic gap refers to the cognitive disparity between low-level, raw sensory data (such as pixels and points) and high-level, human-understandable scene semantics (such as object categories, functional attributes, and spatial relationships), as shown in Figure 6 (d). [153]. The algorithms can directly process pixel colors and 3D coordinates (x, y, z, r, g, b), but they cannot understand concepts like “the window", “the door", “the stair". To bridge this gap, researchers introduce techniques like semantic segmentation and language models as semantic priors [154,155]. This injects high-level knowledge into the sensing pipeline to meet the demands of advanced tasks, such as scene understanding, which is a key requirement from industrial partners [156].

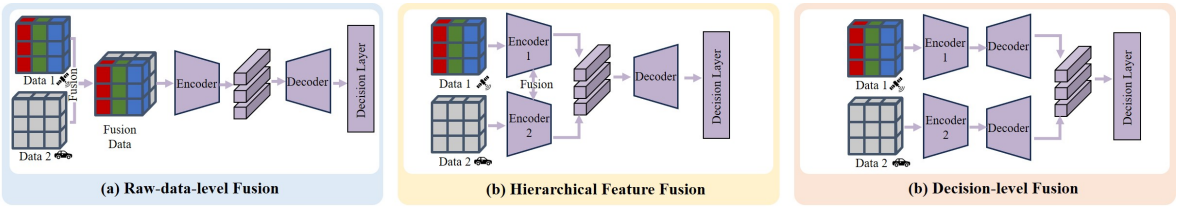

Figure 7. The different fusion strategy in the multi-domain data for Urban Embodied Agents.

4.2. Fusion Strategies for Multi-Domain Data

The fusion strategies designed to bridge domain gaps can be primarily classified into the following categories according to the stage at which they occur in the model fusion pipeline, including Raw-data-level Fusion, Hierarchical Feature Fusion, and Decision-level Fusion [157,158].
• **Raw-data-level Fusion**. In Raw-data-level Fusion, raw or independently preprocessed data from different modalities are merged, such as combining RGB images from satellites and depth maps from drones into a unified RGB-D data stream. This unified representation is then fed into a single encoder to extract high-level features. For instance, Muturi et al. [159] handle heterogeneous representations from the outset to bridge the Representation Gap. Additionally, the depth information can compensate for the lack of texture in images, improving the overall quality and robustness of the data to bridge the Quality & Density Gap. Shang et al. [160] integrate RGB images and depth information (from pre-trained depth models or consumer-level depth sensors) to enhance the performance in few-shot novel view synthesis.
• **Hierarchical Feature Fusion**. Hierarchical Feature Fusion combines features from different domains at various levels of abstraction [115], such as enriching a geometric model with high-level context from a pre-trained semantic segmentation backbone (such as CNN [161,162], ViT [163], PointNet [164]). For instance, the SUDS model [165] utilizes a joint optimization framework through a composite loss function to fuse multiple data domains, enabling the understanding of large-scale dynamic urban scenes and overcoming the limited observability to achieve a city-scale dynamic scene. GAANet [166] unifies cross-domain features in graph space via global graph interaction and local attention alignment, eliminating geometric and semantic gaps for robust fusion. To handle large-scale scenes, VastGaussian [167] and CityGaussian [168] also adopt a “divide-and-conquer" strategy, partitioning the scene into multiple cells for parallel optimization before merging. To address common Quality & Density Gap, such as illumination variation, they introduce a method that uses a CNN during optimization to separate the scene’s stable geometric colors from transient lighting effects.
• **Decision-level Fusion**. Decision-level Fusion, also known as Late Fusion, is a straightforward strategy where each data domain is processed independently through separate network streams. These

 doi:10.20944/preprints202601.2155.v1

individual networks extract high-level features or form initial predictions, and the fusion only occurs at the very end of the decoder or classifier of individual sub-networks. For instance, a model might separately process features corresponding to different levels of detail. Horizon-GS [169,170] uses a "coarse-to-fine" training strategy to establish a geometric model before addressing the gap between the aerial data (sparse 3D data) and terrain data (real street-view images). The City3D [171] method uses a height map from terrain to guide the reconstruction of aerial LiDAR point clouds. This not only ensures the semantic correctness of the result (addressing the Semantic Gap) but also helps handle issues like missing walls (addressing the Quality & Density Gap). CrossView-GS [172] uses a multi-branch architecture, training models on different view sets independently to serve as priors. It tries to capture Internet video to bridge the Quality & Density Gap when using cross-view data with large disparities.

5. Task Application: From Perception to Social Interaction

Having established the challenges of multi-domain data management in Section III and the fusion 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" [173,174]. 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 Spatio-Temporal 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

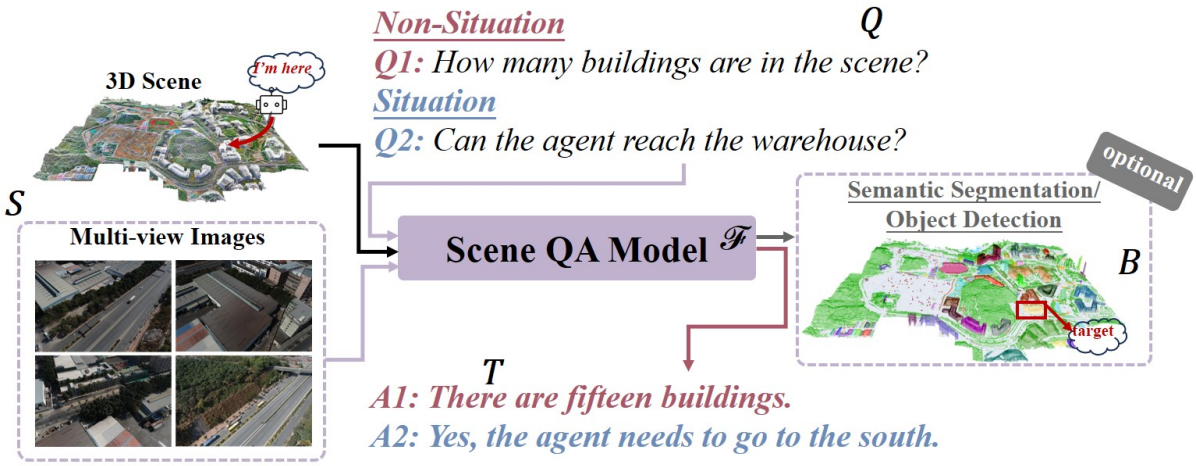

Figure 8. Definition of Urban Scene QA (SQA) for Urban Embodied Agents.

1) Definition. Urban SQA enables intelligent systems to answer queries about their spatial context, making it a key task for environmental interpretation [175]. 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

 doi:10.20944/preprints202601.2155.v1

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:

$$(T, B) = \mathcal{F}(S, Q). \tag{4}$$

Effectively performing SQA requires the model to fuse spatial understanding with multimodal 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.

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 [118], NuInstruct [176], NuPrompt [177], DriveLM [120], and VLAAD [119]. After the data fusion, researchers employ cross-attention or the multi-layer perceptron mechanisms to achieve deep interaction between vision and language [118,120,177]. 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 [74,178–181]. 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 [181] and Sg-CityU [74] construct a structured tree or graph to abstract complex spatial information into objects and their interrelations, which simplifies spatial reasoning. SOBA [178] and EarthVQANet [182] 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 [180] and CityBench [179] align perception models with real Geographic Information Systems (like OpenStreetMap) to provide precise geographic coordinates and place names.

• **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 [123] and CityEQA [183]. 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

 doi:10.20944/preprints202601.2155.v1

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 [183].

5.2. Vision-Language Navigation

1) Definition. Vision-and-Language Navigation (VLN) is a multimodal task requiring an embodied agent to navigate a realistic environment based on natural language instructions [184]. 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:

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

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 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 [48] and map2seq [185] 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 [186]), or by leveraging large language models to synthesize diverse auxiliary data, such as navigation rationales and landmark descriptions (e.g., FLAME [187], NavAgent [188]). Regarding multimodal fusion, strategies have evolved from early methods like direct feature vector concatenation (e.g., RconCAT [189]) and novel style-transfer fusion (e.g., VLN-Trans [190]), 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 [188]).

• **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 [55], OpenUAV [191], CityNav [192]. For instance, CityNav [192] 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 [193] 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 [194] 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.

 doi:10.20944/preprints202601.2155.v1

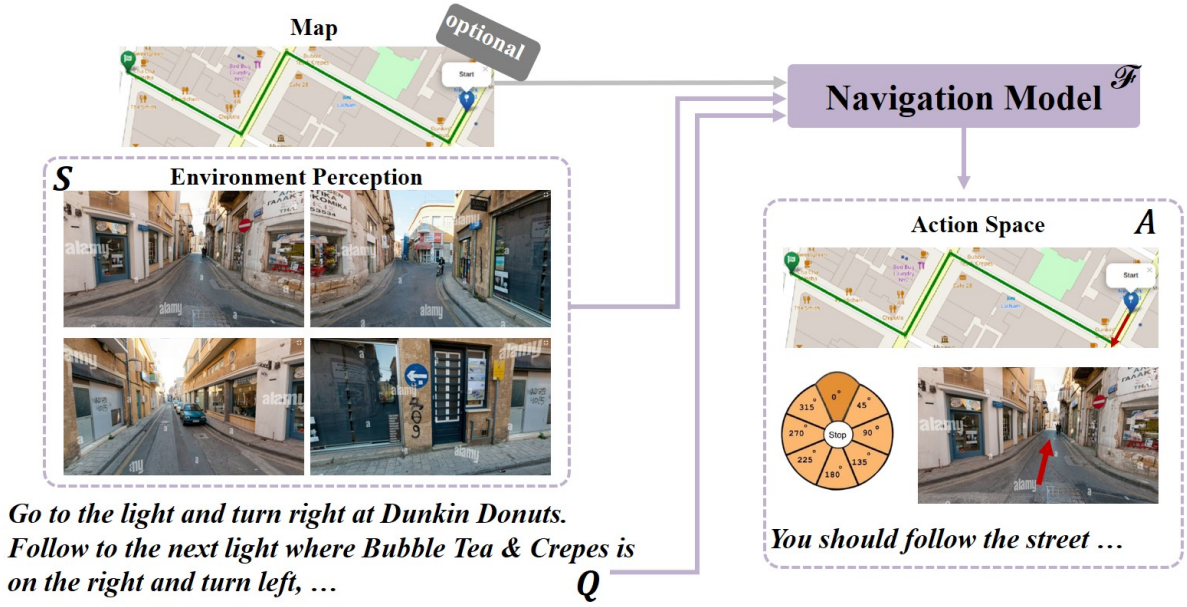

Figure 9. Definition of Vision-Language Navigation for Urban Embodied Agents.

5.3. Human-Agent Collaboration

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 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:

$$O = \mathcal{F}(G, S, I | P, C). \tag{6}$$

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 [195, 196]. In contrast to AI systems, Human-Agent Collaboration (HAC) leverages human strengths to build more efficient, robust, and trustworthy systems[197–199].

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.

 doi:10.20944/preprints202601.2155.v1

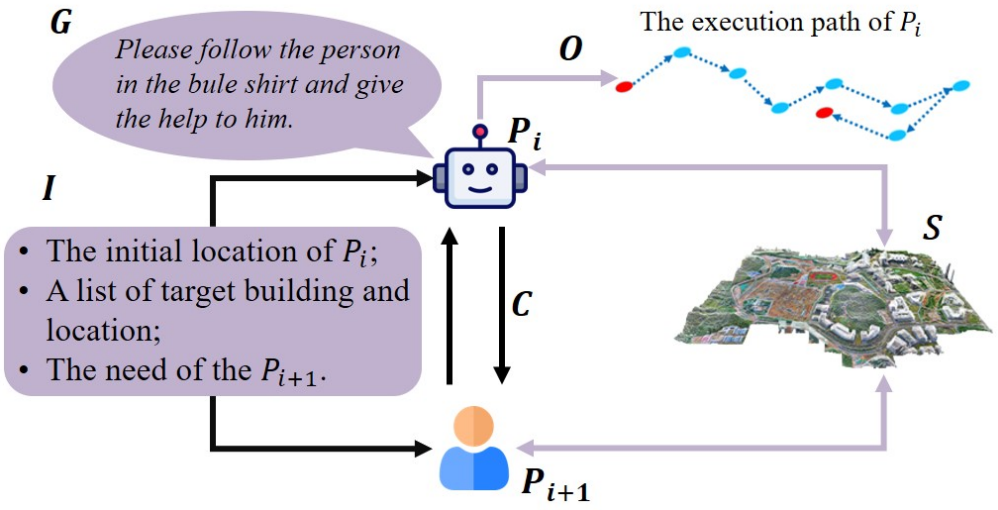

Figure 10. Definition of Human-Agent Collaboration for Urban Embodied Agents.

• **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 [200–204]. 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 [205].

• **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 [206]. 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 [207–210].

5.4. Discussion

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 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.

 doi:10.20944/preprints202601.2155.v1

6. Social Impact

The urban environment is a socio-physical environment that integrates social behavior with physical scene [211,212]. 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 [213, 214]. 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 [215]. For instance, in the traffic flow management [216–218], 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 [219].

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 [220–222]. 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" [223–225].

6.0.3. **Climate Change**

The high-density buildings and surfaces of cities exacerbate problems like the “Urban Heat Island" (UHI) effect. 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 [226]

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 [227]. 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 [228,229].

7. Outlook and Discussion

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

7.1. Method-Level Challenges

7.1.1. Robust Fusion with Imbalanced Multi-domain Data

This is a challenge that directly extends the discussion of the Quality & Density 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) [230].

 doi:10.20944/preprints202601.2155.v1

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 Learning and Incremental City Updates

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 [3,231]. 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.

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 Spatio-temporal Gap from handling minute-long sequences to managing year-long urban evolution.

7.2. System-Level Challenges

7.2.1. High-Fidelity Simulators and the Sim-to-Real Loop

While existing urban simulators like CARLA [16] and MetaUrban [201] 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).

A future direction is to leverage the UrbanEA itself to advance simulator development. By constructing ultra-high-fidelity digital twins from real-world data, we can extract physical materials, lighting properties, and behavioral patterns of traffic agents to build the next generation of more realistic simulators. This forms a "Real-to-Sim-to-Real" loop: real-world data improves the simulator, the improved simulator trains and validates better agents, and those agents then better perceive and influence the real world.

7.2.2. Multi-Agent Collaboration and 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 [232,233].

Future research must address communication bottlenecks, decentralized decision-making consistency, and collaboration among heterogeneous agents. 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. Causal Reasoning by Fusing Social Knowledge

This represents the leap across the Semantic Gap. Current fusion operates at the level of physical sensing (geometry, appearance), whereas a future agent must be able to fuse non-physical, abstract social knowledge [234].

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 fusion 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. Fairness and Data Bias

The challenge of imbalanced multi-domain data transcends a technical concern, posing an issue of social equity. Data perception often mirrors existing socio-economic disparities, leading to the over-

 doi:10.20944/preprints202601.2155.v1

representation of affluent areas and the under-representation of marginalized communities [235,236]. 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 [237–239].

Future research will create fusion methods that can be audited. It must focus on developing fairness-aware fusion 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 [240].

8. Conclusions

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 fusion, 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 multi-domain data fusion strategies, highlighting the three core challenges in this field (environmental variability, scale limitation, and interaction complexity) and the four major Gaps in multi-domain data fusion. 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.
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. 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*.
9. 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.
10. 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.
11. Lu, Y.; Tang, H. Multimodal Data Storage and Retrieval for Embodied AI: A Survey. *arXiv preprint arXiv:2508.13901* **2025**.
12. 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.

 doi:10.20944/preprints202601.2155.v1

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. 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.
16. 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.

 doi:10.20944/preprints202601.2155.v1

32. 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**.
33. 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.
34. El-Omari, S.; Moselhi, O. Integrating 3D laser scanning and photogrammetry for progress measurement of construction work. *Automation in construction* **2008**, *18*, 1–9.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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**.
48. 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.
49. 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.
50. 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.
51. 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.v1

52. 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.
53. 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.
54. 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.
55. 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.
56. 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**.
57. 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**.
58. Zhong, F.; Wu, K.; Wang, C.; Chen, H.; Ci, H.; Li, Z.; Wang, Y. UnrealZoo: Enriching Photo-realistic Virtual Worlds for Embodied AI. *arXiv preprint arXiv:2412.20977* **2024**.
59. 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**.
60. 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**.
61. 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**.
62. 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**.
63. 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. arXiv preprint arXiv:2407.06886, 2024.
64. Zheng, Y.; Yao, L.; Su, Y.; Zhang, Y.; Wang, Y.; Zhao, S.; Zhang, Y.; Chau, L.P. A Survey of Embodied Learning for Object-Centric Robotic Manipulation. arXiv preprint arXiv:2408.11537, 2024.
65. Warren, J.; Marz, N. *Big Data: Principles and best practices of scalable realtime data systems*; Simon and Schuster, 2015.
66. Kreps, J. Questioning the Lambda Architecture. O'Reilly Radar, 2014.
67. Azzabi, S.; Alfughi, Z.; Ouda, A. Data lakes: A survey of concepts and architectures. *Computers* **2024**, *13*, 183.
68. 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.
69. 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.
70. 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.
71. Neo4j, Inc.. Neo4j, 2010.
72. The JanusGraph Project. JanusGraph, 2017.
73. 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.
74. 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.
75. Meta AI Research. FAISS, 2017.

 doi:10.20944/preprints202601.2155.v1

76. 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.
77. 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.
78. Jegou, H.; Douze, M.; Schmid, C. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* **2010**, *33*, 117–128.
79. 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.
80. 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.
81. Timescale. TimescaleDB, 2018.
82. 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.
83. OSGeo, P.P.. PostGIS. https://postgis.net/, 2025.
84. 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.
85. 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.
86. Sawadogo, P.; Darmont, J. On data lake architectures and metadata management. *J. Intell. Inf. Syst.* **2021**, *56*, 97–120.
87. 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.
88. Liu, R.; Isah, H.; Zulkernine, F. A big data lake for multilevel streaming analytics. In Proceedings of the IBDAP. IEEE, 2020, pp. 1–6.
89. Inmon, B. *Data Lake Architecture: Designing the Data Lake and avoiding the garbage dump*; Technics Publications, LLC, 2016.
90. Jarke, M.; Quix, C. On warehouses, lakes, and spaces: the changing role of conceptual modeling for data integration. *Concept. Model. Perspect.* **2017**, pp. 231–245.
91. Ravat, F.; Zhao, Y. Data lakes: Trends and perspectives. In Proceedings of the DEXA. Springer, 2019, pp. 304–313.
92. Lu, J.; Holubová, I. Multi-model Data Management: What's New and What's Next? In Proceedings of the EDBT, 2017.
93. 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.
94. 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.
95. Khine, P.P.; Wang, Z. A review of polyglot persistence in the big data world. *Information* **2019**, *10*, 141.
96. 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.
97. Bimonte, S.; Gallinucci, E.; Marcel, P.; Rizzi, S. Data variety, come as you are in multi-model data warehouses. *Information Systems* **2022**, *104*, 101734.
98. Mihai, G. Multi-model database systems: The state of affairs. *Economics and Applied Informatics* **2020**, pp. 211–215.
99. 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.
100. 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.
101. 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.

 doi:10.20944/preprints202601.2155.v1

102. 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.
103. Kumar Kaliyar, R. Graph databases: A survey. In Proceedings of the International Conference on Computing, Communication & Automation. IEEE, 2015, pp. 785–790.
104. Robinson, I.; Webber, J.; Eifrem, E. *Graph databases: new opportunities for connected data*; " O'Reilly Media, Inc.", 2015.
105. 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*.
106. Sun, J.; Zhang, J.; Li, Q.; Yi, X.; Liang, Y.; Zheng, Y. Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks. *IEEE Transactions on Knowledge and Data Engineering* **2020**, *34*, 2348–2359.
107. 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.
108. 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.
109. 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.
110. 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.
111. 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.
112. 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.
113. 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**.
114. 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.
115. 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.
116. 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.
117. 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.
118. 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.
119. 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.
120. 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.
121. 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**.
122. Tang, D.; Shang, Z.; Elmore, A.J.; Krishnan, S.; Franklin, M.J. CrocodileDB in action: resource-efficient query execution by exploiting time slackness. *Proceedings of the VLDB Endowment* **2020**, *13*, 2937–2940.
123. 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**.

 doi:10.20944/preprints202601.2155.v1

124. 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.
125. 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.
126. 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**.
127. 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.
128. 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.
129. 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.
130. 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.
131. 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**.
132. 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.
133. 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.
134. 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**.
135. 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.
136. Li, S.; Tang, H. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040* **2024**.
137. 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**.
138. Özyeşil, O.; Voroninski, V.; Basri, R.; Singer, A. A survey of structure from motion*. *Acta Numerica* **2017**, *26*, 305–364.
139. 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.
140. 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.
141. 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**.
142. 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.
143. 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.
144. 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.

 doi:10.20944/preprints202601.2155.v1

145. 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.
146. 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.
147. 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.
148. 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.
149. 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.
150. 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.
151. 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.
152. 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.
153. 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.
154. 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.
155. 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.
156. 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.
157. Zhao, F.; Zhang, C.; Geng, B. Deep multimodal data fusion. *ACM computing surveys* **2024**, *56*, 1–36.
158. 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.
159. 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.
160. 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**.
161. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
162. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
163. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations.
164. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660.

 doi:10.20944/preprints202601.2155.v1

165. 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.
166. 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**.
167. 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.
168. 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.
169. 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.
170. 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.
171. Huang, J.; Stoter, J.; Peters, R.; Nan, L. City3D: Large-scale building reconstruction from airborne LiDAR point clouds. *Remote Sensing* **2022**, *14*, 2254.
172. Zhang, C.; Cao, Y.; Zhang, L. CrossView-GS: Cross-view Gaussian Splatting For Large-scale Scene Reconstruction. *arXiv preprint arXiv:2501.01695* **2025**.
173. 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.
174. 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**.
175. 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.
176. 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.
177. 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.
178. 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.
179. 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**.
180. 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.
181. 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**.
182. 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.
183. 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**.
184. 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.
185. 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

 doi:10.20944/preprints202601.2155.v1

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 489–502.
186. 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.
187. 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.
188. 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.
189. 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.
190. 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.
191. 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.
192. 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.
193. 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.
194. 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.
195. 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**.
196. 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.
197. 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**.
198. 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.
199. 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**.
200. 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.
201. 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. *arXiv preprint arXiv:2407.08725* **2024**.
202. 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.
203. 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.
204. Romeu-Guallart, P.; Zamora-Martinez, F. SML2010. *UCI Machine Learning Repository* **2014**.

 doi:10.20944/preprints202601.2155.v1

205. 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**.
206. 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**.
207. 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.
208. 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**.
209. 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**.
210. 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**.
211. 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.
212. Gao, C.; Lan, X.; Li, N.; Yuan, Y.; Ding, J.; Zhou, Z.; Xu, F.; Li, Y. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* **2024**, *11*, 1–24.
213. 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.
214. 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.
215. Mouratidis, K. Time to challenge the 15-minute city: Seven pitfalls for sustainability, equity, livability, and spatial analysis. *Cities* **2024**, *153*, 105274.
216. Hamissi, A.; Dhraief, A. A survey on the unmanned aircraft system traffic management. *ACM Computing Surveys* **2023**, *56*, 1–37.
217. 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.
218. 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.
219. 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.
220. 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.
221. 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.
222. Wang, L.; Shao, J.; Ma, Y. Does China's low-carbon city pilot policy improve energy efficiency? *Energy* **2023**, *283*, 129048.
223. 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.
224. 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.
225. 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.

 doi:10.20944/preprints202601.2155.v1

226. 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.
227. 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**.
228. 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**.
229. 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.
230. 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.
231. 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**.
232. Kennedy, J. Swarm intelligence. In *Handbook of nature-inspired and innovative computing: integrating classical models with emerging technologies*; Springer, 2006; pp. 187–219.
233. 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.
234. 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.
235. 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.
236. 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.
237. 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.
238. 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.
239. 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.
240. 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**.