Simulating Diverse Urban Environments for AI-Driven Mobility and Accessibility Studies

WAYNE WU, HONGLIN HE, and BOLEI ZHOU, University of California, Los Angeles, USA

As cities evolve towards inclusivity, ensuring accessibility for all urban residents is a critical challenge. Intelligent mobility systems, particularly those designed for individuals with disabilities, are essential for navigating complex urban environments. In this paper, we present a simulation-based perspective, aimed at advancing research in AI-driven urban mobility and accessibility. First, we propose MetaUrban, a flexible and scalable *simulation platform* that enables the development, testing, and evaluation of embodied AI systems in dynamic urban environments. MetaUrban can construct an infinite number of interactive urban scenes from compositional elements, covering a vast array of ground plans, object placements, pedestrians, vulnerable road users, and other mobile agents' appearances and dynamics. Then, to facilitate AI-driven *urban mobility* research, we designed point navigation and social navigation tasks as the pilot study using MetaUrban and established various baselines of Reinforcement Learning and Imitation Learning. Further, we outline a broader vision for how MetaUrban can be leveraged to simulate and improve *urban accessibility* across a range of mobility systems. We position MetaUrban as a foundational tool for future urban design, promoting research that enhances the mobility and accessibility of cities. Project page: https://metadriverse.github.io/metaurban

CCS Concepts: • **Computing methodologies** → **Simulation environments**; *Artificial intelligence*; • **Human-centered computing** → *Accessibility systems and tools.*

Additional Key Words and Phrases: Simulation Platforms, Urban Mobility, Urban Accessibility, Embodied AI

ACM Reference Format:
Wayne Wu, Honglin He, and Bolei Zhou. 2024. Simulating Diverse Urban Environments for AI-Driven Mobility and Accessibility Studies. 1, 1 (October 2024), 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Public urban spaces, such as streetscapes, plazas, and parks, are vital components of cities. These spaces serve not only as transit corridors but also as social and economic hubs where interactions between residents occur. Historically, urban spaces have played a key role in shaping social behavior and community engagement [Herbert 1962; Jacobs 1961; Park et al. 1925]. As cities continue to grow and evolve, making these spaces accessible and inclusive to all individuals becomes increasingly crucial for enhancing the quality of life and ensuring equitable access to public resources.

Authors' Contact Information: Wayne Wu, wuwenyan0503@gmail.com; Honglin He, hollis71025@gmail.com; Bolei Zhou, bolei@cs.ucla.edu, University of California, Los Angeles, Los Angeles, California, USA.

ACM XXXX-XXXX/2024/10-ART
https://doi.org/XXXXXXX.XXXXXXX

In recent years, the development of Robotics and Embodied AI has introduced a wide range of AI-driven mobile machines into urban spaces. From autonomous delivery robots to electric wheelchairs, these intelligent systems are becoming a common sight in public areas, navigating alongside pedestrians. Thus, *how to ensure mobile machines' ability to move safely through complex urban environments* and *how to design the urban space to make it accessible to the varied nature of mobile machines* become critical questions. These two ones form a pair of interesting dual questions – one stands in improving mobile machines' ability when facing complex urban environments, and another, in contrast, stands in improving urban environments themselves to better accommodate mobile machines. Urban simulation, which is able to model complex scenarios and interactions in urban spaces, allowing the development and testing of AI systems in a controlled, scalable environment, provides a promising way to improve both AI-driven mobility and accessibility in cities.

Simulation platforms [Deitke et al. 2020, 2022b; Dosovitskiy et al. 2017; Kolve et al. 2017; Krajzewicz et al. 2002; Li et al. 2024, 2022b; Savva et al. 2019; Shen et al. 2021; Szot et al. 2021] have played a crucial role in enabling systematic and scalable training of the embodied AI agents and the safety evaluation before real-world deployment. However, most of the existing simulators focus either on *indoor household environments* [Gan et al. 2021; Kolve et al. 2017; Li et al. 2024; Puig et al. 2018; Savva et al. 2019; Shen et al. 2021] or *outdoor driving environments* [Dosovitskiy et al. 2017; Krajzewicz et al. 2002; Li et al. 2022b]. For example, platforms like AI2-THOR [Kolve et al. 2017], Habitat [Savva et al. 2019], and iGibson [Shen et al. 2021] are designed for household assistant robots in which the environments are mainly apartments or houses with furniture and appliances; platforms like SUMO [Krajzewicz et al. 2002], CARLA [Dosovitskiy et al. 2017], and MetaDrive [Li et al. 2022b] are designed for research on autonomous driving and transportation. Yet, simulating *urban spaces* with diverse layouts and objects, complex dynamics of pedestrians, is much less explored.

To address this gap, we present a simulation-based perspective, demonstrating a practical scenario of how a simulator enhances urban mobility and discussing a future vision of how a simulator can be used for improving urban accessibility. First, we propose a simulation platform, MetaUrban, which can construct an infinite number of interactive urban scenes from compositional elements, covering a vast array of ground plans, object placements, pedestrians, vulnerable road users, and other mobile agents' appearances and dynamics. Then, we make a pilot study using MetaUrban in AI-driven urban mobility with two standard tasks – point navigation and social navigation, and establish extensive baselines for Reinforcement Learning, Safe Reinforcement Learning, Offline Reinforcement Learning, and Imitation Learning. Finally, we explore how MetaUrban can be adapted to address broader issues of urban accessibility, offering a vision for how simulation platforms can be leveraged to create more inclusive cities. We envision this work will

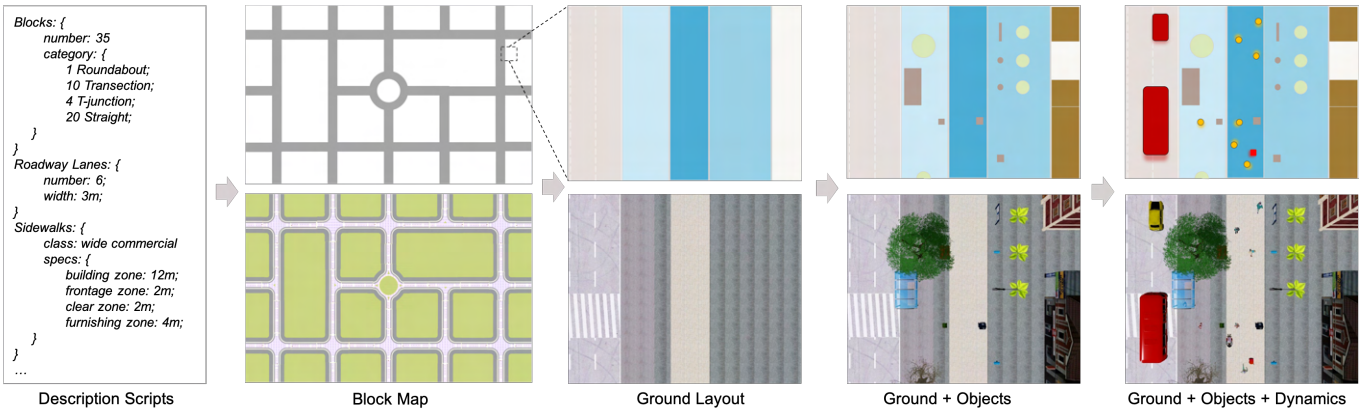

Fig. 1. **Procedural generation.** MetaUrban can automatically generate complex urban scenes with its compositional nature. From the second to the fourth column, the top row shows the 2D road maps, and the bottom row shows the bird-eye view of 3D scenes.

serve as a pioneering exploration that harnesses urban simulation platforms to advance AI-driven mobility in complex environments and enhance the accessibility of public spaces for all individuals.

2 MetaUrban Simulation Platform

MetaUrban is a compositional simulation platform that can generate infinite training and evaluation environments for Embodied AI in urban spaces. Figure 1 depicts the procedural generation pipeline. MetaUrban uses a structured description script to create urban scenes. Based on the provided information about street blocks, sidewalks, objects, agents, and more, it starts with the street block map, then plans the ground layout by dividing different function zones, then places static objects, and finally populates dynamic agents.

This section highlights three key designs in the MetaUrban simulator to support exhibiting three unique characteristics of urban spaces – diverse layouts, particular object distribution, and complex dynamics. Section 2.1 introduces **Hierarchical Layout Generation**, which can infinitely generate diverse layouts with different functional zone divisions and object locations that are critical for the *generalizability* of agents. Section 2.2 introduces **Scalable Object Retrieval**, which harnesses worldwide urban scene data to obtain real-world object distributions in different places, and then builds large-scale, high-quality static objects set with Vision Language Models (VLMs) [Li et al. 2022a] enabled open-vocabulary searching. It is useful for training agents *specialized* for urban scenes. Section 2.3 introduces **Cohabitant Populating**, in which we leverage the advancements in digital humans to enrich the appearances, movements, and trajectories of pedestrians and vulnerable road users, as well as incorporate other agents to form a vivid cohabiting environment. It is critical for improving the *social conformity* and *safety* of the mobile agents.

2.1 Hierarchical Layout Generation

The diversity of scene layout, *i.e.*, the connection and categories of blocks, the specifications of sidewalks and crosswalks, as well as the placement of objects, is crucial for enhancing the generalizability of trained agents maneuvering in public spaces. In the hierarchical layout generation framework, we start by sampling the categories of street blocks and dividing sidewalks and crosswalks and then allocate various objects, with which we can get infinite urban scene layouts with arbitrary sizes and specifications of maps.

Ground plan. We design 5 typical street block categories, *i.e.*, straight, intersection, roundabout, circle, and T-junction. In the simulator, to form a large map with several blocks, we can sample the category, number, and order of blocks, as well as the number and width of lanes in one block, to get different maps. Then, each block can simulate its own walkable areas – sidewalks and crosswalks, which are key areas for urban spaces with plenty of interactions.

As shown in Figure 2 (left), according to the Global Street Design Guide [Initiative and of City Transportation Officials 2016] provided by the Global Designing Cities Initiative, we divide the sidewalk into four functional zones – building zone, frontage zone, clear zone, and furnishing zone. Based on their different combinations of functional zones, we further construct 7 typical templates for sidewalks (Figure 2 (right)). To form a sidewalk, we can first sample the layout from the templates and then assign proportions for different function zones. For crosswalks, we provide candidates at the start and the end of each roadway, which support specifying the needed crosswalks or sampling them by a density parameter. Finally, roadways, sidewalks, and crosswalks can take a terrain map as substrate to form different ground situations.

Object placement. After determining the ground layout, we can place objects on the ground. We divide objects into three classes. 1) Standard infrastructure, such as poles, trees, and signs, are placed periodically along the road. 2) Non-standard infrastructure, such as buildings, bonsai, and trash bins, are placed randomly in the designated function zones. 3) Clutter, such as drink cans, bags, and bicycles, are placed randomly across all functional zones. We can get different street styles by specifying an object pool while getting different compactness by specifying a density parameter.

2.2 Scalable Object Retrieval

Hierarchical layout generation decides the scene's layout and *where* to place the objects. However, to make the trained agents generalizable when navigating through scenes composed of various objects

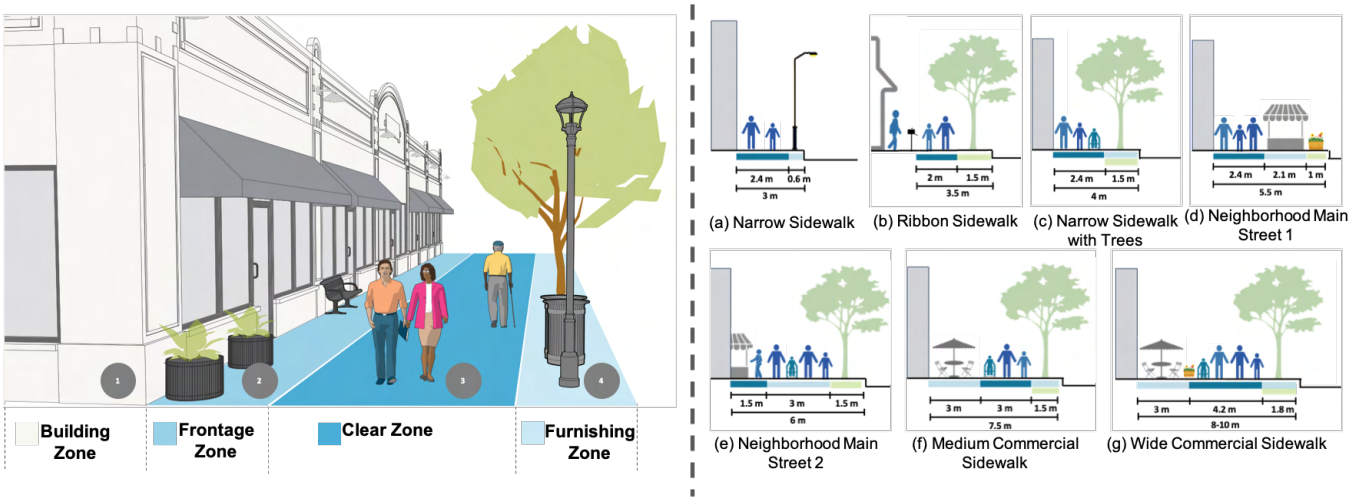

Fig. 2. **Ground plan.** (Left) Sidewalk is divided into four functional zones – building, frontage, clear, and furnishing zone. (Right) Seven typical sidewalk templates – from (a) to (g).

in the real world, *what* objects to place is another crucial question. In this section, we propose the Scalable Assets Retrieval pipeline, in which we first get real-world object distributions from web data, and then retrieve objects from 3D asset repositories through an open-vocabulary search schema based on Vision Language Models (VLMs) [Li et al. 2022a]. This pipeline is flexible and extensible: the retrieved objects can be scaled to arbitrary sizes as we continue to exploit more web data for scene descriptions and include more 3D assets as the candidate objects.

Real-world object distribution extraction. Urban spaces have unique structures and object distributions, such as the infrastructure built by the urban planning administration and clutters placed by people. Thus, we design a real-world distribution extraction method to get a description pool depicting the frequent objects in urban spaces. As illustrated in Figure 3 (a), we first leverage off-the-shelf academic datasets for scene understanding, CityScape [Cordts et al. 2016] and Mapillary Vistas [Neuhold et al. 2017], to get a list of 90 objects that are with high frequency to be put in the urban space. However, the number of objects is limited because of the closed-set definitions in the image datasets. We introduce two open-set sources to get broader object distribution from the real world. 1) Google Street data. We first collect 25,000 urban space images from 50 countries across six continents. Then, we harness GPT-4o [OpenAI 2024] and open-set segmentation model Grounded-SAM [Ren et al. 2024] to get 1,075 descriptions of objects in the urban public space. 2) Urban planning description data. We further get a list of 50 essential objects in public urban spaces through a thorough survey of 10 urban design handbooks. Finally, by combining these three data sources, we can get an object description pool with 1,215 items of descriptions that form the real-world object category distribution.

Open-vocabulary search. The recent development of large-scale 3D object repositories [Deitke et al. 2024, 2023; Wu et al. 2023] enables efficiently constructing a dataset for a specific scene. However, these large repositories have three intrinsic issues to harness these repositories: 1) most of the data is unrelated to the urban scene, 2) the data quality in large repositories is uneven, and 3) the data has no reliable attribute annotations. To this end, we introduce an open-vocabulary search method to tackle these issues. As shown in Figure 3 (b), the whole pipeline is based on an image-text retrieval architecture. We first sample objects from Objaverse [Deitke et al. 2023] and Objaverse-XL [Deitke et al. 2024] repositories to get projected multi-view images. Here, a naive uniform view sampling will bring low-quality harmful images. Following [Luo et al. 2024, 2023], we select and prioritize informative viewpoints, which significantly enhance retrieval effectiveness. Then, we leverage the encoder of a Vision Language Model BLIP [Li et al. 2022a] to extract features from projected images and sampled descriptions from the object description pool, respectively, to calculate relevant scores. Then, we can get target objects with relevant scores up to a threshold. This method lets us get an urban-specific dataset with 10,000 high-quality objects in real-world category distributions. In addition, we provide an interface for customizing training objects in the scene by providing images or text descriptions, taking advantage of recent advances in 3D object reconstruction [Kerbl et al. 2023; Liu et al. 2023b] and generation [Chen et al. 2023; Poole et al. 2023].

2.3 Cohabitant Populating

In this section, we will describe how to populate these static urban scenes with varied agents regarding appearances, movements, and trajectories through Cohabitant Populating. Figure 4 shows the sampled (a) pedestrians, vulnerable road users and robots, (b) movements, and (c) trajectories.

Appearances. Following BEDLAM [Black et al. 2023] and AGORA [Patel et al. 2021], we represent humans as parametric human model SMPL-X [Pavlakos et al. 2019], in which the 3D human body is controlled by a set of parameters for pose θ, shape β, and facial expression ϕ, respectively. Then, built upon SynBody [Yang et al. 2023]'s asset repository, 1,100 3D rigged human models are procedually generated by sampling from 68 garments, 32 hairs, 13 beards,

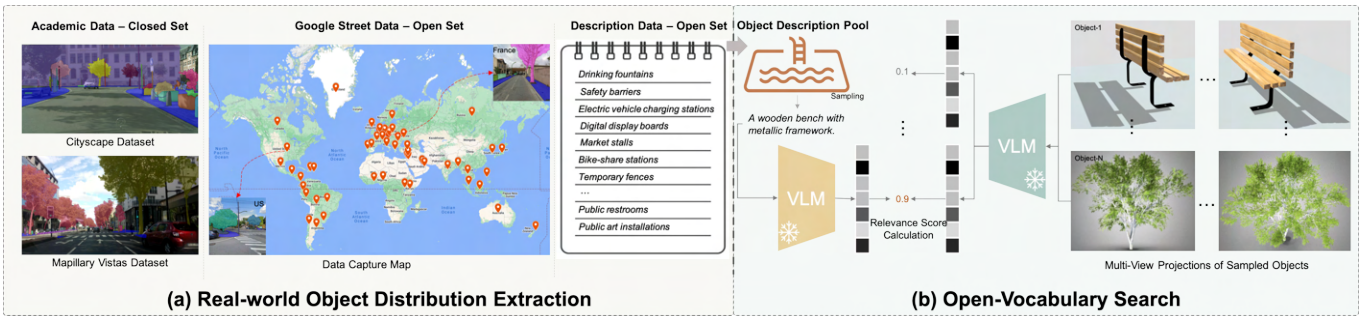

Fig. 3. **Scalable assets retrieval.** (a) Real-world distribution extraction. We get object distribution for urban spaces from three sources: academic datasets, Google Street data, and text description data. (b) Open-vocabulary search. We use the VLM to get image and text embedding, respectively. Then, based on the relevant scores, we can get the objects with high rankings.

46 accessories, and 1,038 cloth and skin textures. To form safety-critical scenarios, we also include vulnerable road users like bikers, skateboarders, and scooter riders. For the other agents, we incorporate the 3D assets of COCO Robotics and Starship's delivery robots, Drive Medical's electric wheelchair, Boston Dynamic's robot dog, and Agility Robotics' humanoid robot.

Movements. We provide two kinds of human movements in the simulator – daily movements and unique movements. Daily movements provide the basic human dynamics in daily life, *i.e.*, idle, walking, and running. Unique movements are the complicated dynamics that appear randomly in public spaces, such as dancing and exercising. We harness the BEDLAM dataset [Black et al. 2023] to obtain 2,311 unique movements.

Trajectories. For humans and other agents with daily movements, we simulate their trajectories using the ORCA [Van Den Berg et al. 2011] social forces model and Push and Rotate algorithm [De Wilde et al. 2014]. ORCA [Van Den Berg et al. 2011] uses a joint optimization and a centralized controller that guarantees that agents will not collide with each other or any other objects identified as obstacles. Push and Rotate (P&R) [De Wilde et al. 2014] is a multi-agent path-finding algorithm that can resolve any potential deadlock by local coordination. In the future, an interesting direction is to endow personal traits like job, personality, and purpose to humans and harness the advantages of LLMs [Achiam et al. 2023] and LVMs [Liu et al. 2023a] to enable social [Puig et al. 2023] and interactive behaviors [Park et al. 2023] of humans in urban scenes.

3 MetaUrban for AI-Driven Mobility

In this section, we present experiments to demonstrate how to ensure mobile machines' ability to move safely through complex urban environments with the MetaUrban simulator. In urban spaces, for all mobile machines, such as intelligent electric wheelchairs and delivery robots, the most foundational demand is moving from one point to another. To this end, in this section, we design two standard tasks, Point Navigation (PointNav) and Social Navigation (SocialNav), and benchmark extensive learning methods – Reinforcement Learning, Safe Reinforcement Learning, Offline Reinforcement Learning, and Imitation Learning, to evaluate how the MetaUrban simulation platform can help in AI-driven urban mobility.

Data. Based on the MetaUrban simulator, we construct the MetaUrban-12K dataset, including 12,800 interactive urban scenes for training (*MetaUrban-train*) and 1,000 scenes for testing (*MetaUrban-test*). Scenes in this dataset are connected by one to three street blocks covering an average of 20,000m^2 areas. There are an average of 0.03 static objects per m^2 and the average distance of objects is 0.7m. There are 10 dynamic agents per street block, including pedestrians, vulnerable road users, and other agents. The average episode length is 410m. These form significantly challenging scenes for agents to navigate through, which are crowded and have long horizons. We further construct an unseen test set (*MetaUrban-unseen*) with 100 scenes for the unseen evaluations. To enable the fine-tuning experiments, we construct a training set of 1,000 scenes with the same distribution of MetaUrban-unseen, termed *MetaUrban-finetune*.

Tasks. In PointNav, the agent's goal is to navigate to the target coordinates in static environments without access to a pre-built environment map. In SocialNav, the agent is required to reach a point goal in dynamic environments that contain moving environmental agents. The agent shall avoid collisions or proximity to environmental agents beyond thresholds to avoid penalization (distance <0.2 meters). The agent's action space in the experiments consists of acceleration, brake, and steering. The observations contain a vector denoting the LiDAR signal, a vector summarizing the agent's state, and the navigation information that guides the agent toward the destination.

Methods. We evaluate 7 typical baseline models to build comprehensive benchmarks on MetaUrban, across Reinforcement Learning (PPO [Schulman et al. 2017]), Safe Reinforcement Learning (PPO-Lag [Ray et al. 2019], and PPO-ET [Sun et al. 2021]), Offline Reinforcement Learning (IQL [Kostrikov et al. 2021] and TD3+BC [Fujimoto and Gu 2021]), and Imitation Learning (BC [Bain and Sammut 1995] and GAIL [Ho and Ermon 2016]).

Evaluation metrics. The agent is evaluated using the Success Rate (SR) and Success weighted by Path Length (SPL) [Anderson et al. 2018; Batra et al. 2020] metrics, which measure the success and efficiency of the path taken by the agent. For SocialNav, except Success Rate (SR), the Social Navigation Score (SNS) [Deitke et al. 2022a], is also used to evaluate the social complicity of the agent. For both tasks, we further report the Cumulative Cost (CC) [Li et al.

Fig. 4. **Cohabitant populating.** (a) Examples of cohabitants in MetaUrban: pedestrians, vulnerable road users like bikers, skateboarders, scooter riders, and mobile machines. (b) Examples of human movements. (c) Examples of trajectories of humans and mobile agents in complex interaction scenarios.

2022b] to evaluate the safety properties of the agent. It records the crash frequency to obstacles or environmental agents.

Results. We build two benchmarks for PointNav and SocialNav tasks. We train 7 typical baselines on the MetaUrban-train dataset and then evaluate them on the MetaUrban-test set. We further make zero-shot evaluations on the MetaUrban-unseen set to demonstrate the generalizability of models while directly tested on unseen environments. Table 1 shows the results in the PointNav and SocialNav benchmarks. From the results, we can draw 4 key observations. 1) The tasks are far from being solved. 2) Models trained on MetaUrban-12K have strong generalizability in unseen environments. 3) SocialNav is much harder than PointNav due to the dynamics of the mobile environmental agents. 4) Safe RL remarkably improves the safety property at the expense of effectiveness. For qualitative results, please refer to our project page[1].

Table 1. **Benchmarks.** The benchmark of PointNav and SocialNav tasks on the MetaUrban-12K dataset. Seven representative methods of RL, safe RL, offline RL, and IL are evaluated for each benchmark. ▇ indicates the best performance among online methods (RL and Safe RL). ▇ indicates the best performance among offline methods (offline RL and IL).[2]

Category	Method	PointNav						SocialNav					
		Test			Zero-shot			Test			Zero-shot		
		SR↑	SPL↑	Cost↓	SR↑	SPL↑	Cost↓	SR↑	SNS↑	Cost↓	SR↑	SNS↑	Cost↓
RL	PPO [Schulman et al. 2017]	0.66	0.64	0.51	0.49	0.45	0.78	0.34	0.64	0.66	0.24	0.57	0.51
Safe RL	PPO-Lag [Ray et al. 2019]	0.60	0.58	0.41	0.60	0.57	0.53	0.17	0.51	0.33	0.08	0.47	0.50
	PPO-ET [Sun et al. 2021]	0.57	0.53	0.47	0.53	0.49	0.65	0.05	0.52	0.26	0.02	0.50	0.62
Offline RL	IQL [Kostrikov et al. 2021]	0.36	0.33	0.49	0.30	0.27	0.63	0.36	0.67	0.39	0.27	0.62	3.05
	TD3+BC [Fujimoto and Gu 2021]	0.29	0.28	0.77	0.20	0.20	1.16	0.26	0.61	0.62	0.32	0.64	1.53
IL	BC [Bain and Sammut 1995]	0.36	0.28	0.83	0.32	0.26	1.15	0.28	0.56	1.23	0.18	0.54	0.58
	GAIL [Ho and Ermon 2016]	0.47	0.36	1.05	0.40	0.32	1.46	0.34	0.63	0.71	0.28	0.61	0.67

4 MetaUrban for AI-Driven Accessibility

In this section, we discuss how to design urban spaces to make them more accessible to the varied nature of mobile machines with the MetaUrban simulator. We outline 5 aspects that could potentially improve urban accessibility to better accommodate mobile machines and, by extension, humans.

Simulating Diverse Accessibility Challenges. Urban spaces present challenges like uneven terrain, curbs, and crowded sidewalks. MetaUrban can simulate these conditions to test how mobility systems and humans adapt to diverse accessibility challenges. Doing so helps optimize urban layouts, ensuring that infrastructure accommodates both AI-driven devices and individuals with mobility impairments.

Modeling Infrastructure Accessibility. Accessibility infrastructure, such as ramps and tactile paving, is critical in cities. MetaUrban can simulate environments with or without these features, helping planners evaluate whether urban spaces meet accessibility standards by evaluating whether mobility systems and humans could navigate these spaces effectively.

Modeling Different Mobile Machines. Urban spaces must support a range of mobility machines, from electric wheelchairs to autonomous delivery bots to robot dogs to humanoid robots. MetaUrban provides various mobile machines and can simulate interactions between them, ensuring that urban designs accommodate diverse mobility solutions and improving navigation for everyone while reducing congestion and conflict in shared spaces.

Modeling Social Interactions. Pedestrian behavior and interactions in crowded urban spaces are often unpredictable. MetaUrban provides diverse pedestrian models and vulnerable road users, simulating their behaviors and interactions with both each other and mobility systems. By modeling these dynamics, MetaUrban enables city planners to design urban spaces that facilitate smoother interactions and accommodate the needs of different road users.

Improving Safety Protocols. Safety is a key concern in busy urban environments. MetaUrban can simulate high-risk scenarios, such as intersections or emergencies, to evaluate and improve the safety protocols of mobility systems. This ensures urban spaces are safer for vulnerable populations, such as individuals with disabilities, during everyday use and critical situations.

[1]https://metadriverse.github.io/metaurban

[2]Results between Test and Zero-shot are not comparable, with the evaluations on different datasets.

5 Conclusion

In this work, we presented MetaUrban, a highly adaptable and scalable simulation platform designed to tackle the challenges of AI-driven urban mobility and accessibility. By enabling the construction of diverse, interactive urban environments, MetaUrban provides a robust foundation for the development, testing, and evaluation of embodied AI systems in complex urban environments. Our pilot study on point and social navigation tasks underscores the platform's potential to advance research in AI-driven urban mobility.

Beyond mobility, MetaUrban offers a forward-looking framework for addressing urban accessibility by simulating a wide range of mobility systems, pedestrian interactions, and infrastructure challenges. As cities increasingly prioritize inclusivity, platforms like MetaUrban will be instrumental in driving the design of AI-enabled, accessible urban spaces that cater to the diverse needs of all residents, including able-bodied individuals, people with disabilities, and robots.

6 Future Work

Future development of MetaUrban will focus on enhancing the platform's ability to simulate accessibility challenges, particularly for mobility-impaired individuals. This will involve incorporating real-world behavioral data and integrating user-driven requirements to improve agent modeling. Additionally, we plan to incorporate insights from urban planning and transportation modeling, enabling MetaUrban to simulate diverse mobility scenarios and environmental conditions more accurately. By adopting an interdisciplinary approach, including feedback from accessibility and urban design experts, MetaUrban aims to better represent the complexities of urban environments and offer more effective AI-driven solutions for accessible cities.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

Peter Anderson, Angel X. Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. 2018. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757* (2018).

Michael Bain and Claude Sammut. 1995. A Framework for Behavioural Cloning. In *MI*, Koichi Furukawa, Donald Michie, and Stephen H. Muggleton (Eds.).

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. 2020. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171* (2020).

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. 2023. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*.

Zilong Chen, Feng Wang, and Huaping Liu. 2023. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585* (2023).

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

Boris De Wilde, Adriaan W Ter Mors, and Cees Witteveen. 2014. Push and rotate: a complete multi-agent pathfinding algorithm. *JAIR* (2014).

Matt Deitke, Dhruv Batra, Yonatan Bisk, Tommaso Campari, Angel X. Chang, Devendra Singh Chaplot, Changan Chen, Claudia Pérez-D'Arpino, Kiana Ehsani, Ali Farhadi, Li Fei-Fei, Anthony G. Francis, Chuang Gan, Kristen Grauman, David Hall, Winson Han, Unnat Jain, Aniruddha Kembhavi, Jacob Krantz, Stefan Lee, Chengshu Li, Sagnik Majumder, Oleksandr Maksymets, Roberto Martín-Martín, Roozbeh Mottaghi, Sonia Raychaudhuri, Mike Roberts, Silvio Savarese, Manolis Savva, Mohit Shridhar, Niko Sünderhauf, Andrew Szot, Ben Talbot, Joshua B. Tenenbaum, Jesse Thomason, Alexander Toshev, Joanne Truong, Luca Weihs, and Jiajun Wu. 2022a. Retrospectives on the embodied ai workshop. *arXiv preprint arXiv:2210.06849* (2022).

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. 2020. RoboTHOR: An Open Simulation-to-Real Embodied AI Platform. In *CVPR*.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. 2024. Objaverse-xl: A universe of 10m+ 3d objects. *NeuIPS* (2024).

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *CVPR*.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. 2022b. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. *NeuIPS* (2022).

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *CoRL*.

Scott Fujimoto and Shixiang Shane Gu. 2021. A minimalist approach to offline reinforcement learning. *NeuIPS* (2021).

Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin T. Feigelis, Daniel Bear, Dan Gutfreund, David D. Cox, Antonio Torralba, James J. DiCarlo, Josh Tenenbaum, Josh H. McDermott, and Dan Yamins. 2021. ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation. In *NeurIPS Datasets and Benchmarks*.

Gans Herbert. 1962. *The Urban Villagers: Group and Class in the Life of Italian-Americans.*

Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *NeuIPS* (2016).

Global Designing Cities Initiative and National Association of City Transportation Officials. 2016. *Global street design guide.* Island Press.

Jane Jacobs. 1961. *The Death and Life of Great American Cities.*

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *TOG* (2023).

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Kembhavi Aniruddha, Gupta Abhinav, and Farhadi Ali. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474* (2017).

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2021. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169* (2021).

Daniel Krajzewicz, Georg Hertkorn, Christian Rössel, and Peter Wagner. 2002. SUMO (Simulation of Urban MObility)-an open-source traffic simulation. In *MESM*.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R. Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. 2024. BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark with 1, 000 Everyday Activities and Realistic Simulation. *CoRL* (2024).

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. 2022b. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *TPAMI* (2022).

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *NeuIPS* (2023).

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023b. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Tiange Luo, Justin Johnson, and Honglak Lee. 2024. View Selection for 3D Captioning via Diffusion Ranking. *arXiv preprint arXiv:2404.07984* (2024).

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. 2023. Scalable 3d captioning with pretrained models. *NeuIPS* (2023).

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*.

OpenAI. 2024. GPT-4o. https://openai.com/index/hello-gpt-4o/.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *UIST*.

Robert Ezra Park, Ernest Watson Burgess, Roderick Duncan McKenzie, and Louis Wirth. 1925. *The City.*

Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. 2021. AGORA: Avatars in geography optimized for

regression analysis. In *CVPR*.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *CVPR*.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2023. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. In *CVPR*.

Xavier Puig, Tianmin Shu, Joshua B Tenenbaum, and Antonio Torralba. 2023. Nopa: Neurally-guided online probabilistic assistance for building socially intelligent home assistants. In *ICRA*.

Alex Ray, Joshua Achiam, and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708* (2019).

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).

Manolis Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, and Vladlen Koltun. 2019. Habitat: A Platform for Embodied AI Research. In *ICCV*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, Tchapmi Micael, Vainio Kent, Wong Josiah, Fei-Fei Li, and Savarese Silvio. 2021. igibson 1.0: a simulation environment for interactive tasks in large realistic scenes. In *IROS*.

Hao Sun, Ziping Xu, Meng Fang, Zhenghao Peng, Jiadong Guo, Bo Dai, and Bolei Zhou. 2021. Safe exploration by solving early terminated mdp. *arXiv preprint arXiv:2107.04200* (2021).

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John M. Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel X. Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In *NeuIPS*.

Jur Van Den Berg, Stephen J Guy, Ming Lin, and Dinesh Manocha. 2011. Reciprocal n-body collision avoidance. In *ISRR*.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. 2023. OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation. In *CVPR*.

Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. 2023. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *ICCV*.