Scenario Generation for Interactive Urban Environments

Paritosh Sharma, Hui-Yin Wu

July 2025

Project members

- Scientific team: Paritosh Sharma, Hui-Yin Wu

1 Context and objectives

The document highlights the work plan for the the WP4 of the ANR Creative 3D [1] project. The expected outcome of this project is to create a generative model that is capable of creating personalized 3D low vision rehabilitation scenarios for urban traffic scenarios which can be used in Virtual Reality (VR) environments.

2 Introduction

Previous works have shown that 3D generative models can be used to create realistic and interactive environments for various applications, including virtual reality (VR) and augmented reality (AR). These can be further used to not only enhance the user experience, but also to provide personalized training and rehabilitation scenarios. In the context of low vision rehabilitation, these models can be particularly useful for simulating urban pedestrian scenarios that are tailored to the specific needs of individuals with visual impairments. However, most generative approaches have focused on self-driving vehicles or general automobile behaviour in urban environments, rather than pedestrian-centric scenarios. This project aims to address this gap by developing a generative model that can create 3D pedestrian urban scenarios with pedestrians as the ego of the scenario.

3 Related Work

In this section, we review the existing literature on generative models for urban scenarios, focusing on grounding techniques, generation paradigms, output types, datasets, validation methods, and limitations. The goal is to provide a comprehensive overview of the current state of research in this area and identify gaps that our project aims to address.

[1] `https://project.inria.fr/creattive3d/`

3.1 Grounding: Input for Generation

Generative models have been applied to traffic scenario generation, where models synthesize plausible urban environments, pedestrian layouts, or vehicle positions from various types of input. An input in the context of a generative model is any data—such as a prompt, image, or scenario graph—provided to influence the output. Grounding, on the other hand, is the process of linking elements of that input to specific, coherent representations in the generated scenario, such as ensuring the "footpath" appears visually plausible, is correctly positioned on the "road", and respects spatial relationships or physical constraints. Grounding ensures the generated scenario isn't just randomly composed but meaningfully aligned with the intended semantics of the input. The common ways to do grounding are,

3.1.1 Rules and Constraints

Rules can be used to give a prescriptive logic that defines how to generate or modify a scenario. Rule-based systems are necessarily procedural, meaning they follow a set of predefined steps or algorithms to create scenarios. Table 1 lists some of the recent rule-based models for scenario generation. Although these systems allow users to define rules for generating scenarios—such as the placement of buildings, roads, and other elements—they often require an intermediate representation which captures the real-world environments. Creating such representations can be challenging, as it requires a deep understanding of the underlying rules and relationships between different elements. Additionally, rule-based systems can be limited in their ability to generate diverse and realistic scenarios, as they rely on predefined rules that may not capture the full complexity of real-world environments.

Model	Technique	Output
CityEngine [35]	Procedural Modeling with Rules	3D scene
Infinigen [36]	Procedural Modeling with Rules	3D scene
MetaUrban [45]	Description Scripts for Scene Layout	3D Scenario

Table 1: Rule-based Models for Urban Scenario Generation

Constraints define conditions that must be satisfied to get a valid scenario. Scenic [18] is a probabilistic programming language that allows users to define constraints for generating scenarios. It uses a declarative approach to specify the properties of the scenario, such as the layout, objects, and their relationships. These can then be rendered using a frontend such as CARLA [12]. However, it suffers from the same limitations of rule-based systems, as it can only generate scenarios that fit within the defined constraints, potentially missing out on the richness and variability of real-world environments.

3.1.2 Prompts

Prompts are textual cues provided to the generative model to guide the scenario creation process. Since the popularity of large language models (LLMs) like GPT-3, prompts have become a common way to interact with generative models. They can be used to specify the desired characteristics of the scenario, such as the type of environment, objects, and their relationships. Table 2 lists some of the recent prompt-based models for scenario generation.

Model	Technique	Output
ScenicNL [14]	Converts LLM Prompts to Scenic Constraints	Scenic Scenario
ChatScene [51]	Conversational Agent for Scenario Definition using Scenic	Scenic Scenario
LayoutGPT [16]	Prompts converted to CSS-like Layout Formatting by LLMs	Layout Representation
ChatDyn [41]	LLM-based layout planning and low-level trajectory generation for Pedestrian and Vehicle	3D Scenario
Work by Feng et al. [17]	JSON to describe layout and 3D models	3D Scene
TTSG [37]	LLM-based planning and retrieval	3D Scenario
3D-SceneDreamer [53]	Prompt to Image followed by Image to 3D	3D Scene
GraphCanvas3D [28]	Uses LLM to create a scene graph and then a 3D scene	3D Scene
Scenethesis [27]	Uses LLM to create a scene graph and then a 3D scene	3D Scene
SceneX [55]	LLM to plan PCG (Procedural Controllable Generation)	3D Scene
Surreal Drivers [22]	Chain-of-thought prompts	3D Scene
Text2nerf [50]	Text prompts	3D Scene
X-Scene [48]	Text prompts	3D Scene

Table 2: Prompt-based Models for Urban Scenario Generation

3.1.3 Layouts

Layouts are structured representations of scenarios that capture the spatial relationships between different elements, such as buildings, roads, and pedestrians. These may include bird's-eye views, top-down maps, or other forms of spatial representations that provide a high-level overview of the scenario. Table 3 lists some of the recent layout-based models for scenario generation.

Model	Technique	Output
CC3D [1]	2D Layout-based 3D Scene Generation	3D Scene
CityDreamer4D [46]	Uses a Layout Generator and a traffic scenario generator	3D Scenario
Infinicube [32]	Text prompts, HD Maps and Bounding Boxes	3D Scenario
Work by Zhang et al. [52]	BEV map	3D Scene
UniScene [23]	BEV map	Multi-view video
Savkin et al. [38]	Scenegraph	Scenario Image

Table 3: Layout-based Models for Urban Scenario Generation

3.1.4 Multimodal

Multimodal inputs combine different types of data, such as text, images, and structured representations, to provide a richer context for scenario generation. Table 4 lists some of the recent multimodal models for scenario generation.

3.2 Generation Paradigms

This section discusses the different paradigms used for generating scenarios in urban environments. These paradigms can be broadly classified into procedural, and diffusion-based approaches.

3.2.1 Procedural Approaches

Procedural approaches use rules or constraints to generate scenes. The most popular procedural approach used for urban scenarios is Scenic [18]. Variations of Scenic have been introduced to allow for more complex scenario generation, such as ScenicNL [14] and ChatScene [51], which use natural language prompts to define scenarios. MetaUrban [45] is a popular urban scenario generation

Model	Input Type	Output
CityX [54]	Prompt, OSM file or Semantic Map	3D Scenario
CityCraft [11]	Layout data and text prompts	3D Scene
Work by Liu et al. [29]	Scenegraph assisted using text prompts	3D Scene
Dynamic City [4]	Commands, Layouts, Object Inpainting and Trajectory	3D Scenario
GAUDI [2]	Conditioning using prompts or Images	3D Scene
MagicDrive3D [19]	Text prompts, Bird Eye View (BEV) map and 3D Bounding Boxes	Reconstructed 3D video
Scene123 [49]	Text prompt, Image or Text Description	3D Scene
StreetScapes [10]	BEV and height map with support for prompts	Video
Urban Architect [31]	3D Layout and Text Prompts	3D Scene
Urban World [39]	Layout (generation) and prompts (refinement)	3D Scenario
Wonderplay [25]	Image and action (physics)	3D Video

Table 4: Multimodal Models for Urban Scenario Generation

framework that combines procedural and declarative approaches to create urban micromobility scenarios. TTSG[37] is another framework that uses an to plan a traffic scenario in JSON and then render it using CARLA [12].

Even though procedural approaches can be highly customizable and allow for precise control over the generated scenarios, they typically suffer from a lack of realism, as they rely on predefined rules that may not capture the full complexity of real-world environments. However, they can be useful in creating diverse scenarios that adhere to specific constraints or requirements. For example, a crossroad with a specific number of lanes, a certain type of road surface, and a defined layout of buildings and other elements. However, no real crossroad would exist in the world that conforms to these specifications.

3.2.2 Generative Approaches

Generative approaches leverage deep learning techniques to learn complex patterns and relationships in data, enabling the generation of scenarios that are more realistic. These models can learn from large datasets of urban environments, capturing the inherent variability and uncertainty present in real-world scenarios. We can further classify these models into the following categories:

Diffusion-based techniques have shown more promise than other generative models such as autoregressors or GANs for 3D generation. These models work by iteratively refining a noisy input to produce a coherent output. They can be trained on large datasets of urban environments, allowing them to capture the complex relationships between different elements in the scene. Table 5 lists some of the recent diffusion-based models for scenario generation.

Model	Technique	Output

Table 5: Diffusion-based Models for Urban Scenario Generation

3.3 Datasets

In this section, we review the datasets that are used to train generative models for road crossing scenarios. Table 6 lists some of the datasets commonly used for training and evaluating generative models in urban environments.

Table 6: Overview of selected datasets for foundation model-based scenario generation and analysis.

Dataset	**Year**	**View**	**Source**	**Dimensionality**
SIND [47]	2022	BEV	–	–
Waymo Open [40]	2020	FPV	Real	3D
Argoverse [8][42]	2023	BEV/FPV	✓	✓
nuScenes [6]	2022	FPV	✓	✓
KITTI [20]	2012	FPV	✓	✓
Cityscapes [9]	2016	FPV	✓	✓
SemanticKITTI [3]	2019	FPV	✓	✓
HoliCity [56]	2020	FPV	✓	✓
OmniCity [24]	2023	FPV	✓	✓
KITTI-360 [26]	2023	FPV	✓	✓
GoogleEarth [21]	2024	FPV	✓	✓
OSM [34]	2024	BEV	✓	✓
CARLA [13]	2017	BEV/FPV	✓	✓
Virtual-KITTI-2 [5]	2020	FPV	✓	✓
CarlaSC [7]	2022	BEV/FPV	✓	✓
CityTopia [15]	2025	BEV/FPV	✓	✓

3.4 Validation

Table 7 lists common qualitative and quantitative validation methods used to evaluate the performance of generative models for urban road-crossing scenarios. These methods assess the quality, realism, and diversity of generated scenarios to ensure they fulfill their intended applications and research needs.

- Forecasting ? (AgentFormer) - XDE for agents?

4 Code

Currently tested code include:

4.1 Scenic

Easy to setup and run since the repository is well maintained.

Table 7: Qualitative and Quantitative Evaluation Methods for Road-Crossing Scenarios

Method Type	**Evaluation Approach**	**Description**	**Use Case**
Qualitative	Human Review	Human experts assess realism, scenario diversity, and layout plausibility.	Validates human-perceived quality and applicability of the scene.
	Scenario Visualization	3D visual inspection or rendered videos showing pedestrian, vehicle, and environment interactions.	Helps detect unrealistic or unnatural behavior/configurations.
	Surveys	Collects subjective feedback on realism, perceived difficulty, or stress levels from participants.	Measures human-centric realism or emotional response.
	Comparison	Compare scenario features (traffic density, gap opportunities, layout) to real-world statistics or distributions.	Validates realism by matching key scenario characteristics to empirical data.
	Failure Cases	Identification and analysis of implausible or unsafe scenarios generated by the model.	Guides iterative model improvement.
Quantitative	Scenario Feature Statistics	Statistical analysis of scenario properties (traffic speed, number/duration of gaps, crosswalk presence, etc.).	Ensures generated scenarios span realistic distributions.
	Coverage and Diversity Metrics	Measures distributional entropy or spread of key attributes across all generated scenarios.	Assesses generalizability, scenario variety, and model's exploration of edge cases.
	Criticality and Opportunity Metrics	Quantifies the frequency and severity of challenging situations (number of safe gaps, minimum feasible gap size, "no-go" cases).	Evaluates risk/challenge spectrum in the scenario catalog.
	Sim2Real Gap (Domain Distance)	Computes metrics like KL-divergence, Earth Mover's Distance, t-SNE or FID/KID between generated and real scenario feature distributions.	Evaluates how closely synthetic scenarios match reality.
	Controllability Metrics	Measures how well the model can generate scenarios based on prompts (CLIP, BLIP, VQA, etc.)	Assesses controllability and flexibility of the generative model.

4.2 MagicDrive

Had issues running due to model size and GPU memory limitations. Also realised only 2D video generation works for now since the code is not uploaded yet for 3D generation.

4.3 MetaUrban/ScenarioNET

Easy to setup and run since docker container is available and working.

4.4 Threestudio

Threestudio [30] is a collection of 3D generative techniques. Easy to setup and run since docker container is provided. It is well maintained by the community and has a good documentation. Thus, good to test different generative techniques for 3D models.

4.5 City Dreamer

CityDreamer4D [46] tried testing the 3D branch with static scenes but encountered issues with the container.

4.6 Urban Architect

Urban Architect [31] could be useful as a starting point for generating the static urban scenario using diffusion models and score distillation sampling. For now the container has dependency issues with pytorch3d.

4.7 Problem and Open Questions

Most existing scenario-generation methods are vehicle-centric, focusing primarily on creating synthetic scenarios to train autonomous driving agents. A few works such as Jaywalkingvr [33] and Creative3D[44] use a set of real and synthetic scenarios to explore pedestrian behaviour in VR. They however rely on a fixed set of scenarios and interactions. Thus, developing a model that generates diverse pedestrian-centric scenarios can be a significant contribution. The open questions are then:

- **What is the best way to ground an interactive scenario for a generative model:** Based on the literature and continuing work by Wu et al. [43], we can use ontological grounding, where we define a scenegraph that captures the relationships between different elements in the scenario, such as pedestrians, vehicles, and the environment. A good example of interactive scenegraph editing and then 3D scene generation is shown in the demo video (`https://www.youtube.com/watch?v=4YRTydsv-qg`) of GraphCanvas3D [28].
- **How do we create these scenarios allowing for diverse pedestrian behaviors:**
- **How to evaluate interactive scenarios:**

5 Work Plan

Appendix

6 Misc

6.1 Meeting Notes

6.2 01/08/2025

- **Clarify Image Usage:** Specify what kinds of images are being referred to in the input for the models (e.g., diagrams, real-world photos, layouts), and justify their relevance.
- **Add details on Scene Graph:** Provide more information about the scene graphs (e.g., format, usage, etc.) in the context of the models.
- **Avoid Overly High-Level Descriptions:** Some explanations are too abstract — add concrete examples and ensure essential implementation and interaction details are included.

- **Improve Focus on Vocabulary:** Review and refine terminology. Ensure technical or domain-specific terms are clearly defined and used consistently.
- **Identify and Specify Missing Interaction Types:** Clearly outline which user/system interactions are missing or underexplored.
- **Sharpen Research Question (RQ):** Make the RQ more concrete and closely tied to your data, methods, and intended contributions.
- **Better Categorization of Literature:** Reorganize cited papers using clear categories such as research themes, methodologies, etc.

References

[1] S. Bahmani, J. J. Park, D. Paschalidou, X. Yan, G. Wetzstein, L. Guibas, and A. Tagliasacchi. Cc3d: Layout-conditioned generation of compositional 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7171–7181, 2023.

[2] M. A. Bautista, P. Guo, S. Abnar, W. Talbott, A. Toshev, Z. Chen, L. Dinh, S. Zhai, H. Goh, D. Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *Advances in Neural Information Processing Systems*, 35:25102–25116, 2022.

[3] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.

[4] H. Bian, L. Kong, H. Xie, L. Pan, Y. Qiao, and Z. Liu. Dynamiccity: Large-scale lidar generation from dynamic scenes. *arXiv e-prints*, pages arXiv–2410, 2024.

[5] Y. Cabon, N. Murray, and M. Humenberger. Virtual kitti 2, 2020.

[6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[7] CarlaSC Team. Carlasc: A dataset and network for real-time semantic mapping in dynamic environments. 2022. Dataset website: https://umich-curly.github.io/CarlaSC.github.io/.

[8] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] B. Deng, R. Tucker, Z. Li, L. Guibas, N. Snavely, and G. Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[11] J. Deng, W. Chai, J. Huang, Z. Zhao, Q. Huang, M. Gao, J. Guo, S. Hao, W. Hu, J.-N. Hwang, et al. Citycraft: A real crafter for 3d city generation. *arXiv preprint arXiv:2406.04983*, 2024.

[12] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[13] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[14] K. Elmaaroufi, D. Shanker, A. Cismaru, M. Vazquez-Chanlatte, A. Sangiovanni-Vincentelli, M. Zaharia, and S. A. Seshia. Scenicnl: generating probabilistic scenario programs from natural language. *arXiv preprint arXiv:2405.03709*, 2024.

[15] Z. et al. Citytopia: A large-scale synthetic dataset for 3d cities, 2025. CityTopia synthetic dataset for 3D cities.

[16] W. Feng, W. Zhu, T.-j. Fu, V. Jampani, A. Akula, X. He, S. Basu, X. E. Wang, and W. Y. Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36:18225–18250, 2023.

[17] Y. Feng, J. Jiang, J. Ren, W. Li, R. Li, and X. Fan. Text-guided editable 3d city scene generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[18] D. J. Fremont, T. Dreossi, S. Ghosh, X. Yue, A. L. Sangiovanni-Vincentelli, and S. A. Seshia. Scenic: a language for scenario specification and scene generation. In *Proceedings of the 40th ACM SIGPLAN conference on programming language design and implementation*, pages 63–78, 2019.

[19] R. Gao, K. Chen, Z. Li, L. Hong, Z. Li, and Q. Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024.

[20] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[21] Google. Google earth, 2024.

[22] Y. Jin, R. Yang, Z. Yi, X. Shen, H. Peng, X. Liu, J. Qin, J. Li, J. Xie, P. Gao, et al. Surrealdriver: Designing llm-powered generative driver agent framework based on human drivers' driving-thinking data. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 966–971. IEEE, 2024.

[23] B. Li, J. Guo, H. Liu, Y. Zou, Y. Ding, X. Chen, H. Zhu, F. Tan, C. Zhang, T. Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11971–11981, 2025.

[24] W. Li, Y. Lai, L. Xu, Y. Xiangli, J. Yu, C. He, G.-S. Xia, and D. Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[25] Z. Li, H.-X. Yu, W. Liu, Y. Yang, C. Herrmann, G. Wetzstein, and J. Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. *arXiv preprint arXiv:2505.18151*, 2025.

[26] Y. Liao, J. Xie, and A. Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.

[27] L. Ling, C.-H. Lin, T.-Y. Lin, Y. Ding, Y. Zeng, Y. Sheng, Y. Ge, M.-Y. Liu, A. Bera, and Z. Li. Scenethesis: A language and vision agentic framework for 3d scene generation. *arXiv preprint arXiv:2505.02836*, 2025.

[28] L. Liu, S. Chen, S. Jia, J. Shi, Z. Jiang, C. Jin, W. Zongkai, J.-N. Hwang, and L. Li. Graph canvas for controllable 3d scene generation. *arXiv preprint arXiv:2412.00091*, 2024.

[29] Y. Liu, X. Li, Y. Zhang, L. Qi, X. Li, W. Wang, C. Li, X. Li, and M.-H. Yang. Controllable 3d outdoor scene generation via scene graphs. *arXiv preprint arXiv:2503.07152*, 2025.

[30] Y.-T. Liu, Y.-C. Guo, V. Voleti, R. Shao, C.-H. Chen, G. Luo, Z. Zou, C. Wang, C. Laforte, Y.-P. Cao, et al. Threestudio: A modular framework for diffusion-guided 3d generation. *cg. cs. tsinghua. edu. cn*, 2023.

[31] F. Lu, K.-Y. Lin, Y. Xu, H. Li, G. Chen, and C. Jiang. Urban architect: Steerable 3d urban scene generation with layout prior. *arXiv preprint arXiv:2404.06780*, 2024.

[32] Y. Lu, X. Ren, J. Yang, T. Shen, Z. Wu, J. Gao, Y. Wang, S. Chen, M. Chen, S. Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. *arXiv preprint arXiv:2412.03934*, 2024.

[33] K. Mukoya, E. Weng, R. Choudhury, and K. Kitani. Jaywalkervr: A vr system for collecting safety-critical pedestrian-vehicle interactions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9600–9607. IEEE, 2024.

[34] OpenStreetMap contributors. Openstreetmap, 2024.

[35] Y. I. Parish and P. Müller. Procedural modeling of cities. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 301–308, 2001.

[36] A. Raistrick, L. Lipson, Z. Ma, L. Mei, M. Wang, Y. Zuo, K. Kayan, H. Wen, B. Han, Y. Wang, et al. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12630–12641, 2023.

[37] B.-K. Ruan, H.-T. Tsui, Y.-H. Li, and H.-H. Shuai. Traffic scene generation from natural language description for autonomous vehicles with large language model. *arXiv preprint arXiv:2409.09575*, 2024.

[38] A. Savkin, R. Ellouze, N. Navab, and F. Tombari. Unsupervised traffic scene generation with synthetic 3d scene graphs. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1229–1235. IEEE, 2021.

[39] Y. Shang, Y. Lin, Y. Zheng, H. Fan, J. Ding, J. Feng, J. Chen, L. Tian, and Y. Li. Urbanworld: An urban world model for 3d city generation. *arXiv preprint arXiv:2407.11965*, 2024.

[40] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[41] Y. Wei, J. Wang, Y. Du, D. Wang, L. Pan, C. Xu, Y. Feng, B. Dai, and S. Chen. Chatdyn: Language-driven multi-actor dynamics generation in street scenes. *arXiv preprint arXiv:2412.08685*, 2024.

[42] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

[43] H.-Y. Wu, F. Robert, T. Fafet, B. Graulier, B. Passin-Cauneau, L. Sassatelli, and M. Winckler. Designing guided user tasks in vr embodied experiences. *Proceedings of the ACM on Human-Computer Interaction*, 6(EICS):1–24, 2022.

[44] H.-Y. Wu, F. A. S. Robert, F. F. Gallo, K. Pirkovets, C. Quere, J. Delachambre, S. Ramanoël, A. Gros, M. Winckler, L. Sassatelli, et al. Exploring, walking, and interacting in virtual reality with simulated low vision: a living contextual dataset (2023), 2023.

[45] W. Wu, H. He, Y. Wang, C. Duan, J. He, Z. Liu, Q. Li, and B. Zhou. Metaurban: A simulation platform for embodied ai in urban spaces. *arXiv e-prints*, pages arXiv–2407, 2024.

[46] H. Xie, Z. Chen, F. Hong, and Z. Liu. Citydreamer4d: Compositional generative model of unbounded 4d cities. *arXiv e-prints*, pages arXiv–2501, 2025.

[47] Y. Xu, W. Shao, J. Li, K. Yang, W. Wang, H. Huang, C. Lv, and H. Wang. Sind: A drone dataset at signalized intersection in china. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2471–2478. IEEE, 2022.

[48] Y. Yang, A. Liang, J. Mei, Y. Ma, Y. Liu, and G. H. Lee. X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. *arXiv preprint arXiv:2506.13558*, 2025.

[49] Y. Yang, F. Yin, J. Fan, X. Chen, W. Li, and G. Yu. Scene123: One prompt to 3d scene generation via video-assisted and consistency-enhanced mae. *arXiv preprint arXiv:2408.05477*, 2024.

[50] J. Zhang, X. Li, Z. Wan, C. Wang, and J. Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 30(12):7749–7762, 2024.

[51] J. Zhang, C. Xu, and B. Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15459–15469, 2024.

[52] J. Zhang, Q. Zhang, L. Zhang, R. R. Kompella, G. Liu, J. Li, and B. Zhou. Urban scene diffusion through semantic occupancy map. *arXiv preprint arXiv:2403.11697*, 2024.

[53] S. Zhang, Y. Zhang, Q. Zheng, R. Ma, W. Hua, H. Bao, W. Xu, and C. Zou. 3d-scenedreamer: Text-driven 3d-consistent scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10170–10180, 2024.

[54] S. Zhang, M. Zhou, Y. Wang, C. Luo, R. Wang, Y. Li, Z. Zhang, and J. Peng. Cityx: Controllable procedural content generation for unbounded 3d cities. *arXiv preprint arXiv:2407.17572*, 2024.

[55] M. Zhou, Y. Wang, J. Hou, S. Zhang, Y. Li, C. Luo, J. Peng, and Z. Zhang. Scenex: Procedural controllable large-scale scene generation. *arXiv preprint arXiv:2403.15698*, 2024.

[56] Y. Zhou, J. Huang, X. Dai, L. Luo, Z. Chen, and Y. Ma. Holicity: A city-scale data platform for learning holistic 3d structures, 2021.