Scenario Generation for Interactive Urban Environments

Paritosh Sharma, Hui-Yin Wu

September 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 training scenarios in urban environments.

2 Introduction

Virtual Reality (VR) and Augmented Reality (AR) technologies have advanced significantly in recent years, enabling the creation of immersive and interactive environments for various applications. These can be further used to provide personalized training and rehabilitation scenarios. In the context of low vision rehabilitation, these models can be particularly useful to study pedestrian behaviours under normal and simulated vision. However, most simulated environments suffer from perceptual gaps between the designer, the user, and the system. The existing GusT-3D framework [58], developed during the Creative3D project, provides a foundation to address this. However, the current framework is still limited by its reliance on a fixed set of urban environments and interactions innhibiting its ability to scale.

In parallel, recent works on 3D generative models for urban scenario generation has shown promising results in generating diverse and realistic urban environments. Additionally, these models have been able to capture the diverse nature and the complexities of an urban setting, including the interactions between pedestrians, vehicles, and the environment. Driven by these extraordinary capabilities, exploring the potential of generative models will enable us to create more diverse and realistic urban scenarios.

3 Related Work

In this section, we first start by reviewing prior work on pedestrian-in-the-loop simulations. Then, we review recent works on generative models for urban scenario generation as well as diffusion-based

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

4D generation. Finally, we discuss common validation methods used to evaluate the performance of these models.

3.1 Pedestrian-in-the-Loop

VR Simulations have been widely used to study pedestrian behavior and interactions in urban scenarios [57, 51, 37, 47]. This human-in-the-loop approach (also referred as pedestrian-in-the-loop[21]) allows researchers to collect data on how pedestrians interact with their environment, including their decision-making processes, movement patterns, and responses to various stimuli. These simulations can be used to study a wide range of scenarios, from simple road crossings to complex urban environments with multiple interacting agents. More recently, JaywalkerVR [37], a VR human-in-the-loop simulator, used CARLA [11] to create four different scenarios: jaywalking, parked cars, four-way stops, and parking lot entrances. Despite the obvious advantages, developing such simulators is still challenging because of the perceptual gaps between the designer, the user and the system as identified by Dourish [12].

The GUsT-3D framework [58] addresses this by first defining the scene using a scenegraph, which captures the relationships between different elements in the scenario (ontology) and then defining the task to be carried out during the course of the scenario using a GUTasks (intersubjectivity). Lastly, it uses a query component for logging and post-scenario analysis of the experience (intentionality). This framework was also applied by creating a dataset of 6 road-crossing scenarios to study pedestrian behavior under normal and low vision [59]. Even though GusT-3D framework addresses the perceptual gaps which were identified by Dourish, it still relies on a fixed set of scenarios and interactions.

3.2 Generative Models for Urban Scenario Generation

3.2.1 Image-based Generation

Diffusion models have shown impressive results in generating high-quality images from textual descriptions. Models like DALL-E 2 [42], Imagen [45], and Stable Diffusion [43] have demonstrated the ability to create diverse and realistic images based on text prompts. Recent works such as GameNGen [52], Genie [5], and DIAMOND [1] have shown the potential of diffusion models to generate game environments in real-time. However, these models suffer from latency issues since they are not truly 3D and generate image-by-image.

3.2.2 Prompts

Prompts are textual cues provided to the generative model to guide the scenario creation process. Since the popularity of 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 1 lists some of the recent prompt-based models for scenario generation.

3.3 Layout-guided Generation

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,

Model	Technique	Output
ScenicNL [13]	Converts LLM Prompts to Scenic Constraints	Scenic Scenario
ChatScene [68]	Conversational Agent for Scenario Definition using Scenic	Scenic Scenario
LayoutGPT [14]	Prompts converted to CSS-like Layout Formatting by LLMs	Layout Representation
ChatDyn [54]	LLM-based planning and low-level trajectory generation for Pedestrian and Vehicle	3D Scenario
Work by Feng et al. [15]	JSON to describe layout and 3D models	3D Scene
TTSG [44]	LLM-based planning and retrieval	3D Scenario
3D-SceneDreamer [70]	Prompt to point-of-view image followed by image to 3D	3D Scene
GraphCanvas3D [32]	Uses LLM to create a scenegraph and then a 3D scene	3D Scene
Scenethesis [31]	Uses LLM to create a scenegraph and then a 3D scene	3D Scene
SceneX [73]	LLM to plan PCG (Procedural Controllable Generation)	3D Scene
Surreal Drivers [23]	Chain-of-thought prompts	3D Scene
Text2nerf [67]	Text prompts	3D Scene
X-Scene [64]	Text prompts	3D Scene

Table 1: Prompt-based Models for Urban Scenario Generation

top-down maps, or other forms of spatial representations that provide a high-level overview of the scenario. Table 2 lists some of the recent layout-based models for scenario generation.

Model	Technique	Output
CC3D [2]	2D Layout-based 3D Scene Generation	3D Scene
CityDreamer4D [62]	Uses a Layout Generator and a traffic scenario generator	3D Scenario
Infinicube [36]	Text prompts, HD Maps and Bounding Boxes	3D Scenario
Work by Zhang et al. [69]	BEV map	3D Scene
UniScene [25]	BEV map	Multi-view video
Savkin et al. [46]	Scenegraph	Scenario Image

Table 2: Layout-based Models for Urban Scenario Generation

Additionally, some works have also combined multimodal inputs with different types of data, prompts, layouts and other structured representations, to provide a richer context for scenario generation. Table 3 lists some of the recent multimodal models for scenario generation.

Model	Input Type	Output
CityX [71]	Prompt, OSM file or Semantic Map	3D Scenario
CityCraft [10]	Layout data and text prompts	3D Scene
Work by Liu et al. [33]	Scenegraph assisted using text prompts	3D Scene
GAUDI [4]	Conditioning using prompts or point-of-view images	3D Scene
MagicDrive3D [18]	Text prompts, Bird Eye View (BEV) map and 3D Bounding Boxes	Reconstructed 3D video
Scene123 [65]	Text prompt, point-of-view image or Text Description	3D Scene
StreetScapes [9]	BEV and height map with support for prompts	Video
Urban Architect [35]	3D Layout and Text Prompts	3D Scene
Urban World [48]	Layout (generation) and prompts (refinement)	3D Scenario
Wonderplay [28]	point-of-view image and action (physics)	3D Video

Table 3: Multimodal Models for Urban Scenario Generation

3.4 Diffusion-based 4D Generation

Diffusion models have emerged as a powerful approach for generating high-quality 3D content by iteratively refining a random noise input into a coherent output. DreamFusion [39] initially introduced text-to-3D synthesis by optimizing a neural radiance field so that rendered views, when noised and denoised by a pretrained text-conditioned diffusion model, produce identical gradients via Score Distillation Sampling (SDS). This framework was extended to dynamic scenes in MaV3D, which employs video-based SDS to animate a time-conditioned radiance field into a 4D scene [49]. However, MaV3D's reliance on a single diffusion prior leads to trade-offs between appearance fidelity, 3D consistency, and motion realism. 4D-fy [3] addresses this by hybridizing three SDS signals—3D-aware image SDS for geometry, standard image SDS for texture, and video SDS for motion—alternating updates to preserve all three qualities. Animate124 [72] further refines single-image animation into 4D using a coarse-to-fine 4D grid backbone optimized first with 2D and 3D image priors, then with video diffusion, and finally with a personalized ControlNet fine-tuning stage to prevent semantic drift. More recently, Trans4D [66] leverages a multimodal large language model to perform physics-aware 4D scene planning—generating object trajectories, rotations, and transition times—and introduces a Transition Network that predicts point-wise appearance or disappearance probabilities to realize complex geometry-aware transitions such as a missile transforming into an explosion cloud. Despite their impressive generative capabilities, these diffusion-based 4D synthesis methods still lack the explicit, structured control afforded by scene graphs, making it difficult to enforce complex object relationships or constraints at generation time.

3.5 Validation

Table 4 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.

3.6 Code

Currently tested code includes:

3.6.1 Scenic

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

3.6.2 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.

3.6.3 MetaUrban/ScenarioNET

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

3.6.4 Threestudio

Threestudio [34] 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 / refer different generative techniques for 3D.

Table 4: 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.

3.6.5 City Dreamer

CityDreamer4D [62] tried testing the 3D branch with static scenes but encountered issues with the docker container.

3.6.6 Urban Architect

Urban Architect [35] - Working. Can generate semantic maps and depth maps from a layout input. Which then can be used to generate the 3D scene. However, the code has to be adapted to generate using multiple GPUs due VRAM limitations.

3.6.7 CLIP

CLIP [40] is a model that can be used to evaluate the quality of generated scenarios. Tested on colab with the pretrained ViT model.

3.6.8 GraphDreamer

GraphDreamer [17] works but generation takes over 30 hours for 1 scenegraph. However, cannot generate complex scenes.

4 Problem and Open Question

While recent generative models have demonstrated strong capabilities in synthesizing urban scenes from static layouts, they remain limited in their ability to model dynamic urban scenarios. In particular, they struggle to provide precise spatial and temporal control over objects and their interactions, which is essential for simulating a realistic traffic scenario.

This leads to the following open research question:

How can we design a generative model that can create dynamic urban scenarios with precise spatial and temporal control over objects and their interactions?

5 Work Plan

5.1 Model Architecture

Figure 1 illustrates the proposed model architecture for generating dynamic urban scenarios. The model takes a scenario prompt as input, which is processed by a multimodal large language model (MLLM) to extract two key components: layout and tasks. Each component is then handled by specialized modules to generate the final 3D dynamic scene.

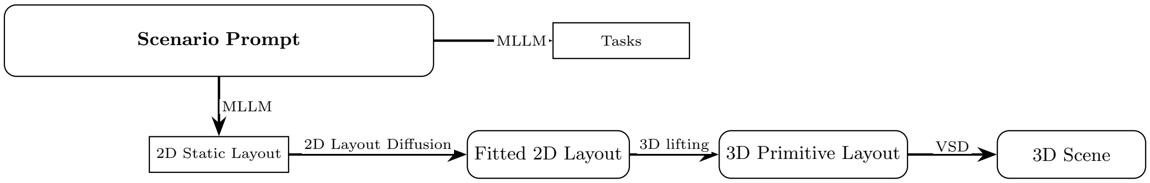


Figure 1: Model Architecture

5.2 Input and Splitting

Since no existing datasets provide paired text prompts with corresponding urban layouts and tasks, we propose leveraging the capabilities of MLLMs to parse and structure input prompts. Similar approaches have been explored in prior works, such as ChatDyn [54] for action planning, LayoutGPT [14] for layout generation, and LLM-grounded Diffusion [29] for LLM-conditioned image synthesis.

We first use a multimodal large language model (MLLM) (eg. Gemini) to process the input prompt and split it into the layout and tasks. The classes in the layout are based on the KITTI-360 [30] dataset. Here is an example of how the input prompt can be split:

Scene Specification Format

Prompt:
You are an urban scenario planning assistant with a pedestrian agent. For any urban scene prompt, parse the scene and output a strictly formatted JSON object containing:

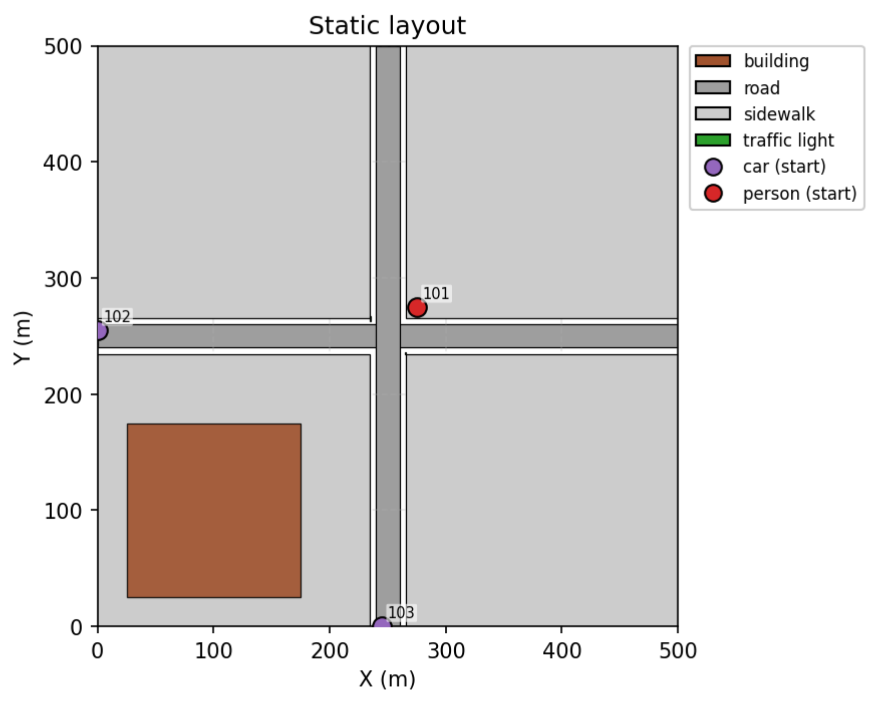
- **static_layout:** A 500 × 500 2D layout map represented as an array of objects. Each object must include:
  - `id`: unique identifier
  - `type`: one of {pole, traffic sign, smallpole, lamp, trash bin, ground, road, sidewalk, parking, building, garage, fence, gate, vegetation, terrain, rail track, wall, box, vending machine, traffic light, rider, bicycle, motorcycle, motorbike, car, truck, bus, van, trailer, caravan, person}
  - `position`: (x, y) coordinates in meters within the 2D plane
  - `orientation`: angle in degrees (0-360)
  - `size`: width and height in meters
- **dynamic_layout:**
  - `trajectories`: Array of dynamic objects representing their movement over time. Each object must include:
    - `id`: unique identifier
    - `type`: one of {traffic light, rider, bicycle, motorcycle, motorbike, car, truck, bus, van, trailer, caravan, person}
    - `initial_position`: (x, y) coordinates at time = 0
    - `trajectory_description`: an array of states, where each state is of the form {”time”: t, ”position”: [x, y]} with time normalized between 0 and 1
- **tasks:** ordered list of actions to be performed by the pedestrian

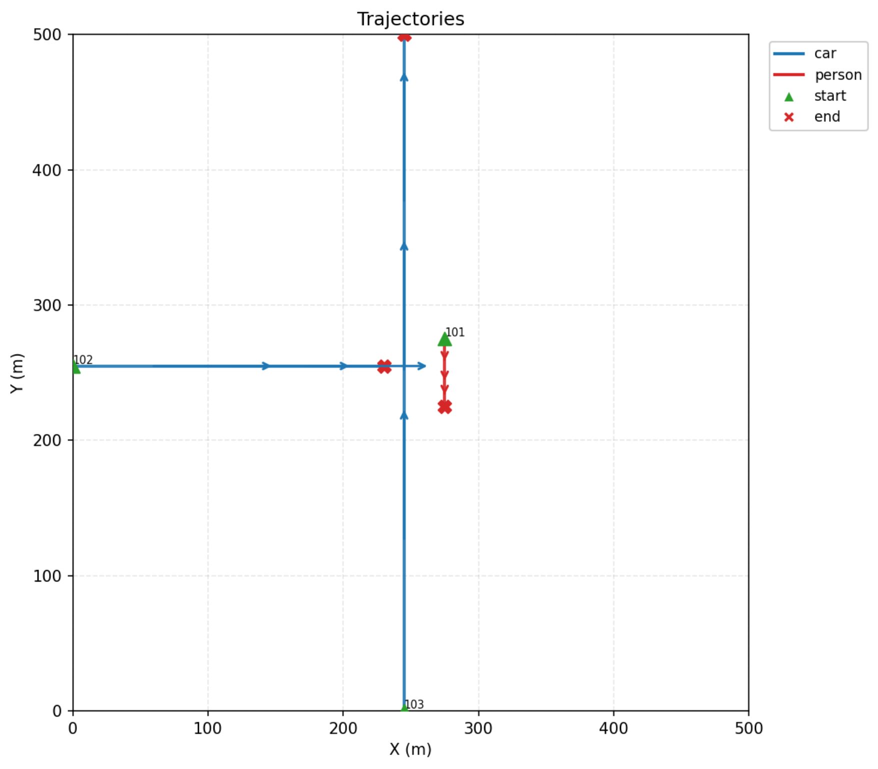
If details are missing, fill with realistic defaults. Only output valid JSON.

Example input:

A busy urban intersection with sidewalks, two cars and a pedestrian crossing. Include a building and a traffic light.
Example Output:

Static layout:





Trajectories:
Tasks:

- Walk from your starting position at [275, 275] to the edge of the crosswalk at [275, 260].
- Wait for the pedestrian signal to allow crossing the horizontal road.
- Cross the street to the opposite sidewalk, arriving at [275, 240].

5.3 Static Scene Generation

The next step is to generate a static 3D scene from the bouding boxes predicted by the MLLM.

1. Preparation of Conditional Inputs Given the predicted bounding box layout predicted by the MLLM, we build a coarse instance-wise semantic segmentation map. We additionally define a *spatial prior mask* encoding object classes (traffic light, car, etc.) as a 2D array. Both the instance layout and the room mask are embedded via trainable embedding layers, yielding conditioning vectors that guide the subsequent diffusion process. The desired room type, if applicable, is also embedded and added to the timestep embedding.

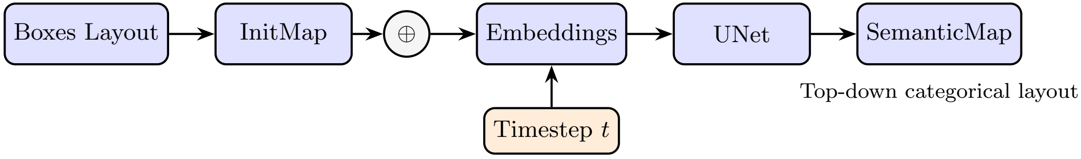
2. Multinomial Diffusion for Layout Synthesis The core of our system is a multinomial (discrete) diffusion model adapted for semantic segmentation. At each denoising timestep t, a

UNet backbone receives:

- the current noisy semantic map (as a categorical one-hot tensor),
- a sinusoidal positional embedding of the diffusion timestep.

The UNet predicts logits for categorical class distributions per pixel. During training, we minimize the variational bound (KL divergence) between the true and predicted posteriors as in [22]. During inference, the layout and mask embeddings provide strong spatial and semantic control, ensuring the generated layout adheres to both the LLM-predicted spatial arrangement and architectural constraints.

3. Output The final output is a top-down semantic map representing the static elements of the 3D scene to be generated. This map can then be used with score distillation sampling (as shown in Urban Architect [35]) to create the full 3D scene.



5.4 Tasks

The tasks predicted by the MLLM are a sequence of actions to be performed by the player.

6 Implementation

Appendix

7 Misc

7.1 Meeting Notes

7.1.1 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).
- **Add details on scenegraph:** Provide more information about the scenegraphs (e.g., format, usage, etc.) in the context of the models.
- **Avoid Overly High-Level Descriptions:** Some explanations are too abstract.
- **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.
- **Better Categorization of Literature:** Reorganize cited papers using clear categories such as research themes, methodologies, etc.

7.1.2 14/08/2025

- **RQ still too high level:** Provide more information about the scenegraphs (e.g., format, usage, etc.) in the context of the models.
- **Clarify the missing pieces in the SOTA:**
- **How to address:** diversity -¿ generative models, realism -¿ ontology, dynamic -¿ tasks
- **Start with WorkPlan:**
- **Remove relation between GusT-3D and generation**

7.1.3 04/09/2025

- **Semantic Information Missing:** Find a way to match the scene to Kitti's prior.
- **Input unclear:** Scenegraph requires too much information. Need to simplify the input.
- **Use Layout:** Graphdreamer based score distillation doesn't work, use layout instead.

7.1.4 29/09/2025

- **Organize docs and notes:** Clarify the layout preferably with images.
- **Add missing citations:** Add missing citations for stuff like scene background generation in the workplan.
- **Detail:** Detail the missing sections more.

7.2 Urban Scenario Generation

This section discusses pre-existing models used for generating scenarios in urban environments. These models can be broadly classified into procedural and deep generative approaches.

7.2.1 Procedural Methods

Procedural methods leverage algorithms and rules to generate urban scenarios, often resulting in highly customizable environments. They can be further divided into two categories:

Classic Procedural Generation use rules or constraints to generate scenes. The most popular procedural approach used for urban scenarios is Scenic [16], which allows users to define scenarios using a probabilistic programming language. Similarly, MetaUrban [60] is a popular urban scenario generation framework to create urban micromobility scenarios using the Metadrive [26] simulator.

LLM-based Procedural Generation can be used to enhance procedural generation with natural language understanding. For example, ScenicNL [13] and ChatScene [68] use LLMs to generate scenic code from prompts and then define the scenarios in Scenic. TTSG[44] is another framework that uses an to plan a traffic scenario using an LLM in JSON and then render it using CARLA [11]. CityX [71] is multi-agent framework that uses LLM to assemble assets using the procedural content generation (PCG) library as well as plan actions for the agents in the scenario.

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 scenario with a crossroad with a specific number of lanes, a certain type of road surface, and a defined layout of buildings. However, no real crossroad would exist in the world that conforms to these specifications.

7.2.2 Deep-Generative Methods

Deep-Generative Approaches have recently become popular for various sorts of 3D generation. To create a complete scenario, some techniques use generative models such as GANs, VAEs or Diffusion models in combination with procedural techniques. These can be broadly classified into the following categories:

Procedural Environments with Deep-Generative Dynamics generate the static aspects of the environment using procedural techniques, while the dynamic aspects such as vehicles and pedestrians are generated using generative models. For example, Chatdyn [54] uses CARLA [11] to construct the traffic environment, then populate it with pedestrians and vehicles, each equipped with an LLM agent to generate a high-level scenario plan. This high-level plan is then executed in a low-level PedExecutor using Text2Motion [20] for pedestrians and VehExectutor using a physics-based, history-aware reinforcement learning controller to produce vehicle trajectories.

Procedural Dynamics with Deep-Generative Environments generate the dynamic aspects such as vehicles and pedestrians using procedural techniques. UrbanWorld [48] converts 2.5D urban layout data into a structured 3D city with separated assets, applying depth-aware, multi-view diffusion-based texture rendering and UV inpainting to achieve high-fidelity visuals. It also uses an urban MLLM for designing the world and provides dynamic elements which are planned using a random tree path planning algorithm. CityDreamer4D [61] modularly integrates autoregressive token-based generation, neural rendering (e.g., NeRF-style volumetrics), and procedural traffic modeling to synthesize large-scale, time-varying 3D cities.

Complete Deep-Generative Approaches recreate the static as well as the dynamic aspects using the learnt representation. Infinicube [36] constructs a large, dynamic 3D voxel world from input HD maps, vehicle positions, and text prompts; then, it generates photorealistic driving videos and reconstructs the scene into a manipulable 3D environment by fusing voxel- and video-based representations. UniScene [25] UniScene generates driving scenes in three modalities—semantic occupancy, multi-view video, and LiDAR—by first producing a controllable, temporally consistent occupancy sequence from BEV layouts. This occupancy then serves as unified geometric–semantic guidance to synthesize realistic videos and point clouds, ensuring cross-modal consistency and editability.

7.3 Scenegraph-Controlled Diffusion

Despite the advancements in urban scenario generation, very few methods have explored the use of scenegraphs to control the generation process. In recent years, the achievements in text-to-image generation has enabled the advancement of text-to-3D generation using Score Distillation Sampling (SDS) [39] which optimizes a 3D model by aligning 2D images rendered at arbitrary viewpoints with the distribution derived from a text-conditioned diffusion model. Subsequent works including ProlificDreamer [53] introduced Variational Score Distillation (VSD) which addresses the over-regularization and mode collapse issues of SDS by introducing a variational formulation that jointly optimizes the 3D representation and a learnable Gaussian noise distribution, enabling more faithful geometry and richer texture generation. GraphDreamer [17] uses the the SDS process with scenegraphs to enable more structured and controllable scene generation. However, it uses a simplified static scenegraph representation as shown in Visual genome [24] in which, nodes represent the objects in the scene, edges represent the relationship between these nodes and attriubtes represent the properties of the node. Example, Elderly (attribute) man (Node) wearing (edge) a hat(node). This limits the ability to generate dynamic scenarios where the relationships between objects change over time.

7.4 Scenegraph-guided Generation

Scenegraph-guided 3D generation is the process of creating 3D environments by leveraging scene graphs, which are structured representations of the objects and their relationships within a scene. A major advantage of scenegraph-guided generation is that it provides a clear relationship between the elements which allows for better generation than other techniques such as simple text prompts or bounding boxes. Scenethesis [31] uses a Vision-Language Model (VLM) to create a scenegraph with parent-child relationships and localizes objects using 3D bounding boxes. Work by Liu et al. [33] defines scenegraphs as graphs where instance nodes represent countable objects with semantic and positional features, a singleton road node encodes global scene structure, and edges capture both physical proximity among instances and connectivity to the road. X-Scene [64] is another work that uses LLMs (Large-Language Models) to create a scenegraph with nodes (objects) and edges (relationships) to facilitate the generation process. Graphdreamer [17] employs scenegraphs structured around the Visual Genome [24] format, where nodes represent objects with associated attributes, and edges encode the relationships between these objects to guide the generation process. Despite the advancements in scenegraph-guided generation, all of these works focus on static scenes and do not consider dynamic scenarios where the state of the objects changes over time.

7.5 Grounding: Input for Scene Generation

Generative models have already been applied to urban 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,

7.5.1 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 5 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.traffic

Model	Technique	Output
CityEngine [38]	Procedural Modeling with Rules	3D scene
Infinigen [41]	Procedural Modeling with Rules	3D scene
MetaUrban [60]	Description Scripts for Scene Layout	3D Scenario

Table 5: Rule-based Models for Urban Scenario Generation

Constraints define conditions that must be satisfied to get a valid scenario. Scenic [16] 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 [11]. 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.

7.6 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.

7.7 Input Scenegraph

We represent a temporal (dynamic) scenegraph as a tuple

$$S = (V, E, \mathcal{T}, \tau) \tag{1}$$

where

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

Dataset	**Year**	**View**	**Source**
SIND [63]	2022	BEV	Real
Waymo Open [50]	2020	FPV	Real
Argoverse [7][55]	2023	BEV/FPV	Real
nuScenes [6]	2022	FPV	Real
KITTI [19]	2012	FPV	Real
Cityscapes [8]	2016	FPV	..
HoliCity [74]	2020	FPV	..
OmniCity [27]	2023	FPV	..
GoogleEarth [61]	2024	BEV	Real
OSM [61]	2024	BEV	Real
CarlaSC [56]	2022	BEV/FPV	Synthetic
CityTopia [62]	2025	BEV/FPV	..

- $V = \{v_i\}_{i=1}^{N}$ is a finite set of nodes (objects). Each node v_i is itself a tuple

$$v_i = \big(\mathrm{id}_i, \mathcal{A}_i^{\text{static}}, \{(I_{i,k}, \mathcal{A}_i^{\text{dynamic}}|_{I_{i,k}})\}_{k=1}^{K_i}\big), \tag{2}$$

where

  - id_i is a unique identifier (string),
  - $\mathcal{A}_i^{\text{static}}$ is the attribute map for invariant attributes:

$$\mathcal{A}_i^{\text{static}} : K_{\text{static}} \rightharpoonup V,$$

e.g., type, material,

  - $\{(I_{i,k}, \mathcal{A}_i^{\text{dynamic}}|_{I_{i,k}})\}$ is an ordered set of temporal states where each

$$I_{i,k} = [t_{i,k}(0), t_{i,k}(1)] \subseteq \mathcal{T}$$

is a closed time interval, and $\mathcal{A}_i^{\text{dynamic}}|_{I_{i,k}}$ are the dynamic attribute values valid for that interval (e.g., position, lid state, traffic-light color).

- $E \subseteq V \times V$ is a set of directed edges. Each edge $e = (v_i, v_j)$ is augmented by

$$e = (v_i, v_j, r_{ij}, C_{ij}), \tag{3}$$

where

  - v_i is the source node,
  - v_j is the target node,
  - r_{ij} is the relation type (e.g., “next_to”, “controls”),

  - C_{ij} is an optional constraint function mapping dynamic attributes of v_i to changes in v_j (e.g., if traffic light changes to red, car stops).

- $\mathcal{T}$ is the global time domain.

- $\tau : \bigcup_i \mathcal{S}_i \cup \bigcup_{i,j} \mathcal{I}_{ij} \to \mathcal{P}$ is an optional grounding map to represent the dynamic aspects w.r.t. time.

Semantics A dynamic scenegraph S encodes both

1. the *static layout* through static attributes $\mathcal{A}_i^{\text{static}}$ and relations r_{ij} that are continuously active (or active on $\mathcal{T}$),

2. the *dynamic behaviour* through temporal attributes $\mathcal{A}_i^{\text{dynamic}}$ and constraint relations C_{ij} between nodes.

Example

Figure 2 illustrates the example dynamic scenegraph.

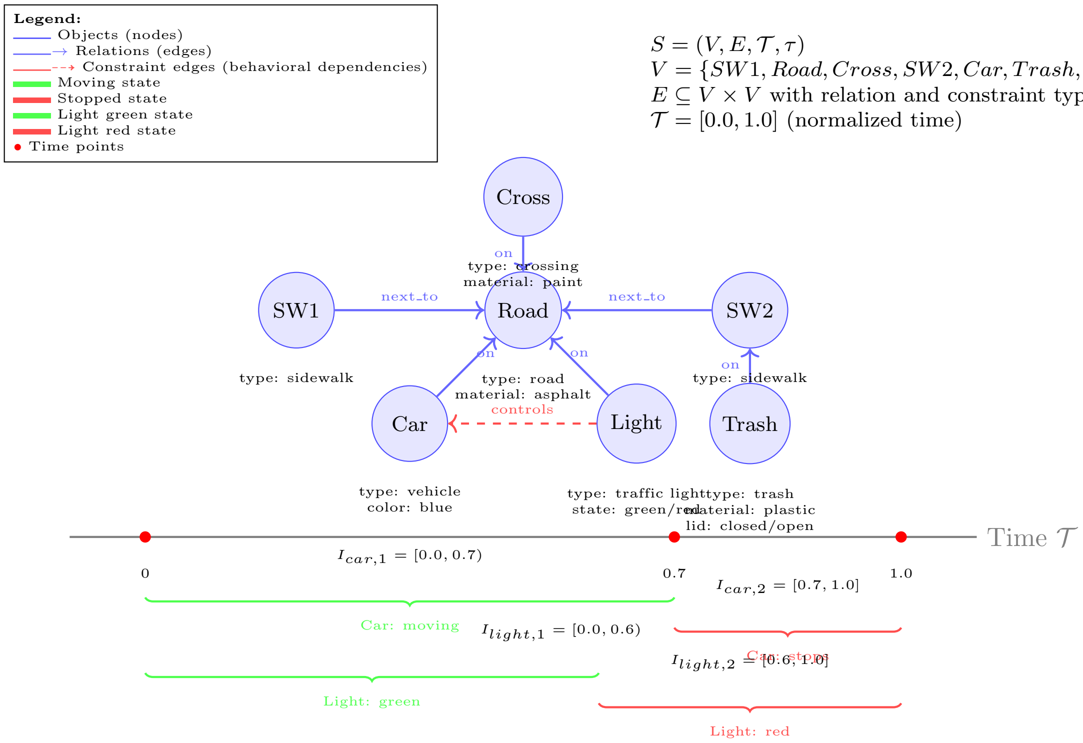


Figure 2: Example dynamic scenegraph with constraints (nodes, relations with active intervals, temporal states, and dependency edges).

References

[1] E. Alonso, A. Jelley, V. Micheli, A. Kanervisto, A. J. Storkey, T. Pearce, and F. Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

[2] 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.

[3] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024.

[4] 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.

[5] J. Bruce, M. D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[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] 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.

[8] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[9] 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.

[10] 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.

[11] 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.

[12] P. Dourish. *Where the action is: the foundations of embodied interaction.* MIT press, 2001.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] G. Gao, W. Liu, A. Chen, A. Geiger, and B. Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21295–21304, 2024.

[18] 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.

[19] 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. IEEE, 2012.

[20] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024.

[21] M. Hartmann, M. Viehweger, W. Desmet, M. Stolz, and D. Watzenig. “pedestrian in the loop”: An approach using virtual reality. In *2017 XXVI International Conference on Information, Communication and Automation Technologies (ICAT)*, pages 1–8. IEEE, 2017.

[22] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.

[23] 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.

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[25] 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.

[26] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3461–3475, 2022.

[27] 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*, pages 17397–17407, 2023.

[28] 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.

[29] L. Lian, B. Li, A. Yala, and T. Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[41] 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.

[42] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[44] 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.

[45] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[46] 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.

[47] S. Schneider and K. Bengler. Virtually the same? analysing pedestrian behaviour by means of virtual reality. *Transportation research part F: traffic psychology and behaviour*, 68:231–256, 2020.

[48] 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.

[49] U. Singer, S. Sheynin, A. Polyak, O. Ashual, I. Makarov, F. Kokkinos, N. Goyal, A. Vedaldi, D. Parikh, J. Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023.

[50] 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.

[51] T. T. M. Tran, C. Parker, and M. Tomitsch. A review of virtual reality studies on autonomous vehicle–pedestrian interaction. *IEEE Transactions on Human-Machine Systems*, 51(6):641–652, 2021.

[52] D. Valevski, Y. Leviathan, M. Arar, and S. Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

[53] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

[54] 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.

[55] 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.

[56] J. Wilson, J. Song, Y. Fu, A. Zhang, A. Capodieci, P. Jayakumar, K. Barton, and M. Ghaffari. Motionsc: Data set and network for real-time semantic mapping in dynamic environments. *IEEE Robotics and Automation Letters*, 7(3):8439–8446, 2022.

[57] H. Wu, D. H. Ashmead, H. Adams, and B. Bodenheimer. Using virtual reality to assess the street crossing behavior of pedestrians with simulated macular degeneration at a roundabout. *Frontiers in ICT*, 5:27, 2018.

[58] 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.

[59] 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.

[60] 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.

[61] H. Xie, Z. Chen, F. Hong, and Z. Liu. Citydreamer: Compositional generative model of unbounded 3d cities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9666–9675, 2024.

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

[63] 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.

[64] 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.

[65] 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.

[66] B. Zeng, L. Yang, S. Li, J. Liu, Z. Zhang, J. Tian, K. Zhu, Y. Guo, F.-Y. Wang, M. Xu, et al. Trans4d: Realistic geometry-aware transition for compositional text-to-4d synthesis. *arXiv preprint arXiv:2410.07155*, 2024.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] Y. Zhao, Z. Yan, E. Xie, L. Hong, Z. Li, and G. H. Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

[73] 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.

[74] Y. Zhou, J. Huang, X. Dai, S. Liu, L. Luo, Z. Chen, and Y. Ma. Holicity: A city-scale data platform for learning holistic 3d structures. *arXiv preprint arXiv:2008.03286*, 2020.