Scene Generation for Low Vision

May 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 road crossing scenes which can be used in Virtual Reality (VR) environments.

2 State of the art

In this section, we start by discussing existing grounding techniques used for road crossing (also referred as traffic scene generation in the context of this project). Next we review the different types of outputs generated by these models as well as their suitability to be used in VR. Lastly, we review the different types of generation paradigms.

2.1 Grounding: Input for Scene Generation

Generative models have been applied to traffic scene 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 scene 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 scene, 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 scene isn't just randomly composed but meaningfully aligned with the intended semantics of the input. The common ways to do grounding are,

2.1.1 Rules

Rules can be used to give a prescriptive logic that defines how to generate or modify a scene.

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

2.1.2 Constraints

Constraints are conditions that must be satisfied to get a valid scene.

- scenic -

2.1.3 Prompts

descriptions

- chatdyn - 3D-SceneDreamer - chatscene - urbanworld

2.1.4 Layout

top-down map-like view

2.1.5 Images / Videos

appearance

2.1.6 Scene Graphs

ontology

LayoutGPT [**?**] introduces a training-free method for leveraging LLMs in layout-based visual planning by formatting layouts using a CSS-like structure that LLMs can easily interpret. To support in-context learning, **CLIP** [**?**] is used to retrieve the most relevant past examples by comparing the similarity between the input prompt and stored text-image pairs.

2.2 Output

Generative models can produce a wide range of outputs depending on their architecture and intended application. These outputs vary in structure and fidelity, ranging from low-level pixel images to high-level semantic representations. In the context of road crossing simulation, choosing the appropriate output type is critical, as it determines how easily the generated content can be interpreted, manipulated, or rendered in downstream systems such as virtual reality.

One common output format is the **semantic layout** or **scene graph**, which captures relationships between entities (e.g., pedestrians, vehicles, sidewalks) in a structured form. Models like **LayoutGPT** [**?**] and **SceneDiffuser** [**?**] generate such layouts before rendering scenes. These intermediate representations provide interpretable and editable control, making them especially useful in applications where spatial relationships must be explicitly preserved. Alternatively, some models generate **RGB images** or **videos** directly from prompts or constraints (e.g., **StoryDALL-E** [**?**]), offering visually rich outputs but less flexibility for interaction or simulation. A few works also explore 3D scene synthesis (e.g., **3D-FRONT** [**?**]) to create spatially accurate environments suitable for embodied AI agents or VR settings.

Ultimately, the choice of output type influences the realism, controllability, and compatibility of generated scenes. For road crossing simulations in VR, outputs that preserve semantic and spatial coherence—such as structured layouts or 3D scene graphs—are more suitable than raw pixel-based generations, which may lack the granularity required for interaction or behavior modeling.

3D scene generation have been used to generate both indoor and outdoor scenes.

2.2.1 Outdoor Scene Generation

Outdoor 3D scene generation techniques have been mostly used to model natural landscapes, urban settings, and other external environments for applications such as autonomous driving simulations.

2.3 4D Scene Generation

4D scene generation techniques are also capable to generate temporal changes (animation key frames). Existing systems can be classified into 3 categroies.

Purely Constraint based eg. scenic

Multi-layered Perceptron based eg. DNF

Hexplane based eg. DreamGaussian4D

2.4 Datasets

2.5 Indoor Scenes

3D-Front,

2.6 Outdoor Scenes

Creative3D, GTASynth etc.

2.7 Validation

2.8 Limitations and Open Questions

3 Work Plan

Appendix

4 Meeting notes