3D Scene Generation using LLM Agents

Kunal Gupta, UC San Diego

Figure 1: We propose leveraging LLM agents to generate 3D scenes from natural language descriptions. These agents plan layouts, optimize design constraints, adjust lighting, and apply materials to create high-fidelity scenes. The framework supports applications like AR, interior design, and autonomous driving, while enabling image and video generation through physically based rendering. Scenes marked with * are preliminary results from our framework with prompts "gym" and "hair salon".

Overview Generating 3D scenes is essential for applications in virtual reality, interior design, and autonomous driving, requiring expertise in layouts, materials, lighting, and domain-specific principles. Traditionally, this process has depended on labor-intensive artistry or automated methods reliant on curated datasets, which are costly to scale and adapt to new features or domains. *Our approach overcomes these limitations by leveraging pretrained Large Language Models (LLMs), such as GPT-4, to create collaborative multi-agent workflows that generate realistic 3D scenes from simple text or image inputs.* This innovation democratizes 3D scene generation, making it accessible to nonexperts, enabling diverse layouts, supporting physically accurate image and video rendering, and providing a scalable solution for dataset generation to train embodied AI systems.

Innovations Our proposal introduces four key innovations to advance 3D scene generation. **First**, we propose a system that prompts LLMs to transform natural language, optionally enhanced with visual inputs, into a scene program— a Python-like script that provides detailed, step-by-step instructions for asset placement and manipulation. Unlike previous methods, our programmatic approach utilizes control structures to manage complex and structured scene elements effectively while integrating specialized tools to improve scene quality. Additionally, it supports iterative refinement, enabling verification and continuous improvement by multiple agents. **Second**, we introduce a Scene Description Language (SDL), a custom-designed language for representing structured scene elements. Unlike natural language used in contemporary methods, SDL leverages control structures to streamline repetitive and conditional asset manipulations. It also incorporates domain-specific constraints, such as ergonomics, along with versatile asset placement and manipulation routines, ensuring that LLM-generated scenes achieve both high visual realism and functional utility. **Third**, drawing inspiration from the success of tool use in enhancing LLM agent capabilities, we curate and develop a comprehensive library of specialized tools tailored for scene generation. These tools support intricate operations such as arranging assets in grids, embedding items within others (e.g., books in a bookshelf), and optimizing for factors like visibility, shadows, or material properties. The library is also easily extensible, allowing seamless integration of additional functionalities from publicly available repositories to address a broader range of use cases. **Lastly**, we propose a multi-agent collaboration framework to improve scene generation efficiency and quality. Specialized agents handle subtasks like layout optimization, aesthetic enhancement, and constraint validation, iteratively refining and resolving conflicts for high-fidelity results. By sharing context and enabling iterative improvements, this framework ensures scalability, adaptability, and robust performance across diverse applications.

Applications *Our research aims to develop efficient and versatile 3D scene generation frameworks powered by LLM agents, targeting applications in augmented reality (AR), interior design, autonomous driving, and image/video generation—key areas.* Leveraging LLMs trained on internet-scale data, the framework can generate diverse scenes across domains, from indoor layouts to complex city traffic simulations. Its integration with tools like Blender can significantly enhance artists' workflows, while its customizability supports fine-grained tasks like warehouse design, accommodating equipment-specific constraints. We believe this research will advance state-of-the-art 3D scene generation, democratize complex scene creation, and establish a robust foundation for high-quality image and video generation pipelines, driving innovation across industries.

Expertise I have extensive experience in graphics and vision, with a strong track record in scene generation, manipulation, and rendering. My work has been recognized with **spotlight** presentations at flagship venues such as NeurIPS [2], SIGGRAPH [1, 4] and Medical Physics Journal [3]. Additionally, I have been a recipient of Qualcomm Innovation Fellowship award, reflecting the impact of my contributions. With deep theoretical and experimental knowledge, as well as hands-on experience in 3D scene generation, deep learning, generative modeling, and LLMs, I am well-equipped to drive innovation in this domain.

[1] Noam Aigerman, Kunal Gupta, Vladimir G. Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. Neural jacobian fields: Learning intrinsic mappings of arbitrary meshes. *ACM Trans. Graph.*, 41(4), jul 2022.

[2] Kunal Gupta and Manmohan Chandraker. Neural mesh flow: 3D manifold mesh generation via diffeomorphic flows. *NeurIPS*, 2020.

[3] Kunal Gupta, Brendan Colvert, Zhennong Chen, and Francisco Contijoch. Difir-ct: Distance field representation to resolve motion artifacts in computed tomography. *Medical physics*, 50(3):1349–1366, 2023.

[4] Kunal Gupta, Milos Hasan, Zexiang Xu, Fujun Luan, Kalyan Sunkavalli, Xin Sun, Manmohan Chandraker, and Sai Bi. Mcnerf: Monte carlo rendering and denoising for real-time nerfs. In *SIGGRAPH Asia 2023 Conference Papers*, SA '23, New York, NY, USA, 2023. Association for Computing Machinery.