Copiloting Creative 3D Scene Modeling and Visualization with Generative Agents

Yiqin Zhao[1*] **Yu Shen**[2] **Stefano Petrangeli**[2] **Matheus Gadelha**[2]
Cuong Nguyen[2] **Gang Wu**[2]
[1]Worcester Polytechnic Institute [2]Adobe Research
yzhao11@wpi.edu
{shenyu,petrange,gadelha,cunguyen,gawu}@adobe.com

Abstract

Creating 3D scenes is a critical yet challenging digital content creation task due to the high demand for professional skills and intense labor efforts. Recently, significant progress has been made in accelerating 3D scene creation through AI-assisted automated design and 3D content generation. However, current systems can not effectively support 3D creation workflow and inspire idea explorations. In this work, we present a novel 3D scene-creating copilot system that allows high-quality and iterative 3D scene modeling and visualization. Under the hood, our system leverages a novel multi-step reasoning workflow to control language- and image-generative agents to create 3D scenes. Compared to traditional 3D modeling and visualization workflows, our system provides more natural and intuitive control and can significantly reduce the time required to create 3D content.

1 Introduction

Creating 3D scenes is a critical task in digital content creation with applications including interior designs, game design, visual effects, product marketing, etc. Recent advancements in generative AI have brought potential improvements to 3D scene creators, as significant progress has been made in AI-assisted automated design and 3D content generation [1, 2, 3]. New developments in AI-assisted 3D content generation include text-to-3D mesh generation [4], procedural model generation [5], and room layout generation [6]. Although current systems can assist 3D content creators in labor-intensive tasks, few research studies have been conducted to improve 3D creation workflow and inspire idea explorations [7]. In this work, we present a novel copilot system that leverages language and image-generative agents to allow high-quality and iterative 3D scene modeling and visualization and accelerate the creation of creative 3D content. Compared to traditional 3D modeling and visualization workflows, our system provides more natural and intuitive control and can significantly reduce the time required to create 3D content.

Our copilot system is designed to tackle the following key challenges in 3D modeling and visualization: *(i)* Traditional 3D editing software typically has *complex user interfaces* that introduces a steep learning curve for content creators. *(ii)* 3D scene modeling and visualization are time-consuming processes that typically require labor-intensive *manual adjustments and fine-tuning* of 3D objects. *(iii)* Generating high-quality, especially large-scale, 3D scenes can be computationally intensive and time-consuming, making iterative evaluating and refining of 3D scene design difficult.

We summarize our key technical contributions as follows: *(i)* we design and implement an easy-to-use copilot system for 3D scene modeling and visualization by leveraging text and image generative agents. Our copilot system allows content creators to control the generation from simple and intuitive

[*]Work partially done while the author interned at Adobe Research.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

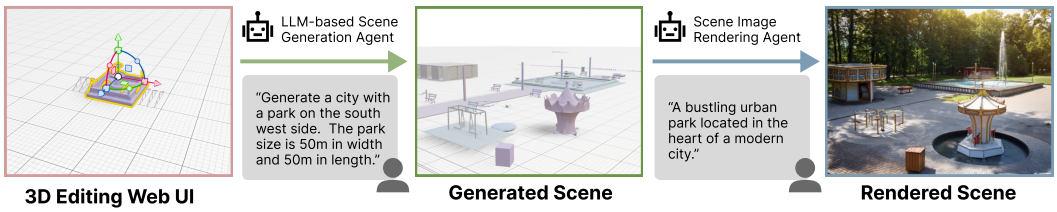

Figure 1: Creation workflow overview. Our system utilizes a web-based interface for 3D scene and object editing. We design an LLM-based scene generation agent to allow easy creation of complicated scenes. The generated 3D scene can then be rendered with a diffusion model-based rendering agent.

inputs, such as natural language text and graphical UI. *(ii)* we develop an LLM-based scene generation agent that leverages a multi-step reasoning workflow to allow high-quality 3D scene creation without labor-intensive manual adjustments or fine-tuning. And *(iii)* we develop a scene subdivision algorithm to allow our scene generation agent to create open-set and large-scale 3D scenes.

2 Related Works

Synthesizing 3D scenes has been a long-standing task due to its practical impacts across various applications, including interior design, game design, etc. In particular, synthesizing indoor room scenes is a center focus that has received a wide range of attention due to its wide impacts [8]. Early works leverage autoregressive models to generate basic scene information such as bounding boxes [9]. Under the autoregressive learning framework, later works [10, 6] improve the generation quality with transformer-based networks. While these works allow high-quality indoor scene generation, the diversity of generation outputs is often heavily limited by the diversity of training data. This limitation discourages autoregressive models from being applied to real-world room design applications, as such applications typically have complicated content generation needs. Recently, new works [11, 12, 13] further advanced generation quality and diversity using large generative models. In particular, large language models (LLMs) have started to show promise for 3D scene generation by leveraging their understanding of semantic relationships and contextual cues derived from textual descriptions [14]. Such abilities allow LLMs to generate both high-quality and diverse details in the generation outputs, particularly open-set scenes. In this work, we investigate leveraging LLMs as the key technique to develop a scene generation system that supports creative 3D scene modeling and visualization.

3 System Design

System overview. Figure 1 shows an overview of the creation workflow of our system. Our system consists of three components: *(i)* a web-based user interface, *(ii)* an LLM-based scene generation agent, and *(iii)* a diffusion model-based rendering agent. Our user interface offers user-friendly 3D scene editing operations, such as asset importing, object manipulation, and layout adjustment. Our LLM-based scene generation agent, the main component in our system, can take natural language instructions from users and generate high-quality 3D scenes. Notably, our scene generation agent can create open-set and large-scale 3D scenes that do not require domain-specific data at any stage. To render a 3D scene, users can trigger the diffusion model-based rendering agent from the current camera view and add any text descriptions as additional guidance to the diffusion model. Unlike traditional rendering workflows, which require highly detailed 3D assets to produce high-quality results, our system leverages a diffusion model-based rendering agent from the Adobe Firefly API [15] to produce high-quality rendering from simple scene setups with basic object geometries.

Scene generation LLM agent. Recent advancements have allowed LLMs to understand complex natural language inputs and generate structured outputs that follow rigorous rules [16]. Leveraging this key capability of LLM, recent works introduce several systems to allow 3D scenes, represented as a scene graph text format, generated from the user's natural language descriptions [11, 5, 13]. However, existing system designs often face quality and generalization issues in real-world applications as they rely heavily on domain-specific data for fine-tuning or in-context examples to ensure high-quality outputs. To address this issue, we design a *multi-step reasoning workflow* that utilizes LLM's built-in knowledge to create high-quality open-set 3D scenes. Specifically, our multi-step reasoning

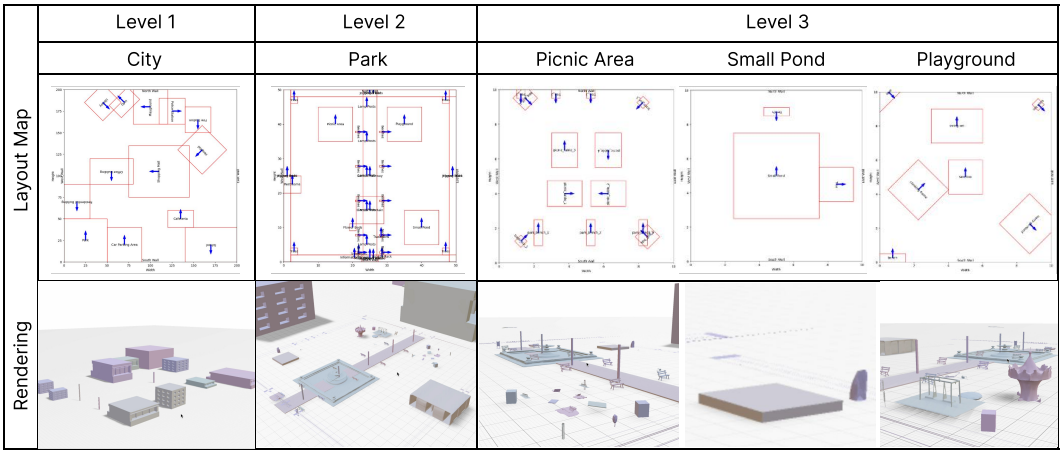

Figure 2: Visualization of the subdivision process in scene generation. Our scene generation agent utilizes a subdivision process to generate high-quality and large-scale scenes. In this example, the entire scene is generated from a single user prompt *"Generate a city with a park on the southwest side. The park size is 50m in width and 50m in length." Our scene generates agent first generates a layout for the city and then subdivides the park into a sub-area to generate. Finally, individual function areas in the park are generated by continuing the subdivision. The subdivision process is determined by each area's size.*

workflow explicitly instructs LLM to plan the scene object positions, orientations, and relative spatial relationships. The generation process can also be configured with multiple conditions, such as object types, number of objects, scene sizes, etc. To scale up the generation to large-sized scenes with many objects, we introduce a scene subdivision algorithm that helps the generation agent break down large generation goals into smaller ones. This design helps the agent to maintain the reasoning complexity and achieve overall high-generation quality. Specifically, the scene subdivision algorithm first generates high-level planning for a scene, such as a city, and then divides the scene into sub-regions, e.g., a park, based on the region sizes. During each subdivision, the algorithm will create a context-inheriting prompt to combine the previous generation's goal with the current one Figure 2 depicts an example scene subdivision process.

Using the multi-step reasoning workflow and the scene subdivision algorithm, our scene generation agent will output a scene graph. This scene graph includes object category information and detailed spatial information such as sizing and positions. Our generation agent can also create object orientation angles, which current scene generation agents often struggle to do. In the subsequent step, 3D model assets will be retrieved and composited into the scene in the web interface. We implement the retrieval in two steps: *(i)* CLIP score [17]-based asset matching between asset name and its rendered image and *(ii)* Object morphology front identification via visual prompting. In the second step, our system will render a set of four object renderings and prompt multi-modal LLM to identify its morphology front so that it correctly aligns with the planned object orientation in the original scene graph. In summary, our scene-generation agent can generate high-quality and open-set scenes ranging from smaller ones, such as living rooms, to large ones, such as a city.

4 Evaluation

We first show example renderings of a living room and a bedroom in Figure 3. Our system can generate indoor rooms with visually coherent layouts, including detailed object position and orientation. Note that our scene-generation agent generates scene objects with respect to their real-world physical sizes. The agent also ensures that objects are placed in contextually appropriate locations, such as arranging furniture according to common room design principles and maintaining functional flow.

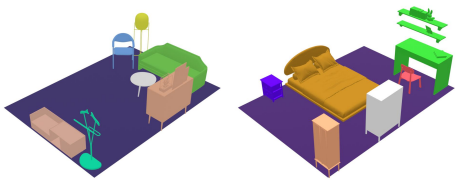

Figure 3: Visualization of indoor rooms generated by our system.

Table 1: Layout generation quality evaluation.

Models	Bedroom (FID ↓)	Living Room (FID ↓)
ATISS	30.02	85.40
LayoutGPT (GPT-3.5)	28.37	76.34
LayoutGPT (GPT-3.5, chat)	29.66	89.40
LayoutGPT (GPT-4)	29.88	78.60
Ours (GPT-4)	**27.51**	**76.21**
Ours (GPT-4o)	**27.92**	**75.11**

Figure 4: Examples of generated 3D scenes and their renderings. We show three additional 3D scene generation results (row two) with their layout maps (left) and scene object geometries (right). In row 1, we show the scene rendering results created by our diffusion-based rendering models. These results highlight that our system can create open-set 3D scenes of high quality and inspire creative 3D editing and idea exploration.

For quantitative evaluation, we evaluate our LLM scene-generation agent on the indoor room generation task against LayoutGPT [11], a recent LLM-based scene-generation system, and ATISS [6], an auto-regressive scene-generation model. We evaluate our system on the 3D-FRONT dataset [18] following LayoutGPT and ATISS. We measure the generation quality regarding the FID scores [19] on the rendered images of the generated scenes. In Table 1, we show the quantitative evaluation results. Our system achieves FID scores of 27.51 and 76.21 on generated bedrooms and living rooms, representing a 5.8% and 9.4% improvement over LayoutGPT and ATISS, respectively.

Additionally, we showcase several examples of our system-generated open-set scenes and their renderings in Figure 4. In the figure, we show that our system can generative tailored scene layouts with sophisticated object layout planning. The generated object layouts can then be used to generate photorealistic rendering with our diffusion-based rendering models. Compared to traditional creative workflow, our system not only requires much simpler user inputs but also delivers high-quality results much faster. These examples highlight our system's capability to generate high-quality, open-set 3D scenes that cater to diverse aesthetic preferences and spatial arrangements. This makes it a valuable tool for creative 3D editing and facilitating the exploration of new design ideas.

Acknowledgments

We thank the reviewers for their thoughtful feedback and valuable suggestions. We also thank our collaborators at Adobe Research for their constructive contributions throughout this project.

References

[1] Ziv Epstein et al. "Art and the science of generative AI". In: *Science* 380.6650 (2023), pp. 1110–1111.

[2] Ramya Srinivasan and Kanji Uchino. "Biases in generative art: A causal look from the lens of art history". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 41–51.

[3] Eric Zhou and Dokyun Lee. "Generative artificial intelligence, human creativity, and art". In: *PNAS nexus* 3.3 (2024), pgae052.

[4] Chen-Hsuan Lin et al. "Magic3d: High-resolution text-to-3d content creation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 300–309.

[5] Mengqi Zhou et al. "Scenex: Procedural controllable large-scale scene generation via large-language models". In: *arXiv preprint arXiv:2403.15698* (2024).

[6] Despoina Paschalidou et al. "Atiss: Autoregressive transformers for indoor scene synthesis". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12013–12026.

[7] Hye-Kyung Lee. "Rethinking creativity: creative industries, AI and everyday creativity". In: *Media, Culture & Society* 44.3 (2022), pp. 601–612.

[8] Song-Hai Zhang et al. "A survey of 3D indoor scene synthesis". In: *Journal of Computer Science and Technology* 34 (2019), pp. 594–608.

[9] Daniel Ritchie, Kai Wang, and Yu-an Lin. "Fast and flexible indoor scene synthesis via deep convolutional generative models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6182–6190.

[10] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. "Sceneformer: Indoor scene generation with transformers". In: *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, pp. 106–115.

[11] Weixi Feng et al. "Layoutgpt: Compositional visual planning and generation with large language models". In: *Advances in Neural Information Processing Systems* 36 (2024).

[12] Ata Çelen et al. "I-design: Personalized llm interior designer". In: *arXiv preprint arXiv:2404.02838* (2024).

[13] Vikram Kumaran et al. "Scenecraft: Automating interactive narrative scene generation in digital games with large language models". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 19. 1. 2023, pp. 86–96.

[14] Yining Hong et al. "3d-llm: Injecting the 3d world into large language models". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 20482–20494.

[15] Adobe System. *Adobe Firefly API - Firefly Services*. `https://developer.adobe.com/firefly-services/docs/firefly-api`. Sept. 2024.

[16] Zhiheng Xi et al. "The rise and potential of large language model based agents: A survey". In: *arXiv preprint arXiv:2309.07864* (2023).

[17] Jack Hessel et al. "Clipscore: A reference-free evaluation metric for image captioning". In: *arXiv preprint arXiv:2104.08718* (2021).

[18] Huan Fu et al. "3d-future: 3d furniture shape with texture". In: *International Journal of Computer Vision* 129 (2021), pp. 3313–3337.

[19] Maximilian Seitzer. *pytorch-fid: FID Score for PyTorch*. `https://github.com/mseitzer/pytorch-fid`. Version 0.3.0. Aug. 2020.