Interactive 3D Room Generation for Virtual Reality via Compositional Programming

Jihyun Kim[1]*, Junho Park[1]*, Kyeongbo Kong[2]*, and Suk-Ju Kang[1]✉

[1] Department of Electronic Engineering, Sogang University, South Korea
[2] Department of Electrical & Electronics Engineering, Pusan National University, South Korea
{junho18.park, wgus5950}@gmail.com
kbkong@pusan.ac.kr
sjkang@sogang.ac.kr

Abstract. We introduce a novel framework, Interactive Room Programmer (IRP), which allows users to conveniently create and modify 3D indoor scenes using natural language. Distinctively, our framework decomposes the challenging task into simpler steps instead of directly handling it. Such approach enables precise control of each attribute of an indoor scene, which was previously unattainable. To support the various decomposed tasks with a unified framework, we employ a large language model (LLM) to comprehend an instruction, select and organize the modules in response. More specifically, inspired by visual programming, we leverage the LLM as a programmer to generate compositional programs. We demonstrate IRP's flexibility in generating and editing 3D room meshes, and prove our framework's superiority compare to an existing model quantitatively and qualitatively.

Keywords: Indoor Scene Synthesis · Text-to-3D Generation · In-Context Learning

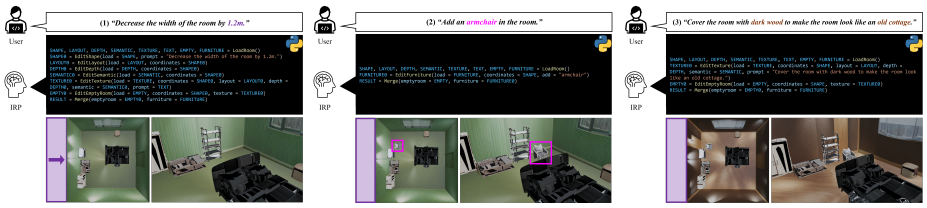

Fig. 1: Editing examples supported by IRP. Pairs of user instructions and corresponding bounding boxes are illustrated in the same color.

1 Introduction

Virtual space is a crucial element of Metaverse. However, generation of 3D assets has traditionally been the domain of graphics experts due to its complexity. In

* Equal contribution.
✉ Corresponding author.

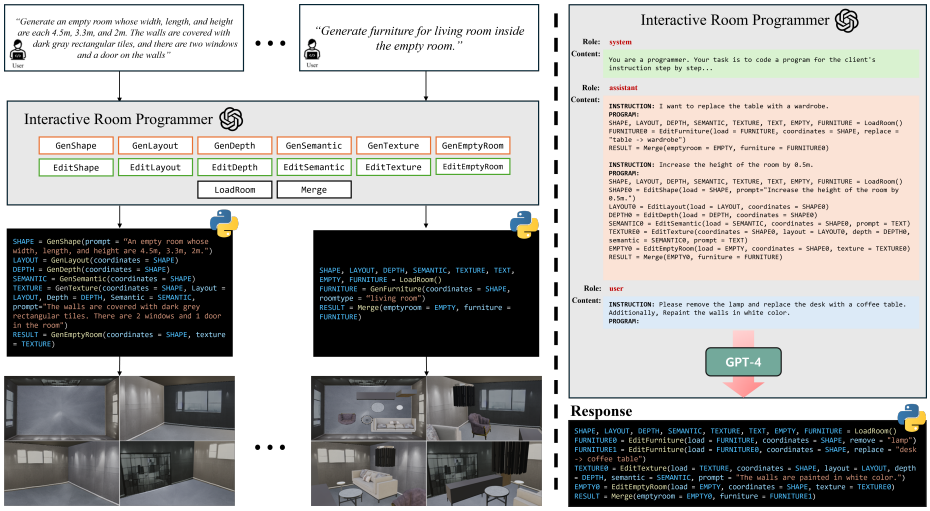

Fig. 2: (left) Overall pipeline. IRP converts a user-provided instruction into a Python program of predefined modules. Users can continuously edit the generated room until they obtain the most satisfying result. (right) Prompts given to an LLM, including supporting examples for in-context-learning are illustrated.

response, research on text-to-3D indoor scene [1] has recently got its attention for enabling non-experts to easily generate 3D room using text. Our framework, Interactive Room Programmer (IRP), takes one step further to provide precise control over the generation and editing of 3D rooms by decomposing the complex process of room building into multiple subtasks: (1) determining 3D coordinates of a room which align with user-provided instructions, (2) generating a panorama texture image, (3) constructing an empty room by integrating the coordinates and the panorama texture image, and (4) arranging appropriate furniture in the room. For example, in Fig. 1, users can change the color and texture of the floor and walls individually without affecting other elements such as the room's shape, furniture, and the location of windows and door.

To support the various decomposed tasks with a unified framework, IRP leverages a large language model (LLM) to comprehend the user instruction, select and organize predefined modules. We developed most of the algorithms and models in the module list: *GenShape, GenSemantic, GenTexture, GenEmptyRoom, GenFurniture, EditShape, EditLayout, EditDepth, EditSemantic, EditTexture, EditEmptyRoom, EditFurniture, LoadRoom,* and *Merge*. On the other hand, for subtasks on which many researches have already been conducted, we employed existing works. As for *GenFurniture*, we adopted the latest model [2] from indoor scene synthesis task, and for *GenLayout, GenDepth* we implemented coventional mathematical algorithms. Inspired by visual programming [3], we represent the organized modules in a Python format, where modules in each line are executed sequentially. In the experiments, we verify our framework can generate and edit 3D room more plausibly compare to an existing model.

2 Method

Due to an absence of training dataset, finetuning LLMs to act as a specialized program generator for our task is not feasible. Thus, we leverage in-context learning ability of LLMs. In-context learning is a way of finetuning LLMs to enhance their performance on certain tasks or domains, without updating the parameters. It is achieved by providing examples within the context of the tasks or domains. Thus, as shown in Fig. 2, we provide the LLM an instruction and a general task description, along with pre-selected supporting examples which consist of diverse pairs of instructions and their corresponding programs. To prevent programming errors and enhance the LLM's understanding, we named each module and variables intuitively.

3 Experiments

3.1 Datasets

For the furniture generation and editing modules, we utilize two datasets: 3D-FUTURE [4] and 3D-FRONT [5]. 3D-FUTURE comprises 5,000 varied scenes and involve 9,992 distinct industrial 3D CAD furniture shapes. 3D-FRONT, is a large-scale, and comprehensive repository of synthetic indoor scenes. It includes 18,797 rooms, each uniquely furnished with a variety of 3D objects.

3.2 Evaluation Metrics

We rendered images of 10 rooms from 5 different views to conduct a user study. We asked 30 participants to score perceptual quality (PQ) and 3D structure completeness (3DS) of the room meshes on scores ranging from 1 to 5. Moreover, we measured the runtime of each model in seconds.

3.3 Experimental Results

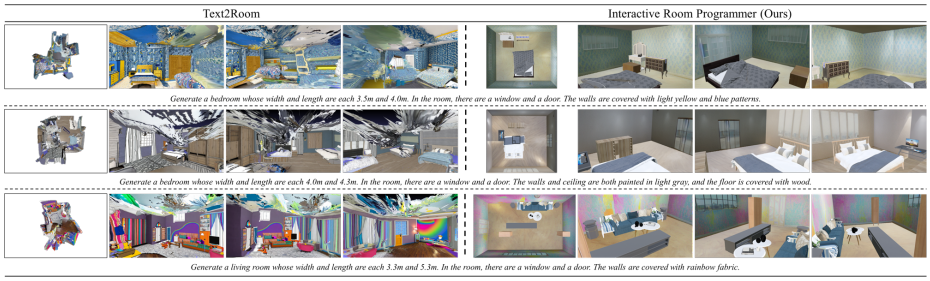

Fig. 3: Qualitative comparisons on 3D mesh generation between Text2Room [1] and IRP.

As shown in Table 1, our method achieves higher scores over Text2Room [1] as the rendered images are more visually pleasing in terms of layout. On the

Table 1: Quantitative comparisons on 3D room mesh generation.

Method	PQ↑	3DS↑	Runtime(s)↓
Text2Room [1]	2.68	2.39	5179.08
IRP (Ours)	**3.57**	**3.82**	**154.61**

contrary, rendered images from Text2Room demonstrate unrealistic room shape and layout, which resulted in a low 3DS score. Moreover, our method better satisfies given texture text prompts, thus achieves higher PQ. For example, in the second case of Fig. 3, the instruction specifies the the walls to be in light gray but the output mesh of Text2Room is in purple and blue colors. In addition, the room mesh generated by Text2Room contains floating artifacts which degrade the perceptual quality of rendered images. In terms of computing speed, our method is approximately 33 times faster than Text2Room as depicted in Table 1.

4 Conclusion

We presented a novel framework for interactive 3D room mesh generation and editing with user-provided instructions in natural language format. For a unified approach to the various subtasks, IRP leverages an LLM to write programs using predefined modules. Consequently, our framework demonstrates strength in editability. The primary limitation of IRP lies in its limited room categories. This is due to furniture generation models from indoor scene generation task which are focused on bedroom and living room. As the research on this task continues, we expect our system's output to be even more realistic and diverse. Additionally, building a diffusion model conditioned on multiple prompts for *GenTexture* – a module which generates texture panorama images for specified room layouts and descriptions, was challenging. For the future work, we will sophisticate the conditioning method.

References

1. Höllein, Lukas and Cao, Ang and Owens, Andrew and Johnson, Justin and Nießner, Matthias. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Int. Conf. Comput. Vis.*, pages 7909–7920, 2023. 2, 3, 4
2. Feng, Weixi and Zhu, Wanrong and Fu, Tsu-jui and Jampani, Varun and Akula, Arjun and He, Xuehai and Basu, Sugato and Wang, Xin Eric and Wang, William Yang. Layoutgpt: Compositional visual planning and generation with large language models. *Adv. Neural Inform. Process. Syst.*, 36, 2024. 2
3. Gupta, Tanmay and Kembhavi, Aniruddha. Visual programming: Compositional visual reasoning without training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14953–14962, 2023. 2
4. Fu, Huan and Jia, Rongfei and Gao, Lin and Gong, Mingming and Zhao, Binqiang and Maybank, Steve and Tao, Dacheng. 3d-future: 3d furniture shape with texture. *Int. J. Comput. Vis.*, 129:3313–3337, 2021. 3

5. Fu, Huan and Cai, Bowen and Gao, Lin and Zhang, Ling-Xiao and Wang, Jiaming and Li, Cao and Zeng, Qixun and Sun, Chengyue and Jia, Rongfei and Zhao, Binqiang and Zhang, Hao. 3d-front: 3d furnished rooms with layouts and semantics. In *Int. Conf. Comput. Vis.*, pages 10933–10942, 2021. 3