RealityCrafter: User-guided Editable 3D Scene Generation from a Single Image in Mixed Reality

Seokyoung Kim
Graduate School of Metaverse
KAIST
Daejeon, Republic of Korea
seokyoung@kaist.ac.kr

Dooyoung Kim
KI-ITC ARRC
KAIST
Daejeon, Republic of Korea
dooyoung.kim@kaist.ac.kr

Taejun Son
Graduate School of Metaverse
KAIST
Daejeon, Republic of Korea
signal725@kaist.ac.kr

Woontack Woo
UVR Lab.
KAIST
Daejeon, Republic of Korea
wwoo@kaist.ac.kr

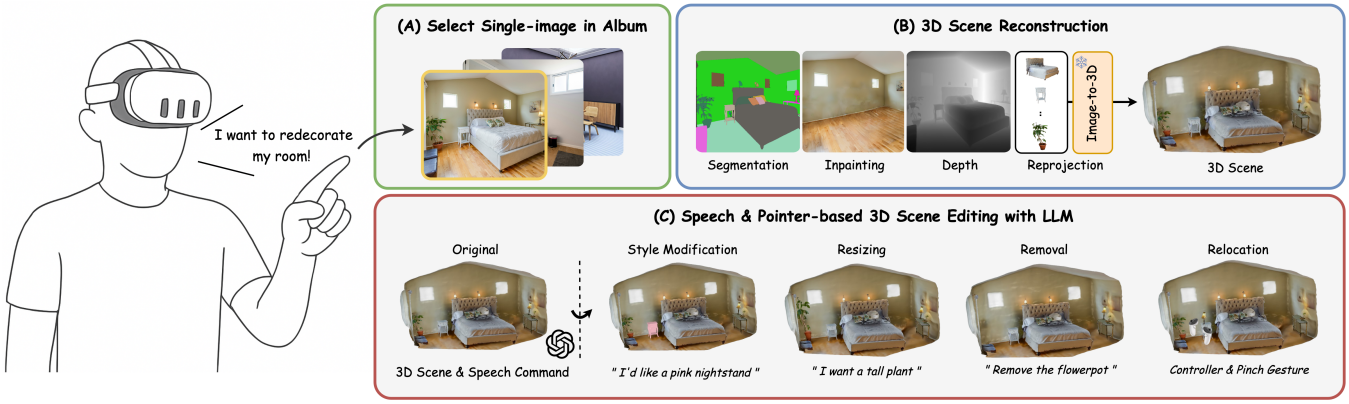

Figure 1: RealityCrafter creates an editable 3D scene driven by user speech from a single real-world image. Our system (a) takes a single image as input; (b) reconstructs it into a 3D scene; and (c) empowers users to author that space via voice commands and pointer interactions.

Abstract

We propose RealityCrafter, a mixed-reality 3D authoring tool that enables users to edit and interact with a reconstructed 3D scene from a single real-world image. Prior research has largely focused on 3D authoring tools for purely virtual spaces, insufficiently incorporating real-world context and thereby hindering user immersion during the creation process. To overcome these limitations, our approach takes a single real-world image as input, generates segmented object-level 3D meshes in a zero-shot manner, and reconstructs a 3D scene where objects can be removed or modified without occlusion through instance mask-based inpainting. We leverage LLMs to interpret user voice commands and update the style, position, scale, and orientation of 3D objects in real time, providing an interactive 3D authoring interface in mixed-reality environments. By using a single image as a baseline, this approach enables effortless generation of realistic 3D scenes and intuitive editing based on user intent, delivering a novel creative experience that seamlessly blends the real and the virtual objects.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools; 3D Authoring**; • **Computing methodologies** → **Mixed reality; Computer vision**.

Keywords

Graphics; Mixed Reality; Generative AI; AI assisted creativity tool

ACM Reference Format:
Seokyoung Kim, Dooyoung Kim, Taejun Son, and Woontack Woo. 2025. RealityCrafter: User-guided Editable 3D Scene Generation from a Single Image in Mixed Reality. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3746058.3758405

UIST Adjunct '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2036-9/25/09
https://doi.org/10.1145/3746058.3758405

1 Introduction

With the recent significance of spatial computing technologies to extend real-world spaces into mixed reality (MR), there is a growing need for authoring tools that allow users to create immersive 3D content directly based on their physical environment [2, 8]. Furthermore, as MR systems increasingly incorporate speech, text, and pointers, interfaces for multi-modal intent inference and adaptive 3D manipulation have become essential [5].

While many studies have explored 3D scene authoring in AR/VR, systems enabling real-world grounded editing and seamless virtual blending remain challenging. [14] introduced generative models into the VR authoring pipeline for AI-user collaborative layout creation, but its reliance on synthetic datasets limited its applicability to real settings. Similarly, [11] adopted a radiance field-based approach for photorealistic editing, yet its functionality was constrained to pre-constructed virtual scenes.

In this paper, we present RealityCrafter, a MR authoring system that reconstructs a 3D scene from a single real-world image and enables user-guided editing and customization. We incorporate amodal completion [12] and object mask-based inpainting [3], allowing occluded regions to be plausibly reconstructed and removed objects to blend seamlessly with the background. User speech is parsed via a large language models (LLMs), enabling intuitive control over object style, scale and position. This allows users to generate and customize an editable 3D scene from a single image of a past, current, or imagined space through a interactive interface.

2 System Implementation

RealityCrafter is an interface system designed to generate and edit 3D scenes from a single real-world image. The overall system consists of two main stages: (1) offline 3D scene reconstruction from a single image, and (2) online user-guided 3D object editing and creative authoring.

2.1 3D Reconstruction from a Single Image

Robust Image Understanding in the Wild. The core challenge of reconstructing 3D scenes from a single image lies in capturing rich features despite occluded or non-visible regions. To address this, we estimate depth [9], camera parameters [6], and semantic segmentation [13] from the input image to recover both the global layout and instance-level information of the scene. Inspired by [4], we adopt a pretrained diffusion model [12] to perform amodal completion, filling in the occluded parts of each object instance based on segmented crop image. This pipeline enables robust object-level feature extraction from real-world images captured in the wild, and generalizes to multi-object scenes in a zero-shot manner.

3D Scene and Background Generation. RealityCrafter constructs the 3D scene by handling object instances and background separately. Objects are individually reconstructed using a single image-to-3D module [7] and reprojected in the scene according to the predicted depth map. To generate the background, we apply image inpainting [3] guided by segmented object masks, ensuring visual consistency even after object removal. The inpainted background is rendered as a continuous surface using an SDF representation

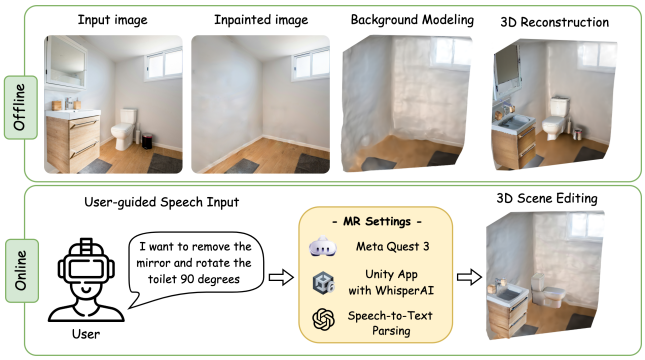

Figure 2: Our system reconstructs a 3D scene from a single image in an offline stage, then moves into an online phase where users direct object edits and craft 3D content.

derived from a depth-based point cloud. Finally, object instances and background are merged as meshes into a unified 3D scene that preserves the original camera viewpoint.

2.2 Interactive Creation of Editable 3D Scenes

LLM-driven Speech Command Parsing. The user's voice commands are transcribed into text in real time using Whisper AI [1] and are treated as interaction inputs for the reconstructed 3D scene. The transcribed text is passed to the GPT API along with a predefined prompt template to extract the command intent and key parameters (e.g., object name, manipulation type and scale). The resulting structured command is then mapped to an internal interpreter, which translates it into concrete actions such as style modification, movement, scaling, or deletion of 3D objects. To ensure reliability, we compute a confidence score by averaging the softmax probabilities of relevant GPT tokens, and request user confirmation when the score falls below a 0.8 threshold to prevent execution errors.

Key Interactions. RealityCrafter seamlessly fuses reconstructed real-world spaces with virtual objects in MR environments and enables intuitive 3D scene authoring through four key interactions:

- *Style modification*: When a speech command like "Make this chair pink" is received, the LLM extracts the style attribute and communicates it to the Python server, which updates the corresponding texture using the InTex [10].
- *Resize & Rotate*: Users can adjust the scale and rotation of real-world-based virtual objects using speech commands.
- *Object removal*: Upon receiving a command like "Delete the table", the system removes the corresponding virtual object and blends the background naturally with the surrounding.
- *Object relocation*: By dragging and dropping via a VR controller, users can move virtual objects to desired positions.

By combining voice-based control with pointer input, this multi-modal interaction effectively integrates real and virtual content in MR environments and enables users to carry out complex 3D editing tasks intuitively.

3 Potential Applications

RealityCrafter enables users to revive spaces connected to past memories and create interactive 3D content using voice and pointer-based editing in MR environments. In VR interior design, users can reconstruct their living spaces, simulate furniture placement in real time, and explore personalized layouts. It can also be extended into a remote collaboration platform, supporting real-time feedback and co-editing in a shared MR space.

4 Conclusion and Future Work

This paper describes RealityCrafter, a 3D authoring system that combines zero-shot 3D scene reconstruction from a single image with user-guided voice command parsing. The proposed approach enables intuitive and immersive 3D editing in mixed reality by immediately reflecting user intent through voice and pointer-based interactions. In future work, we plan to evaluate the system through user studies and deploy it across real-world application scenarios, with the goal of extending it into a collaborative authoring tool for multi-user MR environments.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(IITP-2022(2025)-RS-2022-00156435) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2024-00436398) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2021-NR059444).

References

[1] 2022. Whisper AI. https://github.com/openai/whisper OpenAI.

[2] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. 2024. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21476–21485.

[3] Zhekai Chen, Wen Wang, Zhen Yang, Zeqing Yuan, Hao Chen, and Chunhua Shen. 2024. FreeCompose: Generic Zero-Shot Image Composition with Diffusion Prior. In *European Conference on Computer Vision*. Springer, 70–87.

[4] Andreea Dogaru, Mert Özer, and Bernhard Egger. 2025. Gen3DSR: Generalizable 3d scene reconstruction via divide and conquer from a single view. *International Conference on 3D Vision* (2025).

[5] Chenfeng Gao, Wanli Qian, Richard Liu, Rana Hanocka, and Ken Nakagaki. 2024. Towards Multimodal Interaction with AI-Infused Shape-Changing Interfaces. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–3.

[6] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. 2023. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17307–17316.

[7] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. 2025. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747* (2025).

[8] Pinyao Liu, Alexandra Kitson, Claudia Picard-Deland, Michelle Carr, Sijia Liu, Ray Lc, and Chen Zhu-Tian. 2024. Virtual Dream Reliving: Exploring Generative AI in Immersive Environment for Dream Re-experiencing. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.

[9] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. 2025. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110* (2025).

[10] Jiaxiang Tang, Ruijie Lu, Xiaokang Chen, Xiang Wen, Gang Zeng, and Ziwei Liu. 2024. Intex: Interactive text-to-texture synthesis via unified depth-aware inpainting. *arXiv preprint arXiv:2403.11878* (2024).

[11] Cyrus Vachha, Yixiao Kang, Zach Dive, Ashwat Chidambaram, Anik Gupta, Eunice Jun, and Björn Hartmann. 2025. Dreamcrafter: Immersive Editing of 3D Radiance Fields Through Flexible, Generative Inputs and Outputs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–13.

[12] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. 2024. Amodal completion via progressive mixed context diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9099–9109.

[13] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. 2024. Open-vocabulary SAM: Segment and recognize twenty-thousand classes interactively. In *European Conference on Computer Vision*. Springer, 419–437.

[14] Lei Zhang, Jin Pan, Jacob Gettig, Steve Oney, and Anhong Guo. 2024. Vrcopilot: Authoring 3d layouts with generative ai models in vr. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–13.