An implementation of MagicCraft: Generating Interactive 3D Objects and Their Behaviors from Text for Commercial Metaverse Platforms

Ryutaro Kurai*
Cluster, Inc.
Nara Institute of Science and Technology

Takefumi Hiraki†
Cluster Metaverse Lab
University of Tsukuba

Yuichi Hiroi‡
Cluster Metaverse Lab

Yutaro Hirao§
Nara Institute of Science and Technology

Monica Perusquia-Hernandez¶
Nara Institute of Science and Technology

Hideaki Uchiyama‖
Nara Institute of Science and Technology

Kiyoshi Kiyokawa**
Nara Institute of Science and Technology

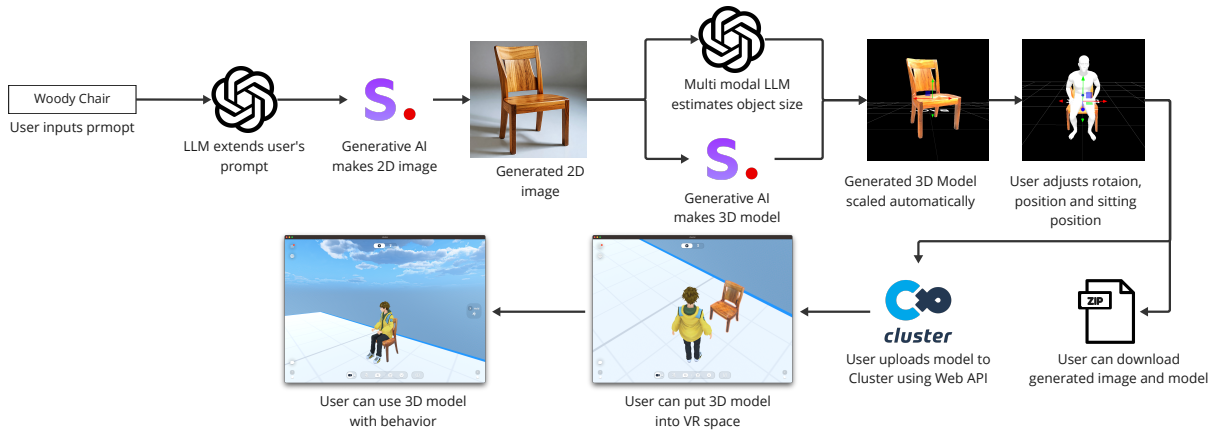


Figure 1: The initial user input is augmented into prompts more suitable for image generation by a large language model (LLM), and this augmented prompt is used as input for an image-generation LLM to generate an image. The resulting image is used as input for an 2D-to-3D to generate a 3D model, and also helps to estimate the real-world size of the object. The 3D model is scaled based on the estimated size, and users can adjust the object's behavior in the 3D graphical user interface and gizmos. The finished object can be immediately uploaded to metaverse services, where users can place or interact with it in virtual space. Users can also download the object as a local file.

Abstract

Metaverse platforms are rapidly evolving to provide immersive spaces. However, the generation of dynamic and interactive 3D objects remains a challenge due to the need for advanced 3D modeling and programming skills. We present MagicCraft, a system that generates functional 3D objects from natural language prompts. MagicCraft uses generative AI models to manage the entire content creation pipeline: converting user text descriptions into images, transforming images into 3D models, predicting object behavior, and assigning necessary attributes and scripts. It also provides an interactive interface for users to refine generated objects by adjusting features like orientation, scale, seating positions, and grip points.

Keywords: Metaverse, 3D Object Generation, Generative AI, AI-Assisted Design

1 Introduction

The rapid development of Virtual Reality (VR) technology is leading to the emergence of metaverse platforms that provide immersive environments for users to interact, collaborate, and express creativity through avatars and virtual content.

User-generated content (UGC), particularly the creation and manipulation of 3D objects in virtual environments, is central to the metaverse experience. However, the development of dynamic and interactive 3D objects is a significant challenge for most users due to the steep learning curve associated with 3D modeling and programming skills. This barrier limits non-experts from engaging in creative activities in the metaverse, reducing the potential diversity and richness of UGC that is essential to the growth of these platforms.

Recent advances in large language models (LLMs) and generative AI show promise in reducing obstacles to content creation. Generative AI models like Stable Diffusion [6], which generate 2D images from natural language, and DreamFusion [5], which produce 3D objects from text or images, have been developed. Moreover, multimodal LLMs such as OpenAI GPT-4 [4] have demonstrated the ability to recognize media like images and to understand and generate code, including scripts that define object behaviors [2].

However, integrating these generative models into metaverse platforms to generate 3D objects that interact with the environment based on users' natural language input is not as straightforward as it may seem. First, ensuring that the generated 3D objects are compatible with the platform's specifications so that they can be smoothly integrated is challenging due to differences in file formats, rendering pipelines, and interaction mechanisms. Second, translating natural language descriptions into functional 3D models with precise behaviors requires coordination among multiple AI models, each with its own limitations, such as image generation fidelity and 3D reconstruction accuracy. Third, defining interactions between the generated 3D objects and users in the metaverse involves ambiguity and diversity, as there is no unique mapping between objects and interactions.

To address these challenges, we propose MagicCraft, a system that enables users to generate functional and dynamic 3D objects

*e-mail: r.kurai@cluster.mu
†e-mail: t.hiraki@cluster.mu
‡e-mail: y.hiroi@cluster.mu
§e-mail: yutaro.hirao@is.naist.jp
¶e-mail: m.perusquia@is.naist.jp
‖e-mail: hideaki.uchiyama@is.naist.jp
**e-mail: kiyo@is.naist.jp

for metaverse platforms from natural language prompts. MagicCraft uses generative AI models to execute the entire content creation pipeline: generating images from user input text descriptions, converting images into 3D models, predicting object behavior, and assigning appropriate attributes and scripts. It also provides an interactive user interface for refining the generated objects, allowing users to adjust interaction points such as orientation, scale, seating positions, and grip points.

2 Implementation

2.1 System Overview

Figure 1 shows an overview of MagicCraft. First, the user imagines the object he wants to create and enters it as a prompt. The system then generates a 2D image based on this prompt. The user can review this image and iterate through the image generation process until the desired object is accurately represented. Once satisfied with the 2D representation, the user can proceed to generate a 3D object from this image. Following the system's guidance, the user can adjust the orientation, scale, and behavior of the generated object. Finally, the finished 3D object can be uploaded directly to a metaverse platform, allowing users to immediately observe and interact with the object in the metaverse environment. This streamlined process enables rapid prototyping and deployment in virtual spaces. In this implementation, we chose Cluster [1] as the metaverse environment. We then implemented a system that creates objects that run on Cluster from users' natural language input.

2.2 Image Generation

Although users tend to input short instructions of 2–3 words, image generation is known to use more detailed instructions, resulting in more complex and physically accurate images of the shape. Therefore, prior to image generation, user input prompts are first sent to the LLM to provide improved prompts tuned for image generation, such as describing the image in more detail.

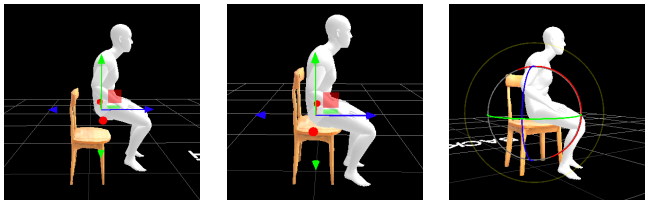
2.3 Object Generation and Scale Adjustment

The generated image is then converted into a 3D object using an image-to-3D generative AI model Stable Fast 3D from stability.ai. This generative AI model generates 3D objects as standardized to a cube of about 1 m^3, regardless of their actual size in the input image. To solve this problem, we introduce an automatic scaling mechanism in our system. We estimate real size of the object from generated image by using OpenAI GPT-4o. Then our system calculates the ratio between the estimated size and the generated object.

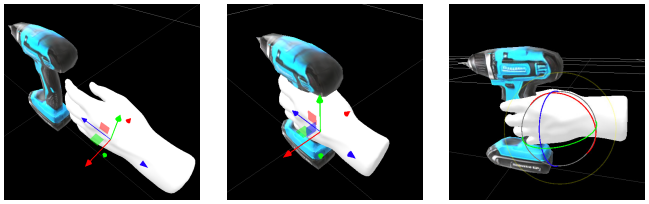
2.4 System-Assisted Object Adjustment

MagicCraft provides functionality to enable behavioral settings necessary for avatar interaction, thereby enhancing the functionality of objects in metaverse spaces. In particular, it is essential to configure settings that allow avatars to perform basic actions such as sitting on or grasping objects. On the other hand, there are also user preferences in setting up such behaviors. For example, different users may have different preferences for sitting astride or sideways against the horse's back.

To address this problem, our system presents a reasonable initial position while providing a user interface that allows the user to adjust the sitting and grasping positions. (Fig. 2). Specifically, by displaying a 170 cm height mannequin object on the 3D interface and adjusting the relative positions of this mannequin and the generated object, the user determines the avatar's interaction with the objects in the metaverse space. These behavioral settings allow the MagicCraft system to use user-created 3D objects more naturally and purposefully in metaverse space.



Sitting Position Adjustment



Grasping Position Adjustment

Figure 2: Position Adjustment User Interfaces

2.5 Automated Scripting for Generated Objects

Our the MagicCraft system automatically generates scripts from initial object images, which is possible to generate object movements and rotations that are appropriate for the object without any special knowledge. The script generation prompt is the same as in [3]: the definition of Cluster Script, the proprietary language of the platform, has been added to LLM.

2.6 Assembly and Upload Process

Our system generates several types of data from user input. These include 2D images, 3D models (consisting of meshes and textures), three vectors defining object position, orientation, and scale, two vectors indicating sitting or grasping positions and orientations, and behavior-defining scripts. We refer to the process of consolidating these heterogeneous data types into a single file called as assembly. This assembly is formatted according to the format of each metaverse platform and uploaded directly to the metaverse platform via web API. The user can then view this object, including its behavior, on the platform.

3 Conclusion

We introduced MagicCraft, a system that allows users to generate functional and dynamic 3D objects for metaverse platforms from natural language prompts. By integrating LLM with image generation, image-to-3D conversion, and automated behavior scripting, MagicCraft enables users with no prior experience in 3D modeling or programming to create complex, interactive objects in virtual environments. Future research will include automating the adjustment of interaction points and refining the generated models and behaviors.

References

[1] Cluster, Inc. Cluster. https://cluster.mu/, 2024. Accessed: 2024-6-15. 2

[2] D. Giunchi, N. Numan, E. Gatti, and A. Steed. DreamCodeVR: Towards democratizing behavior design in virtual reality with Speech-Driven programming. In *Proc. of IEEE VR*, pages 579–589, 2024. 1

[3] R. Kurai, T. Hiraki, Y. Hiroi, Y. Hirao, M. Perusquia-Hernandez, H. Uchiyama, and K. Kiyokawa. Magicitem: Dynamic behavior design of virtual objects with large language models in a consumer metaverse platform. *arXiv*, abs/2406.13242, 2024. 2

[4] OpenAI. Gpt-4 technical report. *arXiv*, abs/2303.08774, 2023. 1

[5] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proc. of ICLR*, 2023. 1

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of CVPR*, pages 10674–10685, jun 2022. 1