InteraCraft: In-HMD Authoring of Interactive 3D Objects on a Commercial Metaverse Platform

Akira Watanabe
University of Tsukuba
Ibaraki, Japan
Cluster Metaverse Lab
Tokyo, Japan
watanabe@mvml.slis.tsukuba.ac.jp

Ryutaro Kurai
University of Tsukuba
Ibaraki, Japan
kurai@mvml.slis.tsukuba.ac.jp

Yuichi Hiroi
Cluster Metaverse Lab
Tokyo, Japan
y.hiroi@cluster.mu

Takefumi Hiraki
University of Tsukuba
Ibaraki, Japan
Cluster Metaverse Lab
Tokyo, Japan
hiraki@slis.tsukuba.ac.jp

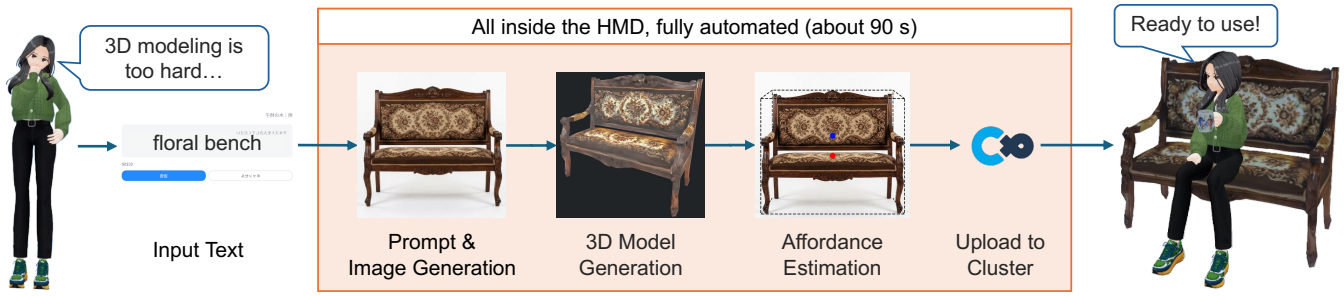


Figure 1: The end-to-end workflow of InteraCraft. From a text prompt entered in VR, the system generates a 3D object, assigns affordance coordinates, and deploys it to Cluster, entirely within the VR-HMD.

Abstract

Creating interactive 3D objects for metaverse platforms requires expertise in both 3D modeling and programming. While text-to-3D services make shape generation accessible, platforms also require affordance coordinates, such as seating positions and grasp points, which still demand manual specification. We present InteraCraft, an end-to-end system for authoring interactive 3D objects entirely inside a VR head-mounted display (HMD): from a single text input, it generates a 3D object, automatically assigns affordance coordinates, and registers it as an interactive item on the commercial metaverse platform Cluster. The system repurposes a general-purpose vision-language model (VLM) as a lightweight, training-free affordance estimator and delivers an object an avatar can sit on or grasp in about 90 seconds from text input.

CCS Concepts

• **Human-centered computing** → **Virtual reality**; *Interactive systems and tools.*

Keywords

Metaverse, Authoring Tools, Affordance, Vision-Language Models, User-Generated Content

ACM Reference Format:
Akira Watanabe, Ryutaro Kurai, Yuichi Hiroi, and Takefumi Hiraki. 2026. InteraCraft: In-HMD Authoring of Interactive 3D Objects on a Commercial Metaverse Platform. In *32nd ACM Symposium on Virtual Reality Software and Technology (VRST '26), November 16–18, 2026, Sendai, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3822517.3849936

This work is licensed under a Creative Commons Attribution 4.0 International License.
VRST '26, Sendai, Japan
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2811-2/2026/11
https://doi.org/10.1145/3822517.3849936

1 Introduction

User-Generated Content (UGC) [6] is central to the experience of metaverse platforms, and avatar–object interactions, such as sitting on a chair or grasping a tool, are a core part of it. Creating such interactive items, however, requires expertise in both 3D modeling and programming, placing them out of reach for most users.

Commercial text-to-3D services [8] now generate a mesh automatically from text, but a mesh alone is not interactive: for an avatar to sit or grasp, the platform needs affordance [2] coordinates, such as a seat center or a grasp point. In prior work [5], these are set by hand in a web application outside VR, forcing users to leave the headset.

We propose InteraCraft, an end-to-end system that generates a 3D object from a single text input, automatically assigns affordance coordinates, and registers it as an interactive item on a commercial metaverse platform, entirely within a VR-HMD. We repurpose a general-purpose VLM as a training-free affordance estimator that outputs these coordinates in metric units. The platform uses them directly, so no step requires leaving the headset.

2 Implementation

To keep the entire workflow inside the HMD, InteraCraft runs as a GPU-less container on Google Cloud Run, which Cluster calls as an external server: the prompt entered in the world is relayed by Cluster to the container, and the finished object is spawned in the world. Gemini 3 [3] serves as the inference engine throughout; both Gemini and the 3D-generation API are swappable modules.

Image generation. Gemini 3 Flash Preview expands the user's natural-language input into a detailed English prompt, from which Gemini 3 Pro Image Preview generates a single orthographic front-view image. This front view is the common reference frame for the rest of the pipeline, keeping the generated mesh and affordance coordinates aligned.

3D generation and affordance analysis. Both run in parallel to reduce latency. Gemini 3 Flash Preview estimates the object's real-world scale and affordance type (*sitting*, *grasping*, or *none*), and Gemini 3 Pro Image Preview annotates the front view with three visual anchors: the bounding box, geometric center, and contact point. Meanwhile, a modularized 3D-generation API (Stability AI [8] in the deployed configuration) produces a textured mesh from the image, which is then scaled to the estimated real-world size, rotated to face front using DINOv2 [7] feature similarity, grounded at $y = 0$, and simplified to meet Cluster's 5,000-triangle limit [1].

Affordance coordinate estimation and upload. Gemini 3 Flash Preview then determines where the avatar should be positioned to sit on or grasp the object and expresses it as a metric offset from the geometric center. It receives the real-world scale, the affordance type, the geometric center of the normalized mesh, and the annotated image, and reasons about the relative pixel positions of the anchors against the real-world scale. This needs no mesh input or task-specific training. The resulting coordinates are embedded as a GLB extension and uploaded to Cluster.

3 Evaluation

3D-generation APIs. Among five commercial 3D-generation APIs (Stability AI, Meshy, Tripo, HiTem3D, and Tencent), we chose Stability AI [8] to keep the in-headset wait short [4]: 27.8 ± 5.5 s from the API call through mesh simplification, against more than 160 s for the others.

Affordance coordinates. We compared the VLM estimates against the affordance coordinates of 13 official Cluster craft items, set and verified by human operators: seven sitting and six grasping. On average, the deployed pipeline's output deviates from these by 14 cm for sitting and 19 cm for grasping in center-relative L2 error. Figure 2 shows one example of each: with a 12 cm error, the avatar sinks slightly into the stool yet still looks seated, whereas with a 19 cm error, a third of the bag's 57 cm width, the hand misses the handle. Since the human-set coordinate is one acceptable placement rather than the only one, this L2 error is a conservative measure of placement accuracy.

End-to-end workflow. We measured pipeline latency per affordance type, with ten trials each for sitting (a wooden chair) and grasping (an electric screwdriver). Three of twenty failed, one from the triangle limit and two from transient API errors. Over the successful runs, the pipeline took 74.9 ± 6.8 s for sitting and 87.0 ± 20.7 s for grasping from prompt submission to upload completion.

4 Conclusion

We presented InteraCraft, an end-to-end system that generates an interactive 3D object from text and deploys it to a commercial metaverse platform. A general-purpose VLM, used as a training-free affordance estimator, replaces the manual coordinate step. This paper does not evaluate with users whether in-HMD authoring is a better experience than out-of-headset authoring, and the numbers reflect current versions of two closed, swappable commercial APIs (Gemini 3 and Stability AI). Future work includes such a controlled user comparison and reducing small-object placement error.

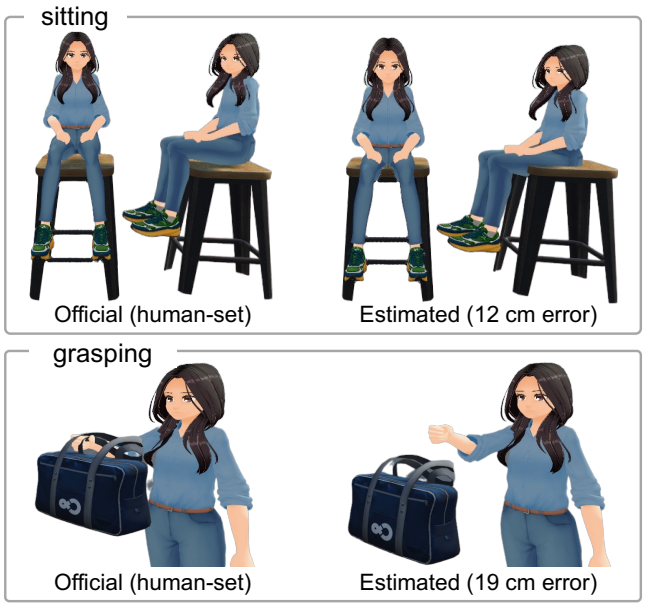


Figure 2: How estimation error appears on the platform, for one stool and one bag. Left: the avatar at the official, human-set affordance coordinates; right: at the estimated ones.

Acknowledgments

This work was supported in part by JST Moonshot R&D Grant Number JPMJMS2013 and JST PRESTO Grant Number JPMJPR2529.

References

[1] Cluster, Inc. [n. d.]. Craft item restrictions. https://docs.cluster.mu/creatorkit/craft-item/limitation/. Accessed: 2026-08-28.
[2] James J. Gibson. 1979. *The Ecological Approach to Visual Perception.* Houghton Mifflin, Boston.
[3] Google. [n. d.]. Gemini models (model list and capabilities). https://ai.google.dev/gemini-api/docs/models. Accessed: 2026-08-28.
[4] Yi-Ting Huang, Chih-Chieh Hsu, and Tzu-Hsuan Wang. 2025. Effects of interactive loading interfaces for virtual reality game environments on time perception, cognitive load, and emotions. *Frontiers in Virtual Reality* 6 (2025), 1540406. doi:10.3389/frvir.2025.1540406
[5] Ryutaro Kurai, Takefumi Hiraki, Yuichi Hiroi, Yutaro Hirao, Monica Perusquía-Hernández, Hideaki Uchiyama, and Kiyoshi Kiyokawa. 2025. MagicCraft: Natural Language-Driven Generation of Dynamic and Interactive 3D Objects for Commercial Metaverse Platforms. *IEEE Access* 13 (2025), 132459–132474. doi:10.1109/ACCESS.2025.3587232
[6] Cory Ondrejka. 2005. Escaping the Gilded Cage: User Created Content and Building the Metaverse. *New York Law School Law Review* 49, 1 (2005), 81–101. https://digitalcommons.nyls.edu/nyls_law_review/vol49/iss1/6/
[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=a68SUt6zFt
[8] Stability AI. [n. d.]. Stability AI Developer Platform. https://platform.stability.ai/docs/getting-started. Accessed: 2026-08-28.