Experiencing Thing2Reality: Transforming 2D Content into Conditioned Multiviews and 3D Gaussian Objects for XR Communication

Erzhen Hu*
University of Virginia
Charlottesville, VA, USA
eh2qs@virginia.edu

Mingyi Li
Northeastern University
Boston, MA, USA
li.mingyi2@northeastern.edu

Xun Qian
Google Research
Mountain View, CA, USA
xunqian@google.com

Alex Olwal
Google Research
Mountain View, CA, USA
olwal@acm.org

David Kim
Google Research
Zurich, Switzerland
kidavid@google.com

Seongkook Heo
University of Virginia
Charlottesville, VA, USA
seongkook@virginia.edu

Ruofei Du†
Google Research
San Francisco, CA, USA
me@duruofei.com

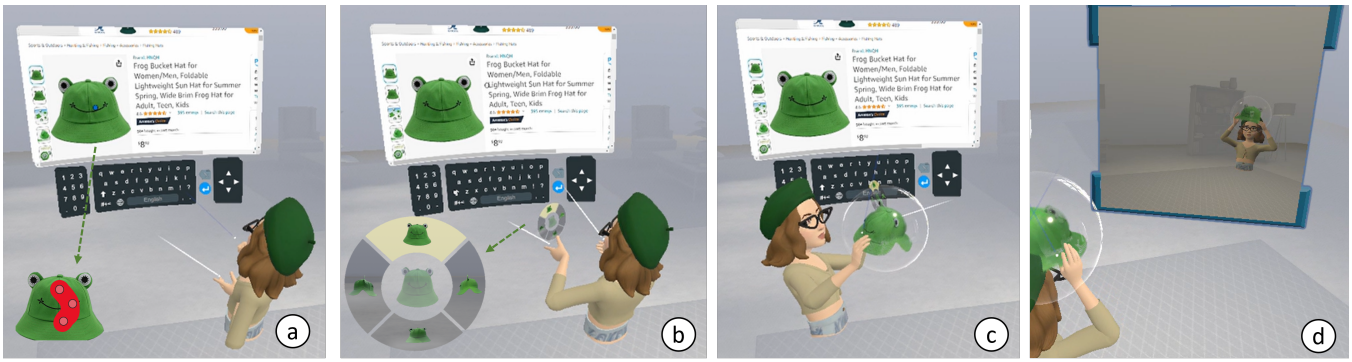

Figure 1: An example user journey: (a) a user begins by selecting a virtual object to bring to reality. This is achieved by marking on the object shown within the web browser or camera feed of the physical space. This virtual object is subsequently processed through progressive stages: starting from a 2D segmented image, evolving into conditioned multi-view renderings, and ultimately, to a 3D Gaussian representation. (b) Meanwhile, the orthogonal multiple views are laid out along the rings of the *Pie Menu*. (c) The 3D Gaussian is summoned after 1-2 seconds. (d) The user can re-position and re-scale it via the *Sphere Proxy*. In this example, the user positions the generated frog hat on her head for the virtual try-on.

ABSTRACT

During remote communication, participants share both digital and physical content, such as product designs, digital assets, and environments, to enhance mutual understanding. Recent advances in augmented communication have facilitated users to swiftly create and share digital 2D copies of physical objects from video feeds into a shared space. However, the conventional 2D representation of digital objects restricts users' ability to spatially reference items in a shared immersive environment. To address these challenges, we propose Thing2Reality, an Extended Reality (XR) communication platform designed to enhance spontaneous discussions regarding both digital and physical items during remote sessions. With Thing2Reality, users can quickly materialize ideas or physical objects in an immersive environment and share them as conditioned multiview renderings or 3D Gaussians. Our system enables users to interact with remote objects or discuss concepts in a collaborative manner.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

extended reality, augmented communication, image-to-3D, remote collaboration, spatial referencing, co-presence

ACM Reference Format:
Erzhen Hu, Mingyi Li, Xun Qian, Alex Olwal, David Kim, Seongkook Heo, and Ruofei Du. 2024. Experiencing Thing2Reality: Transforming 2D Content

*Project conducted when the first author interned at Google.
†Corresponding author.

UIST Adjunct '24, October 13–16, 2024, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-0718-6/24/10.
https://doi.org/10.1145/3672539.3686740

into Conditioned Multiviews and 3D Gaussian Objects for XR Communication. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3672539.3686740

1 INTRODUCTION

Shared artifacts, such as physical objects, printouts, and digital images, play a crucial role in facilitating effective communication and idea generation [3]. They help bridge gaps between collaborators by providing a common spatial reference point and facilitating creative exploration [1]. In addition to physical artifacts, designers often use online platforms like Pinterest and Google to find relevant digital artifacts that can support their design processes [2]. However, using shared artifacts in remote meetings can pose several challenges, especially in scenarios that require quick and spontaneous sharing, such as brainstorming sessions. First, artifacts shared in remote meetings are often in 2D, whether they are captured using a camera or retrieved from an online repository [3]. These 2D representations may not provide the same level of understanding as interacting with a physical object or a 3D model. Second, in physical meetings, participants can easily rotate, manipulate, and interact with the artifacts, which can facilitate creative exploration and idea generation processes [1]. However, in remote meetings, this level of interaction with virtual artifacts generated on-the-fly is often unavailable or limited.

Several methods have been used to address these challenges. One is to prepare 3D models before the meeting by creating or retrieving CAD models, or by 3D-scanning an object [4]. Another is to use a special setup that can capture the physical world in real-time and reconstruct it in 3D [5]. While these methods effectively enable richer sharing of artifacts, they have their own limitations. For instance, using pre-made 3D assets does not effectively support the *spontaneous* sharing of objects, and using a special scanning setup may not be accessible for many people. On the other hand, recent advances in AI-driven text-to-3D and image-to-3D technologies [7] address the need for a more accessible and efficient way of creating and sharing 3D assets. These technologies can significantly lower the barriers to 3D content creation, enabling individuals without specialized skills to contribute to the co-creation process, thereby democratizing access to 3D modeling and enhancing collaboration.

To address the challenges of summoning spontaneous 3D representations into the existing information space, we seek to enable fluid communication in an XR environment comprised of paired 2D and 3D artifacts. In this demonstration, we present Thing2Reality, a distributed communication system that enables users to segment any content from any container (video streams, shared digital screens) in the XR environment (Figure 1a), explore the perspectives with multi-view renderings (Figure 1b), and transform them into shared 3D Gaussians (Figure 1c-d) for 3D manipulation.

2 THING2REALITY WALKTHROUGH

The virtual environment was developed using Unity 2022.3.19f1 and the following SDKs: Oculus Interaction Toolkit, Meta Avatar SDK for rendering avatars, gestures, and lip-syncing, and Photon Fusion and Voice SDK for voice streaming between user avatars. The study program runs on a desktop workstation with an Intel Core i7-13700K processor and an NVIDIA RTX 4070 Ti GPU for MobileSAM [9], text-conditioned [6] and image-conditioned [8] multi-view diffusion models, and Large Gaussian Models [7] to fuse multiview renderings into interactive 3D Gaussians. The Unity program is deployed on two workstations with an Intel Core i9-9900K CPU, 32GB RAM, and an NVIDIA GeForce RTX 4070 GPU, each connected to a Meta Quest 3 headset with two touch controllers. All computing devices were connected to the same local network.

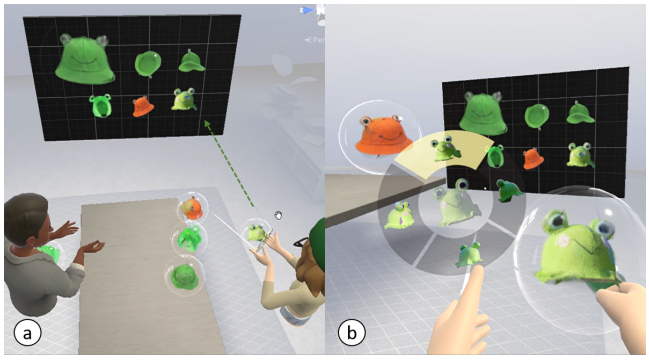

Figure 2: 3D-to-2D: A user can capture snapshots from different perspectives of the 3D Gaussians, and project it on the whiteboard. (a) Third-person view; (b) First-person view.

Here we show the default interaction workflow with an example of digital search. The user journey is presented for the 2D-to-3D (Figure 1) and 3D-to-2D (Figure 2) processes. Different from reconstructing and creating virtual 3D replicas, Thing2Reality focuses on generative methods that turn 2D contents into 3D Gaussians.

Interactive Object Segmentation (Figure 1a). The user can identify object of interest by marking it while holding both the grip and trigger buttons on the controller and moving the pointer along the object. This samples three points to be used for object segmentation. Once MobileSAM [9] segments the object from the image, the segmentation result will be shown to the user. The user can then confirm to send the images for multi-view rendering and 3D Gaussian creation.

2D-to-3D: Creation of Multi-views and 3D Generated Objects (Figure 1b-d). After the user confirms the segmented object, the multiple conditioned views will be rendered on a **2D Pie Menu** (Figure 1b) attached to the user's left controller. The center of the Pie Menu shows the original image being cropped from the data source (*e.g.*, the webviews). The four orthogonal views of the original image, generated with conditioned diffusion models, will be displayed on the top (front view), left (side views), right (side views) and bottom (back views) of the outer ring of the Pie Menu. The user can show or hide it by pressing the "X" button on the controller.

Fusing the multi-view images with 3D Gaussian splatting [7], a 3D generated object will then become available as an shared object in the environment. A semi-transparent sphere, named **Sphere Proxy** (Figure 1c), will be created around the generated 3D Gaussian as a collider for users to grab, move, and resize the object. The Sphere Proxy will become invisible when the controller moves away from the object for clearer view of the object. The orthogonal

views of a 3D objects on the 2D Pie Menu are only visible to the user who created it, but it can be achieved from any shared 3D objects generated in the environment.

3D-to-2D: Projecting Things to Surrounding Whiteboard and Table for Workspace Communication (Figure 2). The user can take a snapshot from any angles of the generated 3D objects under the field of view of the user, and project the snapshot on collaborative surfaces like whiteboard and table (Figure 2a). This is different from the *discrete* orthogonal views due to the *continuous* perspectives an 3D object presented.

Users can press the trigger button to select and drag 2D snapshots on the whiteboard. Rescaling is achieved by selecting the object and adjusting its size using the thumbstick. To delete an object, users can press the "B" button on the controller. Users can select discrete orthogonal views from their 2D Pie Menu and project onto the shared whiteboard. Additionally, the central image can be projected onto the whiteboard to display a 360-degree video of the object.

3 CONCLUSION

In this demonstration, we present Thing2Reality, an XR communication system that allows users to instantly materialize ideas or physical objects and share conditioned multiview renderings or 3D Gaussians for realistic 3D rendering. We believe that XR communication has tremendous promise for co-presence and for bridging distances between humans, and enabling the spontaneous creation and sharing of 3D objects and artifacts in XR will allow for a more fluid and effective exchange of ideas, beyond what is possible in real-world communication.

ACKNOWLEDGMENTS

This work was supported in part by a White Ruffin Byron Center for Real Estate grant.

REFERENCES

[1] Margot Brereton and Ben McGarry. 2000. An observational study of how objects support engineering design thinking and communication: implications for the design of tangible media. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 217–224.

[2] Scarlett R Herring, Chia-Chen Chang, Jesse Krantzler, and Brian P Bailey. 2009. Getting Inspired! Understanding How and Why Examples Are Used in Creative Design Practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 87–96. https://doi.org/10.1145/1518701.1518717

[3] Erzhen Hu, Jens Emil Sloth Grønbæk, Wen Ying, Ruofei Du, and Seongkook Heo. 2023. ThingShare: Ad-Hoc Digital Copies of Physical Objects for Sharing Things in Video Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 365, 22 pages. https://doi.org/10.1145/3544548.3581148

[4] Shahram Izadi, Andrew Davison, Andrew Fitzgibbon, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Dustin Freeman. 2011. KinectFusion: Real-Time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology - UIST '11*. ACM. https://doi.org/10.1145/2047196.2047270

[5] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*. 741–754.

[6] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2023. Mvdream: Multi-View Diffusion for 3d Generation. *ArXiv Preprint ArXiv:2308.16512* (2023).

[7] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2024. LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. *ArXiv Preprint ArXiv:2402.05054* (2024). https://arxiv.org/pdf/2402.05054.

[8] Peng Wang and Yichun Shi. 2023. ImageDream: Image-Prompt Multi-View Diffusion for 3D Generation. *ArXiv Preprint ArXiv:2312.02201* (2023).

[9] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. 2023. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *ArXiv Preprint ArXiv:2306.14289* (2023). https://doi.org/10.48550/arXiv.2306.14289