Thing2Reality: Enabling Spontaneous Creation of 3D Objects from 2D Content using Generative AI in XR Meetings

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

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

Andrew Hong
University of Virginia
Charlottesville, VA, USA
rsv5fd@virginia.edu

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

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

David Kim
Google XR Labs
Zurich, Switzerland
kidavid@google.com

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

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

Figure 1: An example use case of Thing2Reality. Alice and Charlie are discussing room decorations in a shared XR space (b). Alice begins by bringing a shelf from her physical office (a) into the virtual environment. She then searches for a cute cat planter using the web browser interface. With Thing2Reality, she summons 3D Gaussian of the planter and places it onto the virtual shelf. Alice and Charlie then engage in a discussion about various planter designs, projecting 3D Gaussian representations of the planters (c) onto a whiteboard in the space. This allows them to transform 2D images into interactive 3D objects, which can be collectively viewed, manipulated, and compared in real-time, facilitating a seamless and collaborative ideation process.

ABSTRACT

During remote communication, participants often 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, conventional 2D representations of digital objects limits spatial referencing in immersive environments. To address this, we propose Thing2Reality, an Extended Reality (XR) meeting platform that facilitates spontaneous discussions of both digital and physical items during remote sessions. With Thing2Reality, users can quickly materialize ideas or objects in immersive environments and share them as conditioned multiview renderings or 3D Gaussians. Thing2Reality enables users to interact with remote objects or discuss concepts in a collaborative manner. Our user studies revealed that the ability to interact with and manipulate 3D representations of objects significantly enhances the efficiency of discussions, with the potential to augment discussion of 2D artifacts.

*This project was undertaken during the first author's internship at Google XR, with additional support from a Google PhD Fellowship awarded by Google Research.

†Corresponding author

UIST '25, September 28-October 1, 2025, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/2025/09.
https://doi.org/10.1145/3746059.3747621

CCS CONCEPTS

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

KEYWORDS

Extended Reality, Augmented Communication, Image-to-3D, Information Artifacts, Multi-modal, Remote Collaboration

1 INTRODUCTION

Shared artifacts, including digital resources (*e.g.*, text, images, videos), and physical objects (*e.g.*, prototypes, printouts), play a crucial role in facilitating effective communication of spatial concepts and the generation of design ideas, especially in creative fields such as product design, architecture, and marketing strategy development. They provide common spatial reference points that bridge gaps between collaborators, enhancing creative exploration and ideation [4]. Besides physical artifacts, designers frequently use online platforms like Google and Pinterest to source relevant digital artifacts that can support their design processes [24]. However, using shared artifacts in remote meetings often pose challenges, especially in scenarios that require quick and spontaneous sharing, such as brainstorming and early stage design sessions. First, artifacts shared via remote meetings are typically in 2D, whether they are captured via camera or retrieved from online repositories, limiting the understanding compared to interactions with physical objects or 3D models. Second, in physical meetings, participants can easily observe and interact with tangible artifacts, which facilitates creative exploration and idea generation processes [4]. However, in remote meetings, this level of interaction is often unavailable or limited.

Several methods have attempted to address these challenges, such as preparing 3D models in advance via CAD or 3D scanning [34], or employing specialized real-time 3D capture setups [62, 78]. While effective, these approaches have limitations: pre-made 3D assets do not support *spontaneous* sharing, and specialized setups are often impractical for general use. Recent advances in AI-driven text-to-3D and image-to-3D technologies [75] present an accessible and intuitive alternative, lowering barriers to 3D content creation and enabling broader participation in collaborative efforts.

In this paper, we aim to investigate how on-the-fly transformations between 2D content and 3D representations can support object-centric ideation and spatial sense-making in collaborative XR meetings, and how the design of such a system can enhance user experiences by integrating interactive image-to-3D workflows across various XR meeting context.

We designed and implemented Thing2Reality, an XR meeting platform that enables fluid interactions with 2D and 3D artifacts. Thing2Reality allows users to segment content from any source (video streams, shared digital screens) within the XR environment (Figure 1a), generate multi-view renderings (Figure 1b) with conditioned multi-view diffusion models, and transform them into shared 3D Objects with Gaussian splatting for interactive manipulation (Figure 1c). We conducted two user studies: a preliminary study (N=12) analyzing the usability of using digital and physical sources for 3D object creation, and an exploratory study (N=18) exploring how the co-existence and transformation between 2D and 3D formats shape collaboration patterns across tasks such as avatar decoration, spatial layout, and open-ended design brainstorming. Our findings suggest that 3D objects facilitate intuitive explanations, detailed visualization, and interactive collaboration, whereas 2D representations are more often used for final pitch deliverables, suggesting a context-dependent trade-off between the two formats based on task objectives.

In summary, we contribute:

- **Thing2Reality, an XR meeting platform** that provides on-the-fly 3D objects generation by enabling users to present and share spontaneous thoughts, and augment their shared digital and physical artifacts with remote collaborators.
- **Findings from a usability study** (N=12) examining the digital and physical inputs of Thing2Reality.
- **Findings from an exploratory user study** (N=18) evaluating the use of Thing2Reality (both 2D-to-3D and 3D-to-2D workflow) for discussing and presenting both 2D and 3D objects in XR meetings.

2 RELATED WORK

Our work is inspired by prior art on distributed communication around 2D and 3D artifacts, task-space collaboration, and emerging 3D generation techniques that span both computer-mediated work and XR systems.

2.1 Distributed Communication and Collaboration in XR

Remote collaboration platforms have evolved significantly, with researchers exploring various approaches to enhance computer-mediated communication through shared visual content [29, 30]. Drawing on prior research in computer-mediated cooperative work, shared media such as screen sharing and image annotation serve as crucial "common ground" elements that facilitate mutual understanding and support referential communication [53, 56]. However, these 2D channels — while effective for documents or slides — struggle to convey the spatial nuances of physical objects or complex 3D scenes.

XR-based meeting platforms promise higher co-presence and natural referencing, with the ability to place collaborators and shared artifacts in a unified 3D environment [55, 79]. They have been applied to and studied in diverse domains such as education [57, 65], entertainment/gaming [10, 86, 90], and physical task demonstration [45, 78]. Despite these advances, many XR systems rely on static or pre-made 3D assets; they offer limited support for *spontaneously* sharing new objects or seamlessly shifting between 2D and 3D media in real time.

A persistent challenge in both traditional 2D and XR-based communication is supporting "object-focused collaboration", where discussions center around physical artifacts [31, 47]. Research has identified several critical challenges in this domain, including the coordination of viewpoints [22, 64], the communication of gaze [23] and gestural information [76], and the management of object orientations [12]. While XR environments can support multiple shared perspectives and virtual replicas [60], creating these representations typically requires specialized capture equipment or complex setup procedures. This creates a tension between the need for quick, flexible 2D sharing and the desire for richer, more spatially aware 3D collaboration—a challenge that motivates our approach in Thing2Reality.

2.2 Advances in 3D Content Generation

Efficiently creating 3D representations remains a challenge for remote XR collaboration. Traditional modeling with CAD tools is often slow and requires expertise [11, 60]. Depth-based reconstruction using RGB-D sensors [34, 62] enabled faster capture of real-world geometry but requires specialized hardware setups.

More recently, neural representations such as NeRF [33, 70] have improved the fidelity of 3D reconstructions from multi-view images, offering realistic scene rendering. Meanwhile, generative AI approaches have emerged as a promising direction for 3D content creation. Text-to-3D and image-to-3D methods, such as DreamFusion [63], focus on score distillation sampling (SDS) that utilizes pretrained 2D diffusion models to generate 3D content, but faces problems with speed and diversity. Recent advances in large reconstruction models [28, 46] use non-SDS methods. Large Gaussian Models (LGM) [75] use similar methods to [40], with algorithms to convert 3D Gaussian into meshes. Advances using multi-view diffusion models as a prior have also made generation of complex, textured 3D models possible. These generative models provide a foundation for transforming 2D visuals into 3D representations. While foundational models primarily relied on text prompts [44], some approaches [41, 88] incorporate spatial controls through depth maps, skeletons, or point-based inputs.

Thing2Reality leverages these recent AI breakthroughs, integrating Gaussian splatting [40] and LGM pipelines [75] into an XR collaboration context. By enabling users to highlight regions of interest in a 2D snapshot (or otherwise provide spatial prompts), our system can *instantly* produce lightweight, manipulable 3D objects without requiring specialized hardware. We believe that XR meetings can benefit from additional perspectives of shared artifacts, whether sourced from digital content or physical environments. This motivates us to better understand how AI-generated views from single images invite participants to explore novel concepts for idea generation, and how they may support or hinder communication through generated visual details.

2.3 2D and 3D Content in Shared Task Spaces

Building upon Buxton's framework [6], many researchers have looked at ways to embed shared artifacts—whether physical or digital—directly within remote meeting systems, as shared task spaces. The evolution of shared task spaces has seen a progression from simple 2D sharing to increasingly sophisticated 3D representations. Early approaches focused predominantly on 2D visual information. For instance, IllumiShare [38] allowed sharing of physical and digital content on any arbitrary surface, while ThingShare [31] recently incorporated 2D snapshots of physical objects to support remote discussion. Systems such as Visual Captions [49] further enrich 2D conferencing by automatically augmenting language with visual aids, showing how 2D imagery can enhance clarity in communication. These systems, along with research on cohabiting physical and digital artifacts [20, 38, 69], demonstrated how 2D representations could enhance collaborative tasks. Media space research [50, 51, 61] expanded these capabilities through multi-camera setups and shared desk areas. However, these 2D approaches struggled to capture the spatial relationships and physical affordances crucial for object-centered tasks.

To fill this gap, researchers have explored point-cloud renderings and volumetric telepresence, enabling lifelike 3D captures of users or objects [62, 70]. Systems like SharedNeRF [70] and VirtualNexus [33] leverage neural radiance fields (NeRFs) for high-fidelity scene reconstruction. However, they are limited to capturing and reconstructing physical environments and objects, and require either specialized hardware or multi-view captures of the same object. Recent advances in generative AI have enabled the creation of unified 2D and 3D virtual environments that support co-presence in new ways. For instance, BlendScape [66] uses Stable Diffusion and inpainting techniques to synthesize shared virtual spaces, while SpaceBlender [59] transforms user-provided 2D images into immersive 3D environments for telepresence. Despite these advances, Current research has yet to explore the possibilities of creating 3D objects from single images and other 2D content, such as text, sketch, or digital search, during distributed collaboration [58, 67].

In contrast, Thing2Reality adopts a more flexible approach by enabling real-time transformation between 2D and 3D formats. Similar to Loki [78] that integrates 2D and 3D media, Thing2Reality supports 2D images and videos as well as 3D Gaussians, but crucially does so with single 2D image as the input. This design allows users to create 3D representations directly from minimal 2D sources (such as digital search, or camera feed), expanding the possibilities for collaborative workflows. By moving beyond a strict 2D-versus-3D dichotomy, Thing2Reality further focuses on how fluid transitions between these formats (2D-to-3D and 3D-to-2D) can improve remote communication, foster shared understanding, and address a wider spectrum of collaborative needs.

2.4 Summary: A Design Space for Thing2Reality

We situate Thing2Reality into prior literature of collaborative work around 2D and/or 3D artifacts (Table 1). Prior work also explores different ways of creating pre-made or catalog assets, such as using gestures to approximate and imitate the object [27] among a database of known objects, or understanding the role of virtual replicas in remote assistance [60, 79]. We did not include this line of work because we focused on the spontaneity of sharing things during the communication regarding the spatial arrangement, scale, and aesthetic properties of objects.

The media representation dimension (2D vs. 3D) captures how systems support different forms of visual representation, ranging from 2D-only or 3D-only formats to a small number of systems that integrate both. We also distinguish systems by their sharing granularity, comparing scene-level (SL) and object-level (OL) sharing. Our work builds on prior research in *information artifacts*[53] and *object-focused collaboration* [12, 26, 38], in contrast to scene-level customization and editing explored in systems like [14, 48, 59, 66].

Through this design space, we emphasize object-level interactions across three dimensions: spontaneous generation, transformation, and cohabitation of 2D and 3D artifacts. Our approach enables users to identify objects of interest from both digital information spaces and physical environments.

Method	2D	3D	Scene-Level (SL)	Object-Level (OL)
IllumiShare [38]	✓		✓	
Remixed-Reality [48]		✓	✓	
SharedNeRF [70]		✓	✓	
Loki [78]	✓	✓	✓	✓
ThingShare [31]	✓		✓	
Holoporation [62]		✓		✓
Visual Captions [49]	✓		✓	
BlendScape [66]	✓		✓	
Thing2Reality	✓	✓	✓	✓

Table 1: Comparison of related systems based on media representation and sharing granularity.

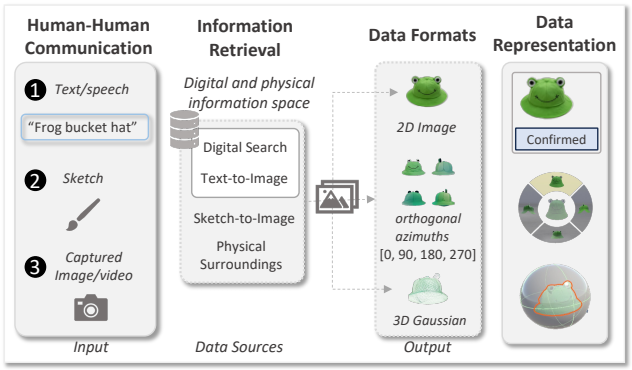

Figure 2: Human-human communication methods: 1) text or speech, 2) sketch, 3) images or videos can be used as *input* to achieve ideal 2D images via digital search, image/video capturing, or GenAI/ML models (*text-to-image*, *sketch-to-image*), which can then be converted to arbitrary segmented image, conditioned multiview renderings, and 3D Gaussian.

3 THING2REALITY SYSTEM OVERVIEW

A key takeaway from our design space highlighted the effectiveness of integrating 3D object affordances with the spatial organization advantages of 2D artifacts, which motivated the design of Thing2Reality. Before delving into the system's design, it is crucial to clarify the definition of "Thing" in the context of our work.

3.1 What the "Thing"? Exploring the Role of User-Generated 3D Assets in Spontaneous Communication

We outline the typical approaches individuals take when spontaneously incorporating various artifacts (*i.e.*, sketches, searched images, and physical objects) into discussions as a source of inspiration, explanation, or clarification (Figure 2).

- **Text-based content (Figure 3 - 1)**: Text-based content uses language to convey ideas, including written descriptions, transcribed speech, and notes. Text-based contents can be transformed into 2D images use text-to-image methods like Imagen[1].
- **Hand-created visual content (Figure 3 - 2):** Hand-created visual content encompasses manually produced images, diagrams, or visual representations, either physical or digital. This includes sketches, drawings, and hand-drawn diagrams, providing intuitive and spontaneous representations of ideas, spatial relationships, or abstract concepts in communication. Current methods such as ControlNet [88] use sketches as one of the ways for controlling image generation.
- **Digital visual content (Figure 3 - 1):** Images found through online searches like Google images or Pinterest, screenshots, and digital artwork stock photos to find images that closely align with their discussion topics, utilizing these images as a reference point [24].
- **Captured real-world content (Figure 3 - 3)**: Photographs or scans of physical objects and environments, which can serve as a powerful means of conveying ideas, but their integration poses challenges for distributed users [4, 31], who might opt to digitally capture and share these items. It is important to note that these digital 3D representations of real-world content do not always capture the specific details of an object as accurately as a virtual replica (*e.g.*, NeRF). Instead, they serve as a proxy for the original object. Furthermore, it can be hard for users to reconstruct an item when some sides of it is not easily capturable (*e.g.*, a large shelf), or when it's difficult to capture at close range.

From an HCI perspective, we aim to explore how images (or objects of interest) can serve as a *middle ground*, connecting a person's inputs—such as text, sketches, and digital searches—with their intentions to communicate (Figure 2). These inputs can be dynamically converted into flexible data representations (spanning both 2D and 3D), facilitating spontaneous and effective communication with others. Transforming these variations of 2D content into 3D objects can help enhance the immediacy and tangible engagement with abstract concepts, such that users can gain a deeper mutual understanding during discussions. Furthermore, text-to-image and sketch-to-image generation methods often produce less predictable results due to the vagueness of such inputs, making precise control challenging. In contrast, searching for images or capturing real-world content indicates more direct and explicit selection.

Recognizing this difference in control between AI-based generation and human-curated or directly captured visual content, Thing2Reality's design primarily focuses on **digital visual content** and **captured real-world content** with image-to-3D approaches. This ensures more precise control over the images used in communication, enhancing the system's reliability and user experience.

3.2 Design Goals

Based on the design space and prior literature, we formulated the following three objectives.

DG1 Spontaneity: Enable Spontaneous Communication Using Digital and Physical 3D Artifacts As Visual Aids. Acknowledging the importance of both physical and digital artifacts during collaboration, we aim to facilitate a seamless conversion of 2D artifacts from diverse data sources (*e.g.*, digital files and physical objects via camera feeds) into 3D representations.

DG2 Cohabitation: Support for Co-Habitation of 2D and 3D Objects During Communication. While remote conferencing typically relies on single data modalities like 2D visuals [1], some

[1]Imagen: https://deepmind.google/models/imagen/

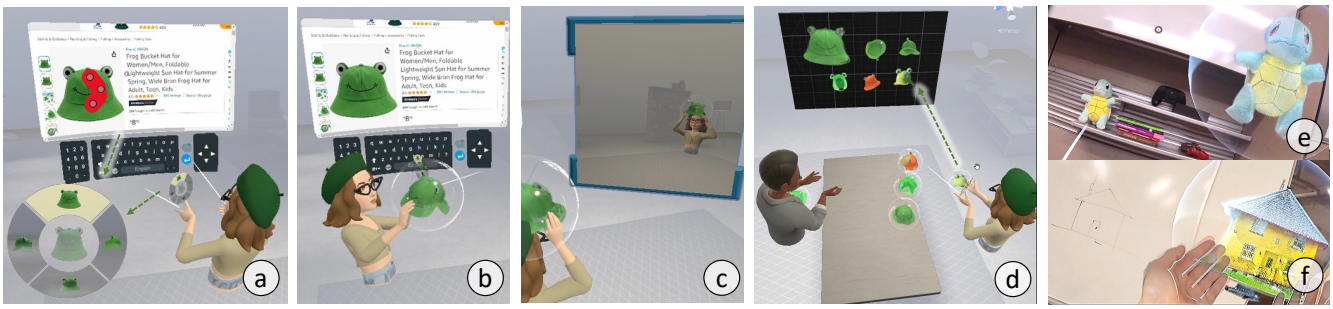

Figure 3: An example user journey: (a) a user selects content by painting regions in the web browser or camera feed. The system processes these selections from 2D segmentation through multi-view rendering to 3D Gaussian representation. The orthogonal views appear in the Pie Menu rings, (b) followed by the 3D Gaussian object appearing within 1-2 seconds. (c) The user can re-position and re-scale it via the *Sphere Proxy*. (d) Users can capture snapshots from different perspectives of the 3D Gaussians, and project it on the whiteboard for collaboration. (e-f) with video see-through mode, users can bring physical objects and sketches to the shared space in XR.

work (*e.g.*, [78]) demonstrates incorporating multiple modalities enhances remote guidance and understanding. To facilitate efficient discussions and collaborative sessions, Thing2Reality should enable users to flexibly choose and combine different data representations to better support their specific communication needs during collaborative sessions.

DG3 Transition: Enable Flexible Bi-Directional Transformations Among Digital Media Forms (i.e., 2D images, videos, and 3D). XR workspaces can be more dynamic and open compared to traditional videoconferencing, and the presence of multiple data modalities may introduce friction for users [15, 77]. Recognizing the diverse needs of remote collaboration with multiple data modalities in **DG2**, it is also essential to allow users to frictionlessly switch between different forms of these representations (2D images, videos, multi-view representations, and 3D models) according to the context of their discussion. This requirement would imply that Thing2Reality should not only store and organize various forms of media but also allow for their easy retrieval and transformation during discussions. By enabling flexible bi-directional transitions between digital media forms (2D-to-3D, 3D-to-2D), users can adapt their communication style to the specific requirements of the task at hand, leading to more efficient and effective collaboration.

3.3 Overview and Interaction Workflow

To balance comfort and input accuracy in collaborative XR sessions, Thing2Reality prioritized controller-based interactions similar to prior work on mixed reality collaboration [19, 78] over other interaction choices of midair gestures or multimodal inputs [25, 36, 42, 52, 81].

Main Components. Three main components were incorporated for the XR workspace.

- *interactive portals*, including browsers for accessing digital content and portals connecting to physical environments;
- *collaborative surfaces*, like whiteboards and tables, which structure user formations, and allows collaboration activities;
- *avatars*, remote participants are represented as embodied avatars who can freely navigate the space, with shared head and hand gestures, body orientation and position in the space.

While existing XR meeting platforms like Mozilla Hubs and Meta Workroom offer common collaborative features as above, such as browsers, shared surfaces, and remote avatars, Thing2Reality uniquely enables fluid transitions between 2D and 3D content. Our system extends beyond traditional XR collaboration by supporting spontaneous identification and bi-directional transformation of content in shared spaces, allowing users to seamlessly switch between formats based on their communication needs.

3.3.1 Workflow. Here we describe the default interaction workflow using the example of a digital search. A key distinction from prior object-focused collaboration work is that our *information artifacts* are not merely brought into meetings but are also actively generated and transformed (*e.g.*, 2D-to-3D, 3D-to-2D) during meetings, enabling on-the-fly illustration of ideas.

Interactive Object Segmentation. Multi-view diffusion models [72] generate images that often contain varied backgrounds, which requires precise object identification and segmentation before 3D fusing. To address this challenge in XR environments, we developed an interactive object segmentation system that builds upon two established approaches: the marking paradigm from *Interactive Graph Cuts* [3], which demonstrated the effectiveness of user-guided mark-based segmentation, and the Segment Anything Model (SAM), a state-of-the-art prompt-based segmentation method. Instead of relying on multiple precise clicks, which is impractical with VR controllers, our system leverages continuous marking gestures to extract three key points. These points act as intelligent prompts for SAM's segmentation algorithm, achieving a balance between segmentation accuracy and continuous user interactions in VR.

Users identify objects of interest through a simple interaction: holding the controller's grip button while using the trigger to mark start and end points. Segmented results are displayed alongside the user's hands for review before proceeding to multi-view rendering

and 3D Gaussian creation. This demonstrates **DG1**, enabling spontaneous identification and transformation of any 2D content into 3D through ad-hoc interactions.

Ad-Hoc Creation of Multi-Views and 3D Generated Objects (2D-to-3D). Prior work demonstrates the value of remote collaboration via 3D virtual replicas in immersive environments [33, 89], and the effectiveness of AI-generated 2D elements in furnishing video meeting activities [66]. Building on this work, we explore how dynamically generated 3D objects created from 2D content can enhance collaborative interactions while maintaining on-demand flexibility.

After the user confirms the object of interest, the multiple conditioned views will be rendered on a 2D **Pie Menu** (Figure 3b) attached to the user's left controller. The center of the Pie Menu shows the original image being cropped from the data source (webviews, images from physical space) or generated from the image generation model. The four orthogonal views, 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. Selecting the central image also displays a 360° video of the object. The user can show or hide it by pressing the "X" button on the controller.

Once generated, the 3D object becomes a shared entity in the environment. Users can interact with it collaboratively, moving, grabbing, or re-scaling it through the semi-transparent Sphere Proxy (Figure 3c), which serves as a collider.

The design of Pie Menu and the 3D Gaussian leverages the cognitive benefits of orthogonal 2D views for understanding 3D structures [7, 80]. The orthogonal views on the 2D Pie Menu remain private to the creator/current holder, where everyone can achieve it from any 3D objects. The combination of private visualization of orthogonal views and shared 3D object interaction enables users to examine object details independently without overwhelming the shared workspace with extraneous visuals. This illustrates the **DG2** that enables the cohabitation of 2D images, videos, and 3D objects in a unified workspace for users to choose and combine flexibly.

Projecting Continuous Perspectives of Shared 3D Gaussians to Surrounding Collaborative Surfaces (3D-to-2D) . The table and whiteboard surfaces can act as shared information space among users [17, 18], where prior work has examined *spatially continuous workspaces* [13, 68] for users to seamlessly move digital 2D content from portable computers, to the table and wall displays as shared workspaces for group collaboration. The system allows users to capture snapshots of generated 3D objects from any angle and project these perspectives onto collaborative surfaces, such as a whiteboard or a table (Figure 3). Unlike the discrete orthogonal views, these snapshots provide *continuous* perspectives by using 3D Gaussians as a proxy. This functionality addresses the need for dynamic visual communication during collaborative tasks.

To interact with projected snapshots, users can drag objects using raycasting, rescale them via the thumbstick's Y-axis, or delete them by pressing the "B" button. Users can also select discrete orthogonal views from their private 2D menu, which can be projected on the whiteboard. The central image can be projected on the whiteboard to show 360-degree videos of the object, supporting a holistic understanding of the object. This demonstrates the bi-directional transformations between 2D and 3D objects (**DG3**) that allowed users to adapt their communication methods based on task requirements for efficient object sharing.

4 IMPLEMENTATION

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 avatars, and ZED SDK for physical space sensing.

System Setup. The system operates on an Intel i7-13700K CPU and NVIDIA RTX 4070 Ti GPU, running MobileSAM [87], text-conditioned [72] and image-conditioned [83] multi-view diffusion models, and Large Gaussian Models [75] to fuse multiview renderings into interactive 3D Gaussians. Due to the privacy issues of capturing data from physical environments via current passthrough technologies of VR/MR HMD, we mounted a ZED Mini stereo camera [2] on the Meta Quest 3.

3D Object Rendering Pipeline. A Unity web browser plugin [82] enables the integration of web content directly into 3D virtual environments through webview functionality. Real-world elements are incorporated into the XR experience using image frames captured by the ZED Mini camera. These frames are seamlessly rendered within the Unity environment, allowing users to interact with both physical and digital content in a unified space.

Hence, users can identify objects in both web-based content and live camera feeds capturing physical spaces via ZED and turn the identified 2D images into 3D. This includes the ability to perform actions such as making strokes or taking snapshots. Interactions within the virtual environment are managed through a custom event listener via a Python Flask server. The original input (Figure 4:3) is the selected image frame with the three points on the selected image frame filtered from the user's stroke interaction using the raycaster of the Meta Quest 3 controller.

The Unity application communicates with the Python Flask Server via HTTP POST requests (Figure 4:4). activating models for quick segmentation. This process identifies objects of interest based on user interactions. The 3D Gaussian output can be visualzied as a 2D 360 video, or a 3D visual effect in the Unity environment. The 3D Gaussian were then imported and visualized in Unity as Gaussian splats, surrounded by a semi-transparent sphere around it as a proxy collider to enable interactions like grab, scale, and move with hands or controllers. System performance metrics are detailed in the Appendix in Figure 11.

5 APPLICATION

While our primary focus was on object-centric collaborative work and meeting scenarios, from lightweight social interactions to focused object discussions, Thing2Reality's applications can extend beyond workplace collaboration. Our exploratory study revealed the potential to enhance casual social interactions through 3D object generation. Users naturally progressed from virtual try-ons to sharing personal items during informal conversations [16], suggesting applications from coffee chats to family gatherings.

[2] ZED Mini: https://store.stereolabs.com/products/zed-mini

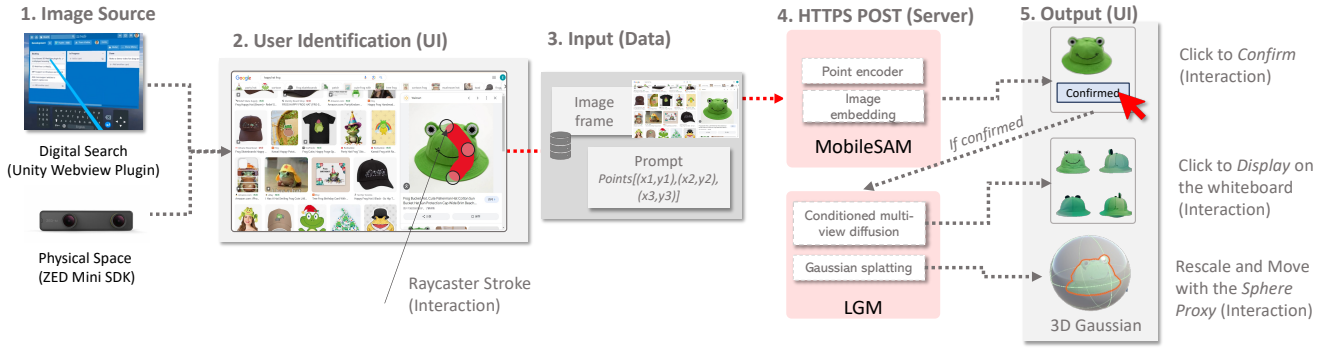

Figure 4: Implementation Diagram of Object-Level Thing-to-3D between Unity and Python Flask Server.

Figure 5: Setup for physical object scanning in Study 1

While 2D emojis, GIFs, and memes are widely used in online communication, Thing2Reality can be used to allow users to transform online images or memes into the 3D version. This feature enhances social gatherings, and VR live streams, moving beyond pre-designed 3D emojis like those in Meta Workroom. Different from Apple's recent Genmojis[3] that enabled 2D text-to-emoji in text-based digital interactions, Thing2Reality might enable more intuitive and spontaneous user-controlled 3D expressions during XR meetings. Building on research in remote reading and social play [35, 86], these capabilities could particularly benefit long-distance and intergenerational connections.

6 SINGLE-USER USABILITY STUDY

The goal of the first IRB-approved study is to evaluate the usability of Thing2Reality and to identify limitations or opportunities for future improvements. We recruited 12 participants (8 male, 4 female, ages 20-33, $\bar{x} = 26.1$) from our university. All sessions consisted of two tasks. In the first task, the experimenter demonstrates all of the basic functionalities of Thing2Reality by going through the workflow of 2D-to-3D, and 3D-to-2D, and communication around objects with the experimenter in the virtual space. Then, the participants were asked to perform example tasks without the experimenter's help, which includes (1) capturing the digital and physical objects, and turn them into 3D, and then (2) decorate the wall by projecting some objects as 2D stickers, and (3) finally find a personalized object to decorate their avatars, and communicate with the experimenter in the virtual space to simulate casual meet up. We chose two examples (room decoration and avatar decoration) as evaluation tasks. Physical objects were placed in the environment for physical object scanning (Figure 5). After the session, we asked the participants to give feedback about the interface and interactions with a survey. Sessions lasted around 45-60 minutes and participants were compensated USD $20.

6.1 Findings

All participants successfully completed the assigned task. Overall, they responded positively to the usability, potential applications, and unique affordances of Thing2Reality. They found the interactions intuitive, easy to learn, and enjoyable, and recognized its potential for supporting ad-hoc communication and object-centered collaboration. Participants rated the overall workflow highly (Mean = 4.9/5), with engagement receiving an average score (Mean = 5/5).

Comparison Between Physical and Digital Inputs. For the physical and digital image sources, participants found that interacting with physical objects was generally simpler than interacting with digital ones. This was largely due to the complexity of web pages, which often include numerous cookies and advertisements—making typing, clicking, and navigating significantly harder than on a laptop. As one participant noted, *"3D objects scanned better when they were in the real space (vs. on Google). They were more accurate, provided they weren't obscured by others (like with the plushies)."* Physical scanning also raised expectations around detail retention and accuracy. In contrast, digital images—such as those from Amazon—were often unfamiliar in 3D form. One participant shared, *"It's helpful to bring in images from Amazon and see where I want it,"* suggesting that digital inputs allowed for quick object-centric ideation brainstorming, even if the 3D rendering wasn't always accurate.

Digital Inputs Provide More Flexibility. Despite this, some participants mentioned the *"lack of options"* in physical environments and suggested that the physical input may be more suitable for casual use cases, rather than professional settings, especially when scanning personal items that cannot easily be found online. In contrast, digital sources were viewed as more flexible: *"It was easy to find the object I was looking for, and I could easily scan it again."*

[3]Genmojis: https://www.apple.com/newsroom/2024/06/introducing-apple-intelligence-for-iphone-ipad-and-mac/

3D Object Quality. In terms of output quality, participants saw Thing2Reality as useful for early-stage design, such as conceptual layouts or rough prototyping. However, they found the current 3D models too blurry for final decision-making. One participant remarked, *"It would be helpful to get a rough idea on how to organize ideas/room design, but higher quality would be better when making a final decision—like whether I actually like that striped pattern on a shirt."* Still, the visual and interactive nature of 3D representations was appreciated for its speed and accessibility: *"For every design and learning scenario, it is easier, faster, and cheaper."*

Finally, participants observed that Thing2Reality primarily supports 3D objects. When working with items like paintings—objects that are more 2D in nature—the system did not perform as well. This suggests that content creation workflows might benefit from being tailored based on whether the object is inherently 2D or 3D.

7 EXPLORATORY USER STUDY

To explore how Thing2Reality supports communication dynamics, we conducted an IRB-approved study with nine user pairs ($N = 18$). The study examined how the coexistence (DG2) and transformation (DG3) of 2D and 3D artifacts impact user behavior, comprehension, mental effort, and interaction in XR. 9 pairs of participants were recruited for three tasks—avatar decoration, furniture arrangement, and workspace demonstration—that required 3D object use and transition between formats (2D-to-3D, 3D-to-2D). Since the focus is on understanding the communication dynamics and user behaviors (*e.g.*, number of objects created, spatial analysis of natural formations, observations of behaviors) afforded by Thing2Reality, all tasks used digital search as the primary method.

7.1 Participants

We recruited 18 participants (7 female, 11 male) aged between 22-29 ($\bar{x} = 26.1$) through the university mailing list. Most of them ($N = 16$) reported being somewhat or moderately familiar with VR technologies (Median = 2, IQR = 1, on a scale from 1 to 5). Regarding interacting with 2D artifacts in VR, the most common 2D artifacts they interacted with were through web pages ($N = 4$) and online searches in a browser ($N = 3$). The study took around 90 minutes per session, and each participant was compensated with $20 USD.

7.2 Procedure, Apparatus, and Study Setup

To explore different facets of object-centric ideation, we designed three use case scenarios. The social avatar decoration scenario simulated informal "water cooler" interactions, where participants searched for personal items online, converted them to 3D, and used them to decorate their avatars while chatting casually [16]. The collaborative furniture layout scenario focused on how participants reached consensus on spatial arrangements without traditional meeting artifacts. In an open virtual room, they searched for furniture images, transformed them into 3D, and discussed placement while managing conflicting layout preferences [26]. The multi-phase ideation and pitch scenario explored structured remote collaboration, including planning, individual preparation, brainstorming and joint presentation around objects. Participants searched for toy ideas, developed short pitches, and co-presented them using 3D objects and the whiteboard. This setup was informed by prior work on fluid transitions between solo and group work [18, 39, 43, 54].

All sessions were video-recorded, and researchers collected field notes on users' communication dynamics and their use of 2D and 3D content. Figure 6 shows the scene configuration and tasks.

7.2.1 Walkthrough & Avatar Personalization (30 min, Figure 6a-b). Participants followed a brief tutorial, then searched for personal items online, converted them to 3D, and decorated their avatars while conversing casually with their partners.

7.2.2 Collaborative Furniture Layout (10 min, Figure 6c). Pairs searched for furniture images, converted them to 3D, and placed them in the environment, collaborating through discussion to decide on layout and selection.

7.2.3 Multi-Phase Ideation and Pitch with 2D and 3D Objects (20 min, Figure 6d). Participants selected toy ideas via image search, prepared elevator pitches, and delivered a joint 2-minute presentation using 3D objects and the whiteboard for support.

After each task, participants completed a short survey. We followed with semi-structured interviews to assess user experience, perceived benefits, limitations, and observed behaviors.

7.3 Methods

We performed a thematic analysis on interviews and observation data. Two researchers independently coded the data to identify initial patterns, then met to discuss, refine, and consolidate the final themes.

7.4 Results

Overall, participants found the system easy to use and collaborate with. They reported it was easy to communicate with their partners (Median = 5, IQR = 0.5), complete the task using the interface (Median = 5, IQR = 1), and navigate the interface itself (Median = 5, IQR = 1). In terms of communication-specific questions, participants felt confident both in effectively showing the 3D object to their partners (Median = 5, IQR = 1) and in understanding the perspectives and object details their partners were referencing (Median = 5, IQR = 1).

Pairs created between 2 to 4 3D objects per session (Mean = 3.00, SD = 0.87). 2D objects used in final presentations ranged from 2 to 8 (Mean = 3.89, SD = 1.83), and the total number of 2D objects created during sessions ranged from 3 to 9 (Mean = 4.00, SD = 2.12) (see Figure 13 for more details in Appendix). In two sessions, pairs found it necessary to create additional 3D objects midway through their process to better support their presentation visuals, which demonstrates how participants adaptively responded to emerging needs during their collaborative work. Furthermore, we observed participants' communication dynamics, and probed participants' perceived usability, comprehension, and mental efforts for three phases: **(P1)** the phase of their self-cognitive and examination process when orthogonal views and 3D Gaussians were generated from their selected image (as the objects are spontaneously generated rather than pre-prepared); **(P2)** the phase when they use 2D snapshots and 3D Gaussians to communicate with their partners; **(P3)** the phase they use 2D and 3D objects for the 2-min presentation.

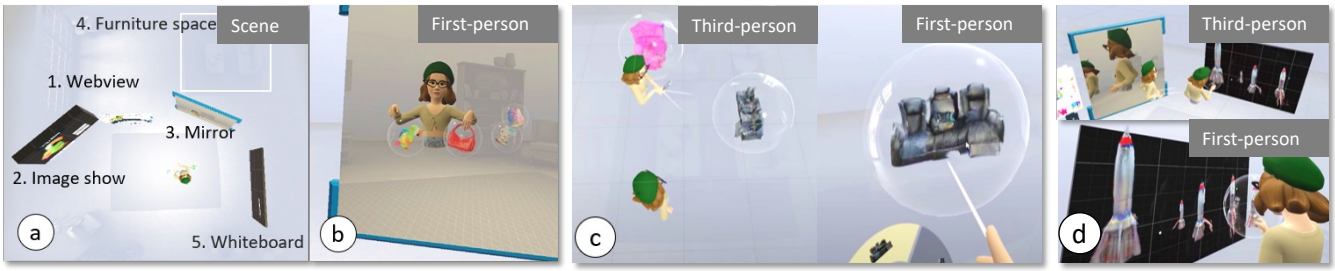

Figure 6: Study Scene and Tasks of Thing2Reality from first-person and third-person views: a) Overview of the immersive environment; b) Warm-up task: avatar decoration; c) Furniture arrangement task; d) Workspace task: Toy pitch.

7.4.1 ***Personal Comprehension, Mental Models, and Ownership of Generated Objects (P1).*** Our analysis revealed contrasting patterns in how users comprehended and claimed ownership of different formats: while 3D objects were strongly preferred for understanding and treated as personal belongings, 2D representations encouraged more shared manipulation and collaborative engagement.

Perceived Benefits of 3D Objects and 2D Menus. Most participants (n = 16) preferred 3D objects over the 2D Pie Menu (n = 2) with orthogonal views. For usage patterns, 3D Objects were used across different phases (P1-P3), while participants mainly looked at 2D orthogonal views (P1) before and after the 3D objects were created, and rarely achieved it during collaboration. When comparing different object representations, the participants generally believed that they could deliver their ideas more clearly with 3D objects (Median = 5, IQR = 1) than the Pie Menu of multiple views (Median = 4, IQR = 2). Participants liked that they can spontaneously turn any images into 3D, which is more helpful than dedicated models, *e.g.*, P10 articulated - "*...using online images was helpful rather than search for dedicated models.*" P18 commented - "*I like its ability to take 2D things on the webpage and convert them to 3D, I think it increases my options to find objects a lot more.*" The 2D Pie Menu, while used less frequently during active collaboration, offered unique advantages during object examination. Its "unfolded effect" provided efficient access to multiple perspectives without requiring extensive manipulation, , *e.g.*, *"see the top, left, right, and bottom makes it easier to retrieve."* Furthermore, P2 highlighted the complementary role, "*2D menu gave me a quick preview of views and took less effort than the 3D manipulation*", making selections easier, *e.g.*, "*Pre-selected views that I select from. Would make working with many items easier and faster.*" (P10), and provided clearer views, *e.g.*, "*the image quality is clearer*" (P1).

When asking about the *mental efforts* required using Pie Menu vs. 3D objects for examining details, participants expressed divided opinions. While most participants found 3D objects requiring less mental effect, a few participants felt that 2D orthogonal views made it easier to understand different perspectives. For example, P10 expressed, "*Easiest is the 2D menu with orthogonal views. Less easy is 2D [snapshots] since I have to select the right view and move it around. Hardest is 3D view since moving changes how the item actually looks.*" In contrast, P17 stated, "*3D object is quite easy to understand, the 2D snapshots and orthogonal views take some effort to use, and I don't think they are as helpful as 3D objects.*" The remaining participants found both options easy to understand, with P4 expressing, "*2D is same as using computer, and 3D is closer to reality.*"

3D Object Ownership and Spatial Boundaries. Most participants demonstrated strong territorial behavior in the shared workspace, treating 3D objects as extensions of personal space. When objects interfered with others' views, participants relied on verbal negotiation rather than direct manipulation. For example, P10 requested "*Can you move your thing?*" instead of moving their partner's object (Figure 8b). This territorial behavior around 3D objects contrasted sharply with 2D snapshot interactions. While participants generally avoided moving others' 3D objects, most of them treated 2D snapshots on the whiteboard as shared territory, collaboratively reorganizing them for decision-making and collective storytelling. We observed two playful exceptions to territorial norms that sparked social bonding and collaboration. In Pair 3 (Figure 7a–b), P5 used her "angry potato" to humorously invade her partner's view, prompting shared laughter and easing interaction. In Pair 7 (Figure 10a–d), P14 initiated a spontaneous object exchange, leading to fluid, co-creative engagement.

7.4.2 ***Collaborative Object Usage Patterns (P2).*** Different collaborative contexts led to distinct patterns in how users leveraged 2D and 3D representations. While spatial arrangement tasks (like furniture placement without whiteboard) relied primarily on 3D manipulation, multi-phase ideation tasks revealed diverse hybrid strategies by allowing the integration of 2D whiteboard.

Our observations revealed three distinct approaches to using 2D and 3D objects during collaborative activities. Two pairs relied solely on 2D objects, reorganizing them to construct narratives. The second approach, observed in three pairs, showed intensive integration of both 2D and 3D objects, switched back and forth between 3D object discussion to whiteboard-mediated communication. The third pattern, seen in four pairs, followed a more sequential workflow, starting with 3D objects for concept discussion and consensus building on the object choices before moving to 2D snapshots for organizing ideas for presentation refinement.

2D Spatial Arrangements Enable Narrative Flexibility. Pairs used 2D-only approaches strategically arranged 2D snapshots to craft and communicate different stories, showcasing the versatility of 2D snapshots as a visual canvas for rapidly building and discussing narratives. For example, Pair 5 developed two distinct presentation approaches (Figure 8d-e) of their two variations of food truck using only 2D snapshots, P10 proposed an interior-exterior

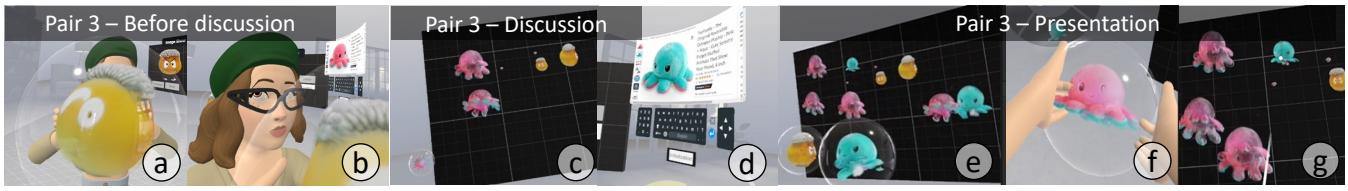

Figure 7: Screenshots of Pair 3: (a-b) P5 used her "angry potato" creation as a tool for playful social interaction, moving the large yellow character extremely close to her partner's view. This unexpected invasion of personal space resulted in shared laughter. (c) Iterative design process: discussing pink octopus on whiteboard and (d) identifying need for a blue variant; (e-g) Integrated use of 2D and 3D octopus models during final presentation.

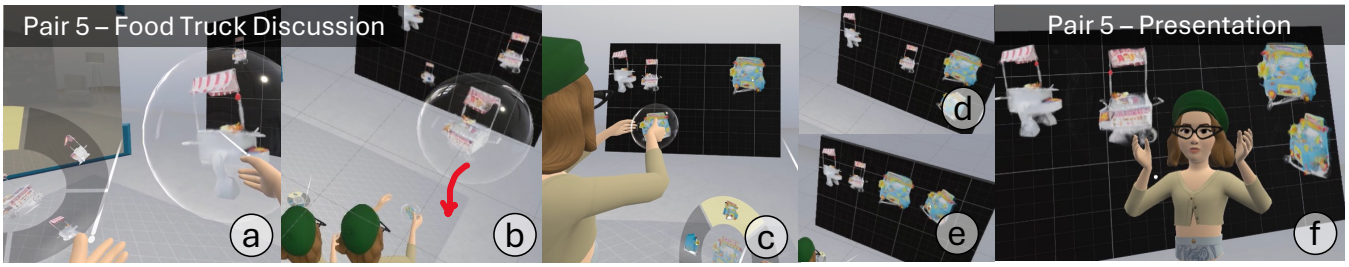

Figure 8: Screenshots of Pair 5: (a) Creating 2D snapshots from 3D models; (b) Object ownership dynamics: P10 requesting "Can you move your thing?"; (c) Capturing snapshots for presentation and discussion of viewpoint choices with P9 explaining back view inclusion: "Just for the vibe! You know you wanna look at every angle!" (d-e) Two contrasting presentation strategies using 2D snapshots: organizing views apart for (d) interior vs. exterior design versus (e) arranging snapshots to showcase product line variety. (f) Presentation.

narrative (Figure 8e) by organizing two variations separately- "*This is the interior of the truck (pink), and this is the exterior of the truck (blue), so we pitch these are different styles, and the kid can ride around this one, and the inside it looks like this.*" P9 instead preferred a product line presentation, where he rearranged the separated 2D snapshots in a line (Figure 8e) to present a chain of food trucks. While these pairs used 3D objects primarily for generating 2D snapshots, a communication challenge emerged: viewers sometimes confused a user's snapshot-taking process (repositioning 3D objects for capture) with attempts to grab attention through object movement.

Hybrid Format Usage Catalyzes Creative Problem-Solving. The hybrid approach enabled creative problem-solving through format integration, as demonstrated by Pair 2's rocket presentation (Figure 9). Their process evolved through several stages, beginning with contrasting discussion styles where P4 used the whiteboard for explaining building blocks while P3 demonstrated his rocket concept through 3D model manipulation. (Figure 9a-b). After selecting the rocket, their discussion and preparation of the presentation structure revealed the complementary use of both formats. When P4 proposed creating a launch animation of the rocket using 2D snapshots, P3 identified a constraint: "*The ladder was there, so it can only be static.*" This challenge led to an effective solution where P4 used the 3D model to demonstrate how rotating to the rocket's back view could hide the ladder (Figure 9c). They looked at the back of the 3D model from the same perspective and reached consensus (Figure 9d). The pair further explored creative presentation techniques, with P4 enlarging the 2D snapshot beyond the whiteboard boundaries to create a dramatic effect, explaining "*See, we can show that it is breaking through the sky!*" (Figure 9e).

Ad-Hoc Format Referencing Bridges Exploration and Refinement. The sequential workflow demonstrated how participants could effectively transition between formats while maintaining focused progression. While primarily sequential, these pairs occasionally made strategic cross-format references when needed. For example, participants effectively complemented 3D object discussions by quickly highlighting specific perspectives with 2D whiteboard. A clear example (Figure 10e-f) emerged in Pair 9's interaction when P17 asked questions about P18's creation, asking "*Is that supposed to be a teddy bear?*" while leaning to examine the 3D object. P18 responded by utilizing the 2D whiteboard, saying "*Yes, but if you look at the side of it, I can show you the image...*" and projected a specific view to clarify the design, showing its imperfections. This seamless switching between 3D and 2D objects emerged as an intuitive communication strategy, allowing participants to direct attention where needed for clearer understanding.

Participants expressed distinct preferences for different formats based on context. Two participants specifically mentioned the desire to hold 3D objects in their hands for better clarification, *e.g.*, P1 shared, *"This was demonstrated more fully when conducting a workspace experiment. When I need to present a toy, holding it in my hand is more sales-oriented"* In contrast, 2D snapshots were seen as

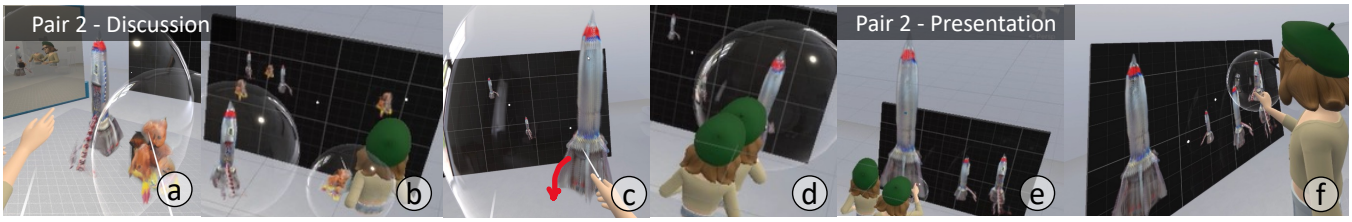

Figure 9: Screenshots of Pair 2: (a) Initial object selection: rocket and building blocks; (b) Contrasting discussion styles: P4 using 2D whiteboard snapshots while P3 demonstrates with 3D model; (c-d) Problem-solving sequence: P4 suggest showing rocket launching animation on the whiteboard, while P3 noting ladder constraint, then discovering solution by rotating 3D model to hide the ladder; (e) Creative scale manipulation showing rocket breaking through whiteboard boundaries; (f) Final presentation combining 3D demonstration and 2D snapshot reference for retractable ladder concept.

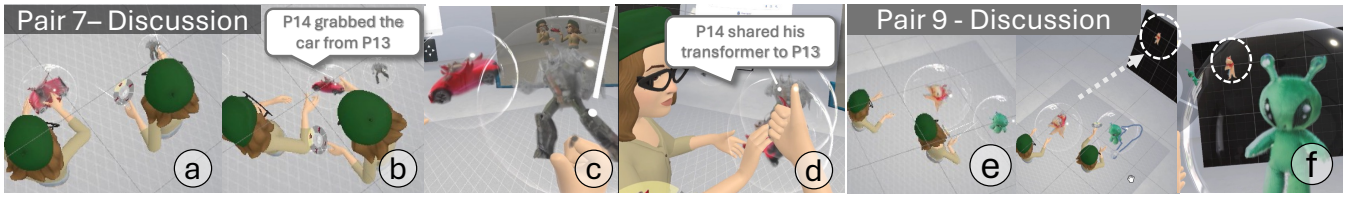

Figure 10: Screenshot of Pair 7: (a-b) P14 grabbed the flying car from P13, saying, "*...these are just, wings or something?*"; (c-d) P14 shared the transformer with P13, noting -"*This one is kind of human-like*" (e-f) For Pair 9, P18 projected one perspective of his 3D object to show the imperfect part.

useful, and a more familiar and usual way for "professional" formats, as they made it easier to explain particular parts of an object without manipulating the 3D view. P6 mentioned, "*it helped us provide the businessperson with multiple views of the object*". The thoughtful consideration given to viewpoint selection became evident in participants' discussions. When P10 questioned the inclusion of the ice-cream truck's back view on the whiteboard, P9 defended the choice with "*Just for the vibe! You know you wanna look at every angle!*" (Figure 8d). In brief, participants found the hybrid and sequential use of 2D and 3D artifacts during the whiteboard task particularly valuable, attributing its advantages to task-specific requirements. Compared to the focused use of 3D objects in the furniture task, the hybrid approach facilitated richer and more dynamic collaboration by effectively blending the strengths of both formats.

7.4.3 ***Presentation Strategies and Format Preferences.*** During the multi-phase task, we observed distinct patterns in how users transitioned between 2D and 3D formats as their presentations evolved. Figure 12 shows the final delivery outcome of the whiteboard presentation. During preparation, 7 out of 9 pairs proactively captured multiple 2D snapshots from their 3D objects to build a comprehensive presentation narrative on the whiteboard. Some pairs went further, using variations of 3D proxy as a creative tool - for instance, generating multiple versions of objects to show transitions (*e.g.*, Pair 3 created happy and angry versions of an octopus toy to show how it will change color from pink to blue, Pair 5 explored food truck variations). This reflects their understanding of how different representations could serve their storytelling needs. Participants also mentioned that the 2D representation on the whiteboard show effectively the variations of these objects.

While participants generally preferred 3D objects for discussion during their pitch preparation, they often used 2D representations in their final pitch delivery. The three pairs who incorporated 3D either to demonstrate dynamic qualities (Pair 1's moving toy car) or to highlight aesthetic appeal (Pair 8's "cute" cat model). The 3 pairs switched between 3D object showcase to referencing 2D artifacts pinned to the whiteboard. For example, Pair 3 seamlessly switched between showing the blue underside of their 3D pink octopus (Figure 7f) and referencing the corresponding 2D whiteboard image (blue octopus) (Figure 7g) to illustrate the character's emotional transition from happy (pink) to angry (blue) (Figure 7f-g), effectively using both formats to tell the story.

When asking about why they used or did not use 3D objects during presentation, participants reported awareness of the different cognitive demands of formats on presenters versus audiences. While some (3/18) noted increased mental effort and attention allocated to transitioning between 2D and 3D elements during presentation, they recognized the value for audience comprehension, *e.g.*, "*when trying to explain a particular part of the object, I think the snapshot made it easier to do since we didn't have to turn the 3D object and show it to the audience's perspective.*" (P2). This resulted in hybrid approaches where presenters would switch between 3D demonstrations and 2D materials, as noted by P2: "*For presentation, 3D objects are better for communicating the actions or interactions with objects,*

while 2D can support a better organization." This highlights the importance of having the flexibility to choose the most appropriate representation based on the specific communication needs.

These patterns reveal that participants don't simply prefer one format over another, but rather develop their own strategies for leveraging each format's strengths across different presentation phases. This validates our **DG3** of enabling flexible transitions between representations, while revealing how users naturally develop best practices for such transitions in professional contexts.

8 DISCUSSION

Through Thing2Reality, we examined both system usability and the broader implications of spontaneous format transformation between 2D and 3D content.

8.1 Spontaneity of 3D Artifacts Creation During XR Meetings

We have investigated the user expectation and interactions of converting objects into 3D from digital vs physical sources. Participants appreciated the flexibility to source objects from online searches and their physical surroundings. While physical object capturing offered realistic references, participants desired more detail and accuracy. Converting digital items into 3D provided a wide variety of items since most items only have one perspective/view point available only, but unpredictable sizing was a challenge, since participants do not have physical copies in their hand. These findings highlight the need for accurate size estimation in digital generation and improved detail retention in physical capture. The combination of 3D Gaussian splatting with multi-view diffusion models enabled new workflows that weren't possible with traditional scanning methods [62, 70]. While sacrificing some accuracy, this approach better supported rapid object-centric design brainstorming and handled challenging capture scenarios (*e.g.*, large objects, occluded views) that prior methods struggle with. This suggests the choice of technical approach should prioritize conversation flow over perfect fidelity. While traditional scanning methods aim for high accuracy, our findings suggest that faster, more flexible approaches better support the spontaneous and evolving nature of collaborative design discussions in XR meetings. Systems should offer variable quality levels matched to different meeting phases - quick generation for initial ideation, higher quality for final presentations. This matches how users transition between informal and formal communication modes during meetings.

8.2 Comparative Use of 2D and 3D Objects for Specific Collaborative Tasks

We have investigated the use of 2D and 3D objects during different stages of communication and discussion process. First, the formats showed complementary strengths in supporting collaboration. 2D snapshots on the whiteboard offered easier viewpoint alignment and consistent perspectives for all users, addressing potential occlusion issues with 3D objects [21]. The 2D Pie Menu with orthogonal views effectively supported spatial understanding. Extending the Pie Menu to include additional views, such as those at 45°intervals for a total of eight perspectives, might potentially enhance spatial understanding but increase cognitive loads.

Second, our findings align with prior work on personal space and territoriality in remote conferencing and co-located spaces with 2D artifacts [30, 71], where we observed shared 3D objects functioning as extensions of personal space for participants who created it. Participants showed clear reluctance to manipulate others' 3D objects, while readily embracing the whiteboard as a shared collaborative space. This behavior might be influenced by *social closeness*, as prior research suggests that sharing personal spaces can strengthen bonds with family and close others [5, 37]. This might indicate the need for providing explicit mechanisms for transitioning objects between personal and shared spaces. We observed that users occasionally misinterpreted 3D object reorientation during 3D-to-2D projection as intentional sharing behavior. While verbal clarification is possible, this highlights the need for designing clearer visual cues to differentiate between intentional sharing and routine 3D object projection in mixed-reality environments.

Third, pairs developed diverse collaboration strategies through format integration. Some only used 2D spatial arrangements for narrative construction, while others combined formats dynamically for creative problem-solving. Several pairs adopted sequential workflows, starting with 3D exploration before transitioning to 2D for presentation refinement. While this flexibility supported various collaboration styles, the mixed use of formats during presentations sometimes increased cognitive load. Our findings extend beyond prior work on 2D artifacts in multi-stage distributed or collocated collaboration [18, 20, 30] by examining how fluid transitions between 2D and 3D formats influence communication dynamics. Our study also demonstrated how users naturally developed diverse strategies for integrating multiple formats based on their collaborative needs, allowing them to tailor artifacts to different contexts and task requirements.

8.3 Limitations and Future Work

8.3.1 ***Scope of User Studies In XR Meetings.*** The two user studies provided valuable insights into Thing2Reality features, focusing on different tasks that involves sharing objects. The lab-based study had limitations, including a small participant sample and a single-day format that may have constrained participants' familiarity with the system, potentially overlooking key challenges and benefits. While a baseline condition (e.g., 2D screen sharing) could offer useful comparisons in terms of both task time and user experience, our current focus is on exploring how participants engage with 2D and 3D objects within these emerging interaction paradigms. As such, we prioritized in-depth, exploratory user experience feedback. Future work can incorporate baseline comparisons to more systematically evaluate how these experiences differ across interaction formats.

8.3.2 ***Limitations in Accuracy and Abstract Visualization.*** Our system faces two primary challenges: accuracy limitations in professional scenarios and difficulties in representing abstract concepts. In contexts where precision is crucial, such as medical diagnosis discussions [73], the current generation method may introduce unacceptable inaccuracies. Similarly, the system struggles to effectively visualize abstract ideas and concepts that lack clear physical

representations. These limitations suggest two key directions for future work. First, improving generation accuracy through automated approaches, such as computer vision for precise object extraction and context-aware suggestion systems based on conversation analysis. Second, developing more sophisticated visualization techniques for abstract concepts, potentially through metaphorical or symbolic representations.

8.3.3 ***Object-Level Versus Scene-Level 3D Gaussian.*** While our system explored 3D object-level interactions, Gaussian splatting can also be used to generate 3D scenes [40]. This can be extended to the exploration of the world of miniatures [8, 74] utilizing the current state of the art. Future work could investigate the scalability of Thing2Reality to handle larger and more complex 3D scenes while maintaining usability and performance.

8.3.4 ***Enhancing Object Fidelity.*** The fidelity of generated 3D Gaussians can be improved by increasing the input density. However, that will result in more time spent on the rendering for the current state [85]. Future work could investigate the integration of these methods with Thing2Reality to improve the visual quality of the generated objects while maintaining interactive performance. Further improvements about the fidelity could come through multiple pathways: integration of point-cloud scanning for enhanced real-world object fidelity [84], physically-based rendering (PBR) for realistic materials and textures [2, 32], and user-guided refinement through 2D snapshots for more controlled generation [91].

9 CONCLUSION

We believe that XR communication has tremendous promise for co-presence and for bridging distances between humans, yet much focus today is on realistic rendering of avatars and remote participants. However, as XR systems mature and become increasingly realistic, it will also become increasingly important to support a similar level of spontaneity with objects and artifacts, as what people experience in real environments. In this paper, we presented Thing2Reality, an XR communication system that allows users to instantly materialize ideas or physical objects and share them as interactive conditioned multiview renderings or 3D Gaussians for realistic 3D rendering. Thing2Reality is one of many necessary building blocks towards increasingly realistic co-presence in XR, and we hope that our work will inspire continued work towards augmented communication in both physical and mirrored world [9].

ACKNOWLEDGMENTS

We would like to thank Chris Ross, Benjamin Hersh, Baosheng Hou for their feedback and discussion on our early-stage proposals. We also thank our reviewers for their insightful feedback. Erzhen Hu was supported by the Google PhD Fellowship. This work was partially supported by research grants from the White Ruffin Byron Center for Real Estate and the Commonwealth Cyber Initiative.

REFERENCES

[1] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: Enhancing Movement Training With an Augmented Reality Mirror. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. ACM, 311–320. https://doi.org/10.1145/2501988.2502045

[2] Raphael Bensadoun, Tom Monnier, Yanir Kleiman, Filippos Kokkinos, Yawar Siddiqui, Mahendra Kariya, Omri Harosh, Roman Shapovalov, Benjamin Graham, Emilien Garreau, et al. 2024. Meta 3d Gen. *ArXiv Preprint ArXiv:2407.02599* (2024). https://doi.org/10.48550/arXiv.2407.02599

[3] Yuri Y Boykov and M-P Jolly. 2001. Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in ND Images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, Vol. 1. 105–112.

[4] 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. https://doi.org/10.1145/332040.332434

[5] Jed R Brubaker, Gina Venolia, and John C Tang. 2012. Focusing on Shared Experiences: Moving Beyond the Camera in Video Communication. In *Proceedings of the Designing Interactive Systems Conference*. 96–105. https://doi.org/10.1145/2317956.2317973

[6] William Buxton. 1992. Telepresence: Integrating Shared Task and Person Spaces. In *Proceedings of Graphics Interface*, Vol. 92. 123–129.

[7] Ming-Chin Chiang and Chih-Fu Wu. 2025. EMPLOYING VARIOUS 2D VISUAL APPEARANCES of 3D OBJECTS to EXPLORE the UNDERSTANDING DIFFERENCES of ORTHOGRAPHIC VIEWS in GRAPHICAL EDUCATION. (2025).

[8] Kurtis Danyluk, Barrett Ens, Bernhard Jenny, and Wesley Willett. 2021. A Design Space Exploration of Worlds in Miniature. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 122, 15 pages. https://doi.org/10.1145/3411764.3445098

[9] Ruofei Du, David Li, and Amitabh Varshney. 2019. Geollery: A Mixed Reality Social Media Platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI, 685)*. ACM, 13. https://doi.org/10.1145/3290605.3300915

[10] Florian Echtler, Vitus Maierhöfer, Nicolai Brodersen Hansen, and Raphael Wimmer. 2023. SurfaceCast: Ubiquitous, Cross-Device Surface Sharing. *Proc. ACM Hum.-Comput. Interact* 7, ISS, Article 439 (nov 2023), 23 pages. https://doi.org/10.1145/3626475

[11] Carmine Elvezio, Mengu Sukan, Ohan Oda, Steven Feiner, and Barbara Tversky. 2017. Remote Collaboration in AR and VR Using Virtual Replicas. In *ACM SIGGRAPH 2017 VR Village*. 1–2. https://doi.org/10.48550/arXiv.2408.02914

[12] Martin Feick, Terrance Mok, Anthony Tang, Lora Oehlberg, and Ehud Sharlin. 2018. Perspective on and Re-Orientation of Physical Proxies in Object-Focused Remote Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, 1–13. https://doi.org/10.1145/3173574.3173855

[13] Andreas Rene Fender, Hrvoje Benko, and Andy Wilson. 2017. MeetAlive: Room-Scale Omni-Directional Display System for Multi-User Content and Control Sharing. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces* (Brighton, United Kingdom) *(ISS '17)*. ACM, 106–115. https://doi.org/10.1145/3132272.3134117

[14] Sean Follmer, Hayes Raffle, Janet Go, Rafael Ballagas, and Hiroshi Ishii. 2010. Video Play: Playful Interactions in Video Conferencing for Long-Distance Families With Young Children. In *Proceedings of the 9th International Conference on Interaction Design and Children* (Barcelona, Spain) *(IDC '10)*. ACM, 49–58. https://doi.org/10.1145/1810543.1810550

[15] Lei Gao, Huidong Bai, Mark Billinghurst, and Robert W Lindeman. 2020. User Behaviour Analysis of Mixed Reality Remote Collaboration With a Hybrid View Interface. In *Proceedings of the 32nd Australian Conference on Human-Computer Interaction*. 629–638.

[16] Carlos Gonzalez Diaz, John Tang, Advait Sarkar, and Sean Rintel. 2022. Making Space for Social Time: Supporting Conversational Transitions Before, During, and After Video Meetings. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–11. https://doi.org/10.1145/3533406.3533417

[17] Jens Emil Grønbæk, Henrik Korsgaard, Marianne Graves Petersen, Morten Henriksen Birk, and Peter Gall Krogh. 2017. Proxemic Transitions: Designing Shape-Changing Furniture for Informal Meetings. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, 7029–7041. https://doi.org/10.1145/3025453.3025487

[18] Jens Emil Grønbæk, Majken Kirkegaard Rasmussen, Kim Halskov, and Marianne Graves Petersen. 2020. KirigamiTable: Designing for Proxemic Transitions With a Shape-Changing Tabletop. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, 1–15. https://doi.org/10.1145/3313831.3376834

[19] Jens Emil Sloth Grønbæk, Juan Sánchez Esquivel, Germán Leiva, Eduardo Velloso, Hans Gellersen, and Ken Pfeuffer. 2024. Blended Whiteboard: Physicality and Reconfigurability in Remote Mixed Reality Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[20] Björn Hartmann, Meredith Ringel Morris, Hrvoje Benko, and Andrew D. Wilson. 2010. Pictionaire: Supporting Collaborative Design Work by Integrating Physical and Digital Artifacts. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. ACM, 421–424.

https://doi.org/10.1145/1718918.1718989

[21] Jörg Hauber, Holger Regenbrecht, Mark Billinghurst, and Andy Cockburn. 2006. Spatiality in Videoconferencing: Trade-Offs Between Efficiency and Social Presence. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*. 413–422. https://doi.org/10.1145/1180875.1180937

[22] Zhenyi He, Ruofei Du, and Ken Perlin. 2020. CollaboVR: A Reconfigurable Framework for Creative Collaboration in Virtual Reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 542–554.

[23] Zhenyi He, Keru Wang, BrandonY. Feng, Ruofei Du, and KenH. Perlin. 2021. GazeChat: Enhancing Virtual Conferences with Gaze-aware 3D Photos. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 769—-782. https://doi.org/10.1145/3472749.3474785

[24] 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

[25] Ken Hinckley, Randy Pausch, Dennis Proffitt, James Patten, and Neal Kassell. 1997. Cooperative Bimanual Action. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 27–34. https://doi.org/10.48550/arXiv.2503.13916

[26] Jon Hindmarsh, Mike Fraser, Christian Heath, Steve Benford, and Chris Greenhalgh. 2000. Object-Focused Interaction in Collaborative Virtual Environments. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7, 4 (2000), 477–509. https://doi.org/10.1145/365058.365088

[27] Christian Holz and Andrew Wilson. 2011. Data Miming: Inferring Spatial Object Descriptions From Human Gesture. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 811–820.

[28] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2023. Lrm: Large Reconstruction Model for Single Image to 3d. *ArXiv Preprint ArXiv:2311.04400* (2023). https://doi.org/10.48550/arXiv.2503.08005

[29] Erzhen Hu, Md Aashikur Rahman Azim, and Seongkook Heo. 2022. FluidMeet: Enabling Frictionless Transitions Between In-Group, Between-Group, and Private Conversations During Virtual Breakout Meetings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, Article 511, 17 pages. https://doi.org/10.1145/3491102.3517558

[30] Erzhen Hu, Jens Emil Sloth Grønbæk, Austin Houck, and Seongkook Heo. 2023. OpenMic: Utilizing Proxemic Metaphors for Conversational Floor Transitions in Multiparty Video Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, Article 793, 17 pages. https://doi.org/10.1145/3544548.3581013

[31] 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)*. ACM, Article 365, 22 pages. https://doi.org/10.1145/3544548.3581148

[32] Xin Huang, Tengfei Wang, Ziwei Liu, and Qing Wang. 2024. Material Anything: Generating Materials for Any 3D Object via Diffusion. *ArXiv Preprint ArXiv:2411.15138* (2024). https://doi.org/10.48550/arXiv.2411.15138

[33] Xincheng Huang, Michael Yin, Ziyi Xia, and Robert Xiao. 2024. VirtualNexus: Enhancing 360-Degree Video AR/VR Collaboration With Environment Cutouts and Virtual Replicas. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. ACM, Article 55, 12 pages. https://doi.org/10.1145/3654777.3676377

[34] 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

[35] Qiao Jin, Ye Yuan, and Svetlana Yarosh. 2023. Socio-Technical Opportunities in Long-Distance Communication Between Siblings With a Large Age Difference. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, Article 94, 15 pages. https://doi.org/10.1145/3544548.3580720

[36] Ricardo Jota, Miguel A Nacenta, Joaquim A Jorge, Sheelagh Carpendale, and Saul Greenberg. 2010. A Comparison of Ray Pointing Techniques for Very Large Displays. In *Graphics Interface*, Vol. 2010. 269–276.

[37] Tejinder K Judge and Carman Neustaedter. 2010. Sharing Conversation and Sharing Life: Video Conferencing in the Home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 655–658.

[38] Sasa Junuzovic, Kori Inkpen, Tom Blank, and Anoop Gupta. 2012. IllumiShare: Sharing Any Surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. ACM, 1919–1928. https://doi.org/10.1145/2207676.2208333

[39] Finn Kensing and Kim Halskov Madsen. 1992. *Generating Visions: Future Workshops and Metaphorical Design*. L. Erlbaum Associates Inc., USA, 155–168. https://doi.org/10.1145/3546155.3546666

[40] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (2023), 1–14. https://doi.org/10.1145/3592433

[41] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *ArXiv:2304.02643* (2023). https://doi.org/10.48550/arXiv.2304.02643

[42] Regis Kopper, Mara G Silva, Ryan Patrick McMahan, and Douglas A Bowman. 2008. *Increasing the Precision of Distant Pointing for Large High-Resolution Displays*. Technical Report. Department of Computer Science, Virginia Polytechnic Institute & State. http://hdl.handle.net/10919/19673

[43] Vijay Kumar. 2012. *101 Design Methods: A Structured Approach for Driving Innovation in Your Organization*. John Wiley & Sons.

[44] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023. Multimodal Foundation Models: From Specialists to General-Purpose Assistants. *ArXiv Preprint ArXiv:2309.10020* 1, 2 (2023), 2. https://doi.org/10.48550/arXiv.2309.10020

[45] Jiannan Li, Maurício Sousa, Chu Li, Jessie Liu, Yan Chen, Ravin Balakrishnan, and Tovi Grossman. 2022. ASTEROIDS: Exploring Swarms of Mini-Telepresence Robots for Physical Skill Demonstration. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, Article 111, 14 pages. https://doi.org/10.1145/3491102.3501927

[46] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. 2023. Instant3D: Fast Text-To-3D With Sparse-View Generation and Large Reconstruction Model. *ArXiv Preprint ArXiv:2311.06214* (2023). https://doi.org/10.48550/arXiv.2311.06214

[47] Christian Licoppe, Paul K. Luff, Christian Heath, Hideaki Kuzuoka, Naomi Yamashita, and Sylvaine Tuncer. 2017. Showing Objects: Holding and Manipulating Artefacts in Video-Mediated Collaborative Settings. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, 5295–5306. https://doi.org/10.1145/3025453.3025848

[48] David Lindlbauer and Andy D. Wilson. 2018. Remixed Reality: Manipulating Space and Time in Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 1–13. https://doi.org/10.1145/3173574.3173703

[49] Xingyu" Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang" Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication With On-The-Fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20. https://doi.org/10.1145/3544548.3581566

[50] Paul Luff, Christian Heath, Hideaki Kuzuoka, Keiichi Yamazaki, and Jun Yamashita. 2006. Handling Documents and Discriminating Objects in Hybrid Spaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. ACM, 561–570. https://doi.org/10.1145/1124772.1124858

[51] Paul Luff, Naomi Yamashita, Hideaki Kuzuoka, and Christian Heath. 2011. Hands on Hitchcock: Embodied Reference to a Moving Scene. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, 43–52. https://doi.org/10.1145/1978942.1978951

[52] Mathias N Lystbæk, Thorbjørn Mikkelsen, Roland Krisztandl, Eric J Gonzalez, Mar Gonzalez-Franco, Hans Gellersen, and Ken Pfeuffer. 2024. Hands-On, Hands-Off: Gaze-Assisted Bimanual 3D Interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–12. https://doi.org/10.1145/3654777.3676331

[53] Jennifer Marlow, Scott Carter, Nathaniel Good, and Jung-Wei Chen. 2016. Beyond Talking Heads: Multimedia Artifact Creation, Use, and Sharing in Distributed Meetings. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1703–1715. https://doi.org/10.1145/2818048.2819958

[54] Bella Martin, Bruce Hanington, and Bruce M Hanington. 2012. Universal Methods of Design: 100 Ways to Research Complex Problems. *Develop Innovative Ideas, and Design Effective Solutions* (2012), 12–13.

[55] Joshua McVeigh-Schultz, Anya Kolesnichenko, and Katherine Isbister. 2019. Shaping Pro-Social Interaction in VR: An Emerging Design Framework. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[56] Terrance Mok and Lora Oehlberg. 2017. Critiquing Physical Prototypes for a Remote Audience. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. ACM, 1295–1307. https://doi.org/10.1145/3064663.3064722

[57] Michael Nebeling, Shwetha Rajaram, Liwei Wu, Yifei Cheng, and Jaylin Herskovitz. 2021. XRStudio: A Virtual Production and Live Streaming System for Immersive Instructional Experiences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, Article 107, 12 pages. https://doi.org/10.1145/3411764.3445323

[58] Nels Numan, Gabriel Brostow, Suhyun Park, Simon Julier, Anthony Steed, and Jessica Van Brummelen. 2025. CoCreatAR: Enhancing Authoring of Outdoor Augmented Reality Experiences Through Asymmetric Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI*

'25). ACM, Article 1232, 22 pages. https://doi.org/10.1145/3706598.3714274
[59] Nels Numan, Shwetha Rajaram, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D Wilson. 2024. SpaceBlender: Creating Context-Rich Collaborative Spaces Through Generative 3D Scene Blending. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. ACM, Article 41, 25 pages. https://doi.org/10.1145/3654777.3676361
[60] Ohan Oda, Carmine Elvezio, Mengu Sukan, Steven Feiner, and Barbara Tversky. 2015. Virtual Replicas for Remote Assistance in Virtual and Augmented Reality. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. ACM, 405–415. https://doi.org/10.1145/2807442.2807497
[61] Jasmin Odenwald, Sven Bertel, and Florian Echtler. 2020. Tabletop Teleporter: Evaluating the Immersiveness of Remote Board Gaming. In *Proceedings of the 9TH ACM International Symposium on Pervasive Displays* (Manchester, United Kingdom) *(PerDis '20)*. ACM, 79–86. https://doi.org/10.1145/3393712.3395337
[62] 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.
[63] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-To-3D Using 2D Diffusion. *ArXiv* (2022). https://doi.org/10.48550/arXiv.2406.04322
[64] Xun Qian, Feitong Tan, Yinda Zhang, BrianMoreno Collins, Alex Olwal, David Kim, Karthik Ramani, and Ruofei Du. 2024. ChatDirector: Enhancing Video Conferencing with Space-Aware Scene Rendering and Speech-Driven Layout Transition. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 16. https://doi.org/10.1145/3613904.3642110
[65] Shwetha Rajaram and Michael Nebeling. 2022. Paper Trail: An Immersive Authoring System for Augmented Reality Instructional Experiences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, Article 382, 16 pages. https://doi.org/10.1145/3491102.3517486
[66] Shwetha Rajaram, Nels Numan, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D Wilson. 2024. BlendScape: Enabling End-User Customization of Video-Conferencing Environments through Generative AI. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 40, 19 pages. https://doi.org/10.1145/3654777.3676326
[67] Julian Rasch, Julia Töws, Teresa Hirzle, Florian Müller, and Martin Schmitz. 2025. CreepyCoCreator? Investigating AI Representation Modes for 3D Object Co-Creation in Virtual Reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, Article 144, 14 pages. https://doi.org/10.1145/3706598.3713720
[68] Jun Rekimoto and Masanori Saitoh. 1999. Augmented Surfaces: A Spatially Continuous Work Space for Hybrid Computing Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. ACM, 378–385. https://doi.org/10.1145/302979.303113
[69] Jun Rekimoto, Brygg Ullmer, and Haruo Oba. 2001. DataTiles: A Modular Platform for Mixed Physical and Graphical Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. ACM, 269–276. https://doi.org/10.1145/365024.365115
[70] Mose Sakashita, Bala Kumaravel, Nicolai Marquardt, and Andrew D. Wilson. 2024. SharedNeRF: Leveraging Photorealistic and View Dependent Rendering for Real-Time and Remote Collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI), 2024*. https://doi.org/10.1145/3544548.3581444
[71] Stacey D Scott, M Sheelagh T Carpendale, and Kori Inkpen. 2004. Territoriality in Collaborative Tabletop Workspaces. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. 294–303.
[72] 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). https://doi.org/10.48550/arXiv.2503.21694
[73] Maurício Sousa, Daniel Mendes, Soraia Paulo, Nuno Matela, Joaquim Jorge, and Daniel Simões Lopes. 2017. VRRRRoom: Virtual Reality for Radiologists in the Reading Room. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, 4057–4062. https://doi.org/10.1145/3025453.3025566
[74] Richard Stoakley, Matthew J. Conway, and Randy Pausch. 1995. Virtual Reality on a WIM: Interactive Worlds in Miniature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 265–272. https://doi.org/10.1145/223904.223938
[75] 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://doi.org/10.48550/arXiv.2402.05054
[76] John C Tang. 1991. Findings From Observational Studies of Collaborative Work. *International Journal of Man-Machine Studies* 34, 2 (1991), 143–160. https://doi.org/10.1016/0020-7373(91)90039-A
[77] Theophilus Teo, Louise Lawrence, Gun A Lee, Mark Billinghurst, and Matt Adcock. 2019. Mixed Reality Remote Collaboration Combining 360 Video and 3d Reconstruction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[78] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. ACM, 161–174. https://doi.org/10.1145/3332165.3347872
[79] Huayuan Tian, Gun A Lee, Huidong Bai, and Mark Billinghurst. 2023. Using Virtual Replicas to Improve Mixed Reality Remote Collaboration. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2785–2795. https://doi.org/10.1109/TVCG.2023.3247113
[80] Maria C Velez, Deborah Silver, and Marilyn Tremaine. 2005. Understanding Visualization Through Spatial Ability Differences. In *VIS 05. IEEE Visualization, 2005.* 511–518. https://doi.org/10.48550/arXiv.2507.05450
[81] Daniel Vogel and Ravin Balakrishnan. 2005. Distant Freehand Pointing and Clicking on Very Large, High Resolution Displays. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*. 33–42.
[82] Vuplex. 2024. Vuplex. 3d Webview: The Ultimate Cross-Platform Web Browser for Unity. https://www.vuplex.com/
[83] Peng Wang and Yichun Shi. 2023. ImageDream: Image-Prompt Multi-View Diffusion for 3D Generation. *ArXiv Preprint ArXiv:2312.02201* (2023). https://doi.org/10.48550/arXiv.2404.17419
[84] Zeyu Wang, Cuong Nguyen, Paul Asente, and Julie Dorsey. 2023. PointShopAR: Supporting Environmental Design Prototyping Using Point Cloud in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, Article 34, 15 pages. https://doi.org/10.1145/3544548.3580776
[85] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. 2024. Grm: Large Gaussian Reconstruction Model for Efficient 3d Reconstruction and Generation. *ArXiv Preprint ArXiv:2403.14621* (2024). https://doi.org/10.48550/arXiv.2403.14621
[86] Ye Yuan, Peter Genatempo, Qiao Jin, and Svetlana Yarosh. 2024. Field Trial of a Tablet-Based AR System for Intergenerational Connections Through Remote Reading. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–28.
[87] 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
[88] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847. https://doi.org/10.1109/CVPR52729.2023.01368
[89] Xiangyu Zhang, Xiaoliang Bai, Shusheng Zhang, Weiping He, Peng Wang, Zhuo Wang, Yuxiang Yan, and Quan Yu. 2022. Real-Time 3D Video-Based MR Remote Collaboration Using Gesture Cues and Virtual Replicas. *The International Journal of Advanced Manufacturing Technology* 121, 11 (2022), 7697–7719. https://doi.org/10.1145/3313831.3376550
[90] Yongxin Zhang, Charlotte Mejlvang Guldbæk, Christian Fog Dalsgaard Jensen, Nicolai Brodersen Hansen, and Florian Echtler. 2024. TableCanvas: Remote Open-Ended Play in Physical-Digital Environments. In *Proceedings of the Eighteenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '24)*. ACM, Article 74, 7 pages. https://doi.org/10.1145/3623509.3635255
[91] Jingyu Zhuang, Di Kang, Yan-Pei Cao, Guanbin Li, Liang Lin, and Ying Shan. 2024. TIP-Editor: An Accurate 3D Editor Following Both Text-Prompts And Image-Prompts. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–12. https://doi.org/10.1145/3658205

A APPENDIX

The top table of Figure 11 breaks down the total system latency into three components: SAM, multiview diffusion models, and Gaussian splatting, averaged across 20 objects with 10 trials each. The bottom panel illustrates the relationship between system performance (FPS/ms) and the number of 3D Gaussians in the virtual environment with two clients. Note that during our studies, none of the sessions generated more than ten 3D objects, which ensured the

performance during the study. We believe the performance and the speed of multi-view diffusion models will become better in the future. Figure 12 and Figure 13 showed the outcome and number and type of generated 2D and 3D objects during the study 2.

Timings (s)	Mean	Sd
System latency	7.51	0.55
SAM runtime	0.28	0.01
Multiview runtime	5.87	0.08
Gaussian splatting runtime	0.31	0.05

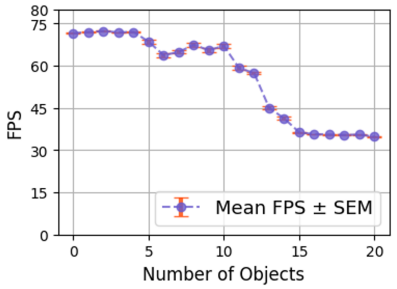

Figure 11: System Performance: (top) shows the overall system latency, and the runtime of SAM, multi-view diffusion models, and Gaussian splatting. Bottom shows the chart of rendering performance as the number of objects ranges from 0 to 20.

Figure 12: Outcomes of Multi-Phase Ideation and Pitch Task with the specific objects that participants eventually chose to present on the board.

Session	# 3D Objects	# 2D Objects Presented	# 2D Objects Created	Objects	Collaboration	Presentation
S1	4	3	3	spider-man, **cars**	3D + 2D	2D + 3D
S2	2	4	9	**rocket**, building blocks	3D + 2D	2D + 3D
S3	3	8	8	**octopus**, potato	2D	2D + 3D
S4	4	4	4	**Pikachu**, octopus	3D → 2D	2D
S5	2	4	6	**food trucks** (fast food, ice cream truck)	2D	2D
S6	4	3	4	**balls**, **cars**	3D → 2D	2D
S7	3	5	7	**transformer**, car	3D → 2D	2D
S8	2	2	3	**cat**, dog	3D → 2D	2D + 3D
S9	2	5	6	teddy bear, **alien**	3D + 2D	2D

Figure 13: Number and Types of Objects During Multi-Phase Ideation and Pitch Sessions. Bold text indicates participants' final choices of objects. During each session, two participants created 2D or 3D objects and presented them on a shared board. The table lists the number of objects in both formats (2D or 3D) that they created, collaborated, and presented.