CollabVisAdapt: Spatio-Temporal Context-Aware Adaptation of Shared Object Visualization for MR Telecollaboration

Xuanyu Wang, Ye Wang, Weizhan Zhang, Shuaichen Guo, Caixia Yan, Shuming Yang, Haipeng Du, Wangdu Chen, and Qi Wang

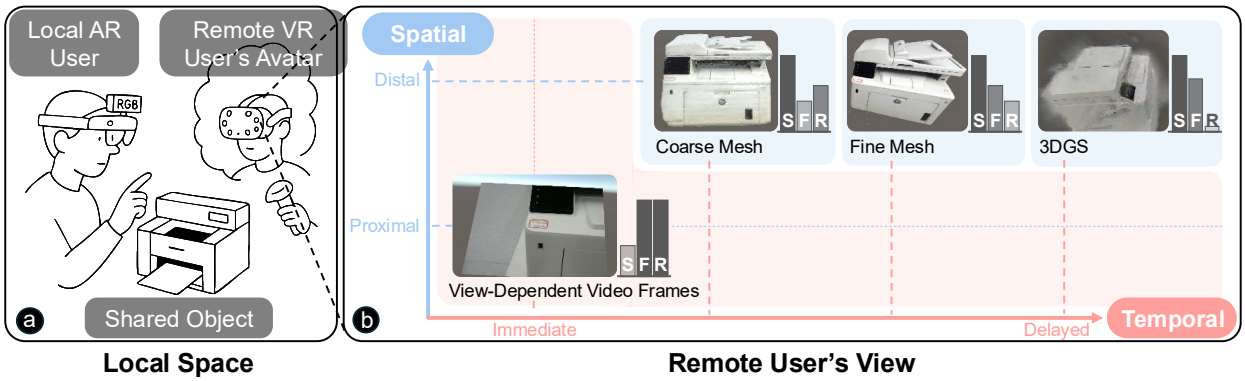

Fig. 1: a A local AR user and a remote VR user are collaborating on a local object. A virtual replica of the object is generated from images captured by the camera on the local user's HMD, and visualized for the remote user in real time. b In the remote user's view, the shared object visualization adapts across four modalities, each prioritizing different aspects of Spatiality (S), Fidelity (F), and Real-time performance (R), based on spatio-temporal contexts. The annotated bars beside each visualization modality indicate its relative priority level in each aspect. View-Dependent-Video-Frames (VDVF) is consistently employed in proximal interactions at any time, and also in distal interactions when the preferred 3D modality is unavailable. Delayed 3D modalities are used in distal interactions when available, shifting from Coarse Mesh, Fine Mesh, to 3DGS regarding generation time.

Abstract—Mixed Reality (MR) telecollaboration aims to enable users to share local objects as real-time synchronized virtual replicas to remote partners and collaborate on physical tasks as if they were co-located. However, in everyday scenarios with mobile and easy-to-setup MR devices, visualizing shared objects in a single modality, ranging from 2D images to 3D reconstruction, struggles to simultaneously optimize all the aspects of Spatiality, Fidelity, and Real-time performance. To overcome this issue, existing methods explore integrating multiple visualization modalities to leverage their respective advantages in subsets of the three aspects. However, they focus on fixed modality combinations without considering user-centered task contexts and workflow, where users may prioritize different aspects of the visualization across task phases. Moreover, they lack support for switching or require manual switching across modalities, which could become disruptive and tiring. In this paper, we propose adapting object visualization based on spatio-temporal contexts in telecollaboration. Specifically, we first couple task type with the user's relative viewing distance as the *spatial context*, and examine its impact on users' prioritized visualization aspects, and the corresponding switching thresholds. With differing generation speeds of modalities, we then explore *temporal* switching schemes when the preferred modality is not immediately available. With the obtained design choices, we implement CollabVisAdapt, a proof-of-concept prototype that supports automatic adaptation of object visualization based on spatio-temporal contexts in MR telecollaboration. A user study in remote maintenance verifies the effectiveness of the proposed workflow with adaptive visualization and the usability of the system.

Index Terms—Mixed reality, telecollaboration, spatio-temporal contexts, adaptive object visualization.

1 Introduction

MR telecollaboration allows remote users to work on an object that is only co-located with one of the users as if they were physically together. With the potential to significantly enhance the immersion and efficiency of remote collaboration, it is increasingly gaining traction in both industry and research in various applications such as education, design, and maintenance.

One of the most critical challenges in telecollaboration is that, unlike face-to-face collaboration, the shared object is not physically available to remote users. It is thus expected to provide them a virtual replica of the object that is 1) freely manipulatable and view-dependent (high Spatiality), 2) of high visual Fidelity, and 3) generated and dynamically updated in real time (high Real-time performance). To simultaneously optimize the visualization in these three aspects, existing methods leverage volumetric capturing and reconstruction via specifically designed pipelines [31,43]. However, they mandate substantial hardware resources and are thus too costly and complex for mobile and everyday scenarios. Targeting these effortless-to-setup scenarios, other solutions employ sparse RGB-D cameras and point cloud-based

- *Corresponding author: Weizhan Zhang.*
- *Xuanyu Wang is with the State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University and the XJTU-POLIMI Joint School. E-mail: xuanyuwang@xjtu.edu.cn.*
- *Ye Wang, Weizhan Zhang, Shuaichen Guo, Caixia Yan, and Haipeng Du are with the School of Computer Science and Technology, MOEKLINNS, Xi'an Jiaotong University. E-mail: {wy1999 | 753693417}@stu.xjtu.edu.cn and {zhangwzh | yancaixia | duhaipeng}@xjtu.edu.cn.*
- *Shuming Yang is with the State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University. E-mail: shuming.yang@mail.xjtu.edu.cn.*
- *Wangdu Chen and Qi Wang are with MIGU Video Co., Ltd. E-mail: chenwangdu@migu.chinamobile.com and 18702100986@139.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

rendering [21, 22, 54]. They offer a certain level of spatiality, but often suffer from rather lower fidelity compared to conventional 2D video-based solutions. The advent of Neural Radiance Fields (NeRF) [39] and 3D Gaussian Splatting (3DGS) [24] supports photorealistic 3D object reconstruction. However, the time-consuming training process, even with significant acceleration [41] and optimization [44, 49], remains prohibitive for real-time high-fidelity reconstruction of dynamic objects, especially with input views being typically sparse and uncontrolled in MR. Furthermore, objects with complex textures and primarily planar surfaces (e.g., screen contents, circuit boards) do not inherently require 3D modalities, which inefficiently consume limited computational resources on mobile MR devices.

While the aforementioned visualization modalities can not simultaneously optimize Spatiality, Fidelity, and Real-time performance (S-F-R) when singly employed in mobile setups, researchers explore mixing multiple modalities to complement each other. They combine mesh and 360° video to balance spatial perception and fidelity [19, 52] or integrate point clouds with NeRF to handle both dynamic and static objects [47]. However, the pre-defined rigid modality-function mappings limit accommodating flexible switching across modalities, as user preferences for modalities could differ regarding collaboration contexts (e.g., task phase and object structure), especially in MR [17, 26], and manual switching could increase cognitive load and disrupt collaboration continuity. Similar autonomous switching techniques exist in 3D User Interface (UI) studies, where researchers exploit adaptation of UI components based on user-centered contexts such as cognitive load [35], head rotation and eye gaze [8, 36, 37], visual angle [63], and response time [18] to improve task efficiency and reduce cognitive load [6, 7]. Informed by this string of research, we explore applying such adaptation to shared object visualization in telecollaboration. Observing that users constantly adjust viewing distance to monitor and interact with the object for task-phase-specific demands, such as detail checking and overview inspection, we couple relative viewing distance and task type as the *spatial context*. This leads to our **RQ1: How does spatial context affect user prioritization across the S-F-R visualization aspects, and how to computationally adapt modalities accordingly?** To answer this question, we conduct our first user study and find the most suitable and preferred modality across varying spatial contexts and their corresponding adaptation thresholds.

In addition to spatial context, we need to further address modality adaptation according to temporal context. The generation of 3D modalities incurs inevitable and variable latency, leading to situations where some modalities, although being the most suitable and preferred in the current spatial context, are not yet accessible. This thus leads to our **RQ2: How to adapt visualization modality with the existence of generation delays to facilitate collaboration?** To answer this question, we conduct the second user study to explore modality switching schemes and find the "coarse-to-fine" strategy to best balance efficiency and modality preference. Note that we focus on investigating visualization modalities that span different regions within the S-F-R design space considering MR telecollaboration in practice, instead of comparing certain representations or specific methods. New techniques could be seamlessly integrated into the adaptation framework regarding their S-F-R characteristics.

With the design distilled from the two studies, we then implement *CollabVisAdapt*, a proof-of-concept prototype to demonstrate MR telecollaboration with adaptive shared object visualization based on spatio-temporal contexts. It consists of a telecollaboration component, a hybrid visualization component, and a modality adaptation component for generating and adapting the visualization. Users can join the session using mobile and commodity MR HMDs. We conduct ablation studies using the prototype to verify the effectiveness of the proposed adaptation and evaluate the system. Results show that our proposed system significantly improves the efficiency of telecollaboration. We also demonstrate the system's practical applicability in broader application scenarios.

The main contributions of this work are threefold.

- We identify the practical necessity of adapting shared object visualization modalities in real-time MR telecollaboration and formulate a spatio-temporal context-aware adaptation paradigm to facilitate telecollaboration workflow.
- Through two user studies, we examine user prioritization in the S-F-R dimensions of visualization, and find the corresponding modality switching thresholds regarding spatial context, and the switching scheme with the existence of generation delays regarding temporal context.
- With the obtained design choices, we build the CollabVisAdapt proof-of-concept prototype system that implements the proposed adaptation techniques, and conduct a comparative study to verify the effectiveness of the system.

2 Related Work

2.1 Object-Centered Telecollaboration

Object-centered remote collaboration is typically asymmetric, in which the remote user has more knowledge about the object and task, while the local user has a more comprehensive view of the workspace and can directly interact with the physical object [14, 56, 67]. Therefore, it is critical to enable the remote user to freely observe and manipulate the virtual shared object with the same fidelity and real-time responsiveness as the local user. Traditional methods utilize video/audio-mediated communication to share linguistic and visual cues between local users and remote collaborators (typically remote experts) [11, 29] for physical tasks. However, it lacks important non-verbal cues, such as depth perception of the physical environment and authenticity of space and objects [1, 30, 33, 40, 57]. To address these problems, immersive telecollaboration systems have emerged with the prevalence of AR/VR/MR HMDs.

Existing methods support collaboration between local users and remote assistants in 3D virtual environments, allowing enhanced communication cues, e.g., annotations, 3D models, gaze, gestures, and text, to be seamlessly integrated into real-world task scenarios [19, 43, 45, 47, 61]. However, these methods often require additional complex and expensive equipment, such as one [47, 61] or multiple [43] RGB-D cameras for capturing and reconstructing 3D virtual environments or objects. This prevents them from being scaled to enter daily work and life, where users require a telecollaboration experience using easy-to-access, lightweight, and portable devices. We focus on telecollaboration using solely MR HMDs in this paper. Other research explores using markers and annotations within a shared task view to enhance remote collaboration performance [1, 12]. They do not focus on the study of object visualization modalities, which is the main focus of this paper.

2.2 Hybrid Object Visualization

To reduce the need for complex capture equipment for 3D model generation, generative methods [2, 34] enable 3D outline creation from a single RGB image. These methods rely on training datasets, showing limited generalization capability to new objects, and require long processing times, necessitating minutes to hours for a single model generation. Optimization solutions like TripoSR [55] and SF3D [4] could reduce static 3D mesh generation time to seconds. To enhance generalization, large-scale 3D models [64] expand object category range by training on massive datasets. Although these generative models can rapidly produce rudimentary 3D shapes of objects, achieving photorealistic detail and preserving the authentic appearance remains a challenge. The models generated by these methods have a very high spatiality, but relatively low fidelity and real-time performance, as shown in Fig. 1.

Against this backdrop, volumetric rendering techniques based on NeRF [39] and 3DGS [24] offer significant advantages, synthesizing photorealistic views from standard RGB camera inputs. However, conventional NeRF and 3DGS require hours of training and dense multi-view data. Optimization methods like Instant-NGP [41], Plenoxels [10], and 3DGS-LM [20] reduce training times and enable rapid reconstruction from sparse views [9, 32]. Despite the improvements, this photorealistic modality remains limited to static scenes, as existing extensions for dynamic scenes [13, 50, 62, 65] still fall short of real-time

training and rendering of volumetric scenes with in-situ real-time captured input images. It has a high spatiality, relatively high fidelity, but poor real-time performance, as shown in Fig. 1.

Existing approaches mentioned above demonstrate an inherent limitation that using a single modality in real-time MR telecollaboration cannot simultaneously optimize the visualization in all S-F-R dimensions. Researchers thus integrate multiple modalities to leverage their respective strengths. Teo et al. [51, 53] combine 360° video and 3D mesh to balance environment scanning speed and spatial cognition. Another group of works focuses on enhancing the reconstruction quality of static objects. Kim et al. [27] improve object boundaries with SAM-guided segmentation. RoomRecon [28] uses generative AI to enhance texture quality. Thing2Reality [17] supports rapid mobile modeling by combining multi-view images with generative meshes. To deal with dynamic scenes, SharedNeRF [47] combines point clouds and NeRF for real-time performance, and VirtualNexus [19] enhances 360° video with mesh for environment sharing and uses NeRF for object copying.

While these approaches explore integrating multiple modalities, they mainly focus on fixed combinations of modalities coupled with specific functions or require manual switching [48]. It is not enough since recent studies underscore the potential necessity of different visualization modalities across platforms and user content preferences in VR [26]. Our study aims to fill this gap by examining how task phases and corresponding user intent and behavior influence user prioritization of the visualization from the S-F-R dimensions and proposing the subsequent automatic modality adaptation.

2.3 Adaptive UI in MR

An increasing number of studies focus on utilizing user behavior and intent for UI adaptation in MR. Firstly, task-driven adaptive UI content adjustments can automatically modify the information density parameters of UI elements, such as text volume and animation complexity, by analyzing user behavior (e.g., operation time and error rate) [18] and physiological indicators such as pupil diameter [35]. This approach enables adaptive granularity adjustment of task guidance content. Secondly, adaptive UI adjustments based on user gaze interaction have been developed to address the Midas Touch problem [23]. By monitoring the user's gaze (head gaze and eye gaze) [8, 36, 37, 46] and visual behaviors such as blinking [7], these systems can distinguish between incidental gaze and intentional interactions. This allows for modifications in the position and visual prominence of application interfaces, such as calendar, weather apps, and video billboards [59]. Finally, existing work adaptively adjusts the Level of Detail (LOD) and physical attributes (such as position and transparency) of UI components based on the relationship between users and spatial semantics. This is achieved by detecting the distance between the UI and objects, the proportion of the Field of View (FoV) they occupy [63], and the positions of the user, UI, and environment [38].

While these works focus on UI component adaptation, we explore user-centered adaptation of object visualization in MR telecollaboration. Informed by the observations of user-UI interaction in these studies, we formulate spatio-temporal contexts that reflect user behavior and interaction intent and investigate the corresponding modality adaptation.

3 Spatial Context-Aware Adaptation

As mentioned previously, we observe that users have different interaction demands, which is also indicated by existing studies, and would continuously adjust viewing distance across task phases. To further explore the impact of such coupled spatial contexts on object visualization, we conduct the first study to investigate user prioritization in the S-F-R visualization dimensions and the corresponding modality switching. The results of this study answer RQ1 and distill design choices for the prototype implementation. Note that although we denote the modalities in this study after their representations, the focus is on examining S-F-R characteristics of the visualization rather than comparing specific representations or methods, as mentioned previously. The study and the subsequent two studies were approved by the institutional review board of the university, and consent from the participant was obtained.

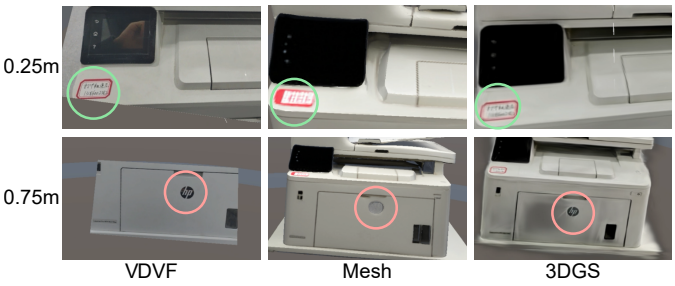

Fig. 2: Observation tasks in Study 1 (0.25m and 0.75m scenarios). The images illustrate the visualizations of the observation tasks using VDVF, Mesh, and 3DGS. In the 0.25m, participants will observe the red label numbers, while in the 0.75m, they will observe the brand logos on the printer's surface. More observation tasks and visualization results can be found in the supplemental file and video.

3.1 Spatial Context and Visualization Modalities

To examine the impact of the spatial context, we need to determine 1) task types and the range of user viewing distance, and 2) the modalities feasible for shared object visualization in real-time MR telecollaboration. We focus on the desktop collaboration scenario involving a single complex shared object in this study.

Targeting detailed inspection and overall observation tasks, we define the range of user viewing distance in the study as $0.25m \sim 1.0m$, considering the Least Distance of Distinct Vision (LDDV) for visual comfort, the typical human arm reach (∼0.58m), and the need to maintain co-presence. We set four evenly distributed points in this range to examine user prioritization in object visualizations, as will be introduced in Sec. 3.3.

Following the discussion of various visualization modalities in Sec. 2.2, we select feasible modalities in this study by examining their relative priority levels in the S-F-R dimensions and the collaboration process in practice. At the beginning of the collaboration, sharing the local user's (LU's) real-time first-person view with the remote user (RU) facilitates the immediate start of the task. To give RU the freedom to control viewing angle, **View-Dependent Video Frames (VDVF)** modality could be employed. It caches video frames captured by LU's camera and the corresponding camera angles, and presents RU the frame with the closest capturing angle to RU's current viewing angle. Similar concepts are also adopted in previous studies [17, 26]. With the real-time updating capability, VDVF could present RU with a dynamic view of the object when observing from the LU's viewing angle registered with LU's head avatar, as shown in the supplemental video. This modality presents a certain level of S since it supports freely adjusting viewing angle, and achieves high F and R, as annotated in Fig. 1. For 3D modalities, input images for reconstruction would be captured with only sparse and limited views due to LU's uncontrolled movement, especially in the early stage. As time progresses, images with more views could become available. Therefore, 3D modalities that can be rapidly generated from sparse views and refined with later-added views are needed in practice. Generative models [4, 64] can reconstruct objects in **Mesh** from a single image in seconds, outperforming other methods such as accelerated NeRF [39, 41] in speed and fidelity under the sparse-view constraint. This modality could be feasible with high S and middle F and R. With more input views, **3DGS** [24] significantly outperforms other methods [3, 10, 41] in photorealism, with high S, relatively high F, and very low R, as annotated in Fig. 1. With the indication in similar studies [17, 26, 60], we select these three modalities mentioned above to represent different regions of the S-F-R design space explored in this study. We provide implementation details in Sec. 3.3.

3.2 Participants

We recruited 16 graduate students (12 males and 4 females; average age: 25.0 (SD = 1.3)) from the local campus. The sample sizes in the studies align with recommendations from relevant research in HCI studies [5]. We recruited young college students as participants in this study and

the subsequent two studies since they are sensitive to MR technology and represent our primary target users in the early stage. It is also a common composition in MR studies [16,17,47]. We mainly focused on the diversity of participants' AR/VR experiences in all studies. Among them, 3 had no prior AR/VR experience, 11 had experienced several times, and 2 were AR/VR application developers.

3.3 Study Scenario and Experiment Setup

We conduct our first study in a remote maintenance scenario, where participants assume the role of a remote VR user and perform inspection tasks. We choose a printer as the target object as it is representative in everyday work and life, and has detailed surface features (e.g., text, buttons, ports) and irregular 3D shapes, supporting both detailed examination and holistic observation task phases.

Informed by similar VR studies [25, 66], we probe the effect of spatial context by setting 4 viewing distances (/m) (0.25, 0.5, 0.75, 1.0), which are maintained by dynamically adjusting the object's position and providing a visual arc trajectory as a guidance cue for movement across different viewing angles. At each distance, users complete proximal (0.25m and 0.5m) detail examination tasks focusing on text and patterns, or distal (0.75m and 1.0m) overall observation tasks focusing on shape and structure, through 3 visualization modalities (**VDVF**, **3DGS**, and **Mesh**). We set the task informed by visual perception principles and LOD theory, which posit that users prefer closer distances for detail inspection and farther distances for shape recognition. In each task, users interact with the object and answer a question about it, as detailed in Sec. 2.2 in the supplemental file.

We implement **VDVF** by capturing 3×360 images of the object covering X-axis (0°-360°) and Y-axis (0°, 45°, 90°) using an RGB camera and a Microsoft HoloLens 2, and presenting corresponding frames based on user viewing angle. Following the reasoning in Sec. 3.1, we implement **3DGS** modality using sparse-view and SfM-free InstantSplat [9] with 12 input views, and render it using a Unity plugin [42]. The model was trained for 1000 iterations, taking 54 seconds. We adopt a state-of-the-art large 3D asset generation model TRELLIS [64] to implement the **Mesh** modality, with 4 input views trained in 21 seconds. We train them respectively on an Nvidia GeForce RTX 4090 GPU. We calibrate the sizes of the visualizations referring to the physical printer through HoloLens 2. Participants use HTC VIVE Pro headsets and controllers in a 2m $\times$ 2m space. We show their dynamic effects in the supplemental video.

3.4 Metrics

We record participants' task completion time for different modalities at each distance and their overall accuracy for all tasks as objective performance metrics. Accuracy is the ratio of correct answers in all tasks. Observation time is measured from the moment the user verbally confirms understanding the task question to the moment they know the answer and press the controller button. For subjective measurements, we collect participants' preferred modalities at each distance, considering the task load and visual comfort. This n-alternative forced-choice data is used to form a psychophysical experiment to detect the distance threshold of the switching between modalities. Then, based on the selected optimal modality at the current distance, we instruct users to freely move and select an optimal observation distance for the current modality and task, and record the optimal distance.

3.5 Procedure

First, we introduce participants to the basic concepts of telecollaboration and familiarize them with the three modalities by presenting example visualizations of a distinct object (to prevent prior exposure). Next, we explain the observation tasks described in Sec. 3.3 and the metrics outlined in Sec. 3.4. Then we walk them through a practice task to get familiar with the controls and experience. After the preparation, we proceeded to the formal experiment.

The experiment adopts a within-subject design, having 12 observation tasks for each user (4 distances $\times$ 3 modalities). Orders of task questions (6 proximal and 6 distal), viewing distance, and modalities were counterbalanced using a Balanced Latin Square. After completing all 12 tasks, each participant took part in a 5-minute semi-structured interview to discuss their experiences, perception of modalities, reasons for their choices, and suggestions on improving the system.

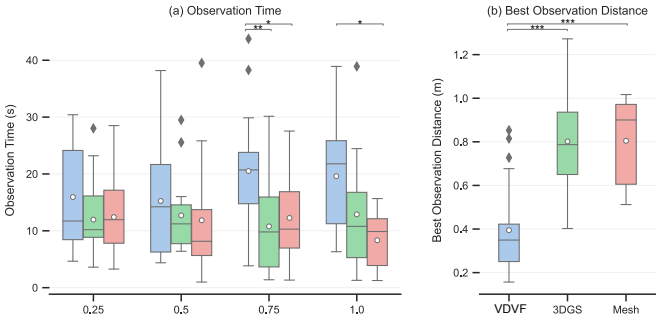

Fig. 3: The comparison of **VDVF**, **3DGS**, and **Mesh** task performance regarding (a) Observation time, (b) Best observation distance in Study 1. **VDVF** had a significantly longer observation time at long distance compared to **3DGS** or **Mesh**, while its best observation distance was significantly shorter than **3DGS** or **Mesh**. We discuss the results more in Sec. 3.7.

We make the following hypotheses:

H1: In the proximal context, VDVF will be significantly more accurate than 3DGS and Mesh due to higher fidelity.

H2: In the proximal context, VDVF will be significantly more efficient due to higher fidelity, while in the distal context, 3DGS and Mesh will be significantly more efficient due to higher spatiality.

H3: The optimal observation distance of VDVF will be significantly shorter than that of Mesh and 3DGS, showing distinct S-F-R prioritization in different spatial contexts.

H4: Users will significantly prefer VDVF in the proximal context and prefer 3DGS in the distal context.

3.6 Results

With Shapiro-Wilk normality tests finding non-normal distributions, we ran non-parametric Friedman Tests and Post-hoc Nemenyi Tests to compare observation time and optimal observation distance, respectively. The same analysis process is applied in Sec. 4 as well. To analyze the results of the psychophysical experiment, we fit psychometric functions to detect the switching threshold between modalities. We will report specific results within the confidence interval of $p < .05$ in the analysis sections. Significant differences between conditions are marked at the top of each figure ($*p < .05$, $**p < .01$, $***p < .001$) in Fig. 3.

Observation Accuracy In the proximal context ((0.25 m, 0.5 m), VDVF (100%, 93.8%) has significantly higher accuracy compared to 3DGS (18.8%, 12.5%) and Mesh (12.5%, 6.2%). In the distal context (0.75 m, 1.0 m), there is no significant difference in accuracy between VDVF (93.8%, 100%) and 3DGS (87.5%, 93.8%), while Mesh has a relatively low accuracy (56.2%, 62.55%). We will discuss the reasons for this difference in Sec. 3.7

Observation Time We present the task observation time for each modality at each distance in Fig. 3(a). Detailed data is provided in the supplemental file. Friedman tests show significant differences in observation time at 0.75m ($\chi^2 = 11.375, p = 0.0034$) and 1.0m ($\chi^2 = 6.500, p = 0.0388$), but not at 0.25 m or 0.50 m. Reasons for this are discussed in Sec. 3.7. Post-hoc Nemenyi tests at 0.75 m reveal significant differences between VDVF and 3DGS ($p = 0.0075$), and VDVF and Mesh ($p = 0.0130$), with no significant difference between Mesh and 3DGS ($p = 0.9829$). At 1.0 m, VDVF differs significantly from Mesh ($p = 0.0356$), but no differences are found between VDVF and 3DGS ($p = 0.1805$) or between 3DGS and Mesh ($p = 0.7593$).

Optimal Observation Distance We present the optimal observation distances for the three modalities in Fig. 3(b). Friedman test reveals a significant difference among them ($\chi^2 = 26.600, p < 0.001$). Post-hoc Nemenyi tests show significant differences between VDVF and 3DGS ($p < 0.001$) and between VDVF and Mesh ($p < 0.001$), while 3DGS

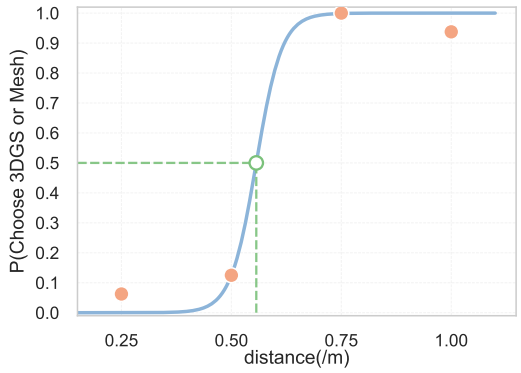

Fig. 4: Fitted psychometric function(line) of the mean estimated threshold values. points represent the probability of users choosing 3DGS or Mesh at different distances, while points indicates the switching point between VDVF and 3DGS/Mesh. We discuss more in Sec. 3.7.

and Mesh have no significant difference ($p = 0.92$). We discuss this lack of significance in Sec. 3.7.

Switching Threshold Based on the abovementioned significance of differences in optimal observation distance between modalities, we fit a psychometric function to determine the switching threshold between VDVF and 3DGS/Mesh, as 3DGS and Mesh are not significantly different in optimal distance. The fitting data is the possibility of choosing 3DGS/Mesh (6.25%, 12.5%, 100%, 93.75%) at each distance $/m$ (0.25, 0.5, 0.75, 1.0), as shown in Fig. 4. We use the standard logistic psychometric function, as shown in Eq. (1).

$$f(x) = \frac{1}{1 + e^{-\beta(x-\alpha)}} \tag{1}$$

From the fitting results, we find the Point of Subjective Equality (PSE) at 0.56m, where users' preference for VDVF and 3DGS/Mesh modalities are equal. Additionally, we use the Bootstrap method to obtain the 95% confidence interval for the threshold α to be (0.52, 0.63).

3.7 Discussion

H1 is supported. In the proximal context, VDVF outperforms 3DGS and Mesh in accuracy. Participants note that "VDVF is simple and clear for close-up observation," while "3DGS and Mesh distort too much, making observation frustrating." This is because VDVF provides high-fidelity ground-truth views ideal for detail observation, whereas 3DGS and Mesh suffer from significant detail loss and fidelity degradation (see Fig. 2). This highlights that the Fidelity dimension is the top priority in the proximal context, where accurate authenticity and detail preservation are critical for task performance and user comprehension. The VDVF modality is currently most feasible in this context with its high F and high R for dynamic objects (further explored in Sec. 4). Future 3D approaches based on 3DGS or novel primitives could also be adopted if they can achieve equally high fidelity with sparse input views, while enabling real-time training and rendering of dynamic content on a mobile MR device solely, which poses the most significant challenge. At longer distances, Mesh accuracy drops due to its insufficient fidelity. Participants reflect that "Mesh lacks realism and conveys little information, making tasks harder". Accuracies of VDVF and 3DGS are similar, suggesting 3DGS can match VDVF in distal tasks. Users also mention that "3DGS has a strong 3D effect at long distances, enhancing the observation experience."

H2 is partially supported. In the proximal context, there is no significant difference in observation time across the three modalities. We consider that it is mainly because users feel unable to get detailed information using 3DGS/Mesh modalities, and finish the task with wrong answers quickly. It is also due to the narrow Field of View (FoV) of VDVF and the frequent movement required in the task. This indicates the necessity for wider FoV capturing, e.g., using a 360° camera, for VDVF implementation. In the distal context, 3DGS and Mesh offer stereoscopic views with high spatiality, requiring less movement with wide FoVs, whereas VDVF requires more head tilting and movement. This is in line with participant reflections: "At long distances, VDVF's limited FoV leads to frustration and lack of global context, while Mesh and 3DGS provide a better sense of structure".

H3 is supported. VDVF's optimal observation distance is significantly shorter than those of 3DGS and Mesh. This significant difference indicates the existence of a modality switching threshold in viewing distance. VDVF's high fidelity and limited FoV make users prefer to use it in proximal contexts, prioritizing fidelity for detail inspection over spatiality. While users still could complete tasks in distal contexts using VDVF, they report a "noticeable flatness", feel it "less like a real object", "provides insufficient information", and "more tiring". The outliers with larger values for VDVF indicate its potential to be also suitable for distal contexts if enhanced in FoV and resolution. There is no significant difference between 3DGS and Mesh in the optimal distance, as users view both as "stereoscopic representations capturing the overall shape", allowing them to comfortably observe the entire object from a greater distance. Users often step back after close-up observations of 3DGS and Mesh, and select a farther distance as their optimal viewing distance. This is also due to the relatively lower fidelity of these two modalities. We thus group 3DGS and Mesh in the psychometric function fitting to determine the switching threshold.

H4 is partially supported. VDVF is significantly more preferred in the proximal context, as shown in Fig. 4 and Fig. 3 (b), with its significantly shorter optimal distance, and the low possibility of users preferring 3DGS and Mesh in the proximal context. However, there is no significant difference in optimal viewing distances between 3DGS and Mesh. Moreover, the respective possibilities of selecting 3DGS and Mesh converge, revealing no significant preference for 3DGS over Mesh. It is also in line with divided user feedback: "I don't choose Mesh because it lacks realism" and "3DGS feels more realistic, like a real object." while others reflect "Floating artifacts around 3DGS interfere with observation, while the Mesh feels more cohesive". These comments indicate that fidelity is also important in distal contexts, and floating artifacts of sparse-view 3DGS and insufficient surface detail of Mesh equally undermine user experience. Although not subjectively more preferred, 3DGS achieves a higher task accuracy than Mesh.

In conclusion, the results show that users prioritize Fidelity in proximal contexts and prioritize Spatiality in distal contexts. Adaptively employing VDVF and 3DGS or Mesh modalities in the corresponding spatial context is necessary to facilitate task efficiency with reduced execution time and error rates. The switching threshold identified through the psychophysical experiment provides a quantitative foundation for telecollaboration workflow with automatic spatial-context-aware adaptation across modalities, guiding the design and implementation of our prototype system in Sec. 5.

4 Temporal Context-Aware Adaptation

Study 1 mainly explores how spatial context influences user prioritization in the S and F dimensions and the corresponding modalities. As mentioned previously, we need to further consider how temporal context affects users' modality choice and workflow, as the preferred modality is not always available due to generation time. Therefore, we conduct the second study to explore the temporal modality switching scheme given inevitable practical generation delays. The results address RQ2 and refine the modality adaptation workflow from the temporal perspective.

4.1 Switching Scheme Design

The results of Study 1 show objectively higher distal task completion accuracy of 3DGS than that of Mesh. However, the generation time of 3DGS [9] is too long (approximately 60 seconds in the implementation introduced in Sec. 3.3, very low in Real-time performance), preventing immediate access during task initiation or following dynamic changes to the object that require visualization updates. The generation of Mesh is relatively faster (around 20 seconds, introduced in Sec. 3.3,

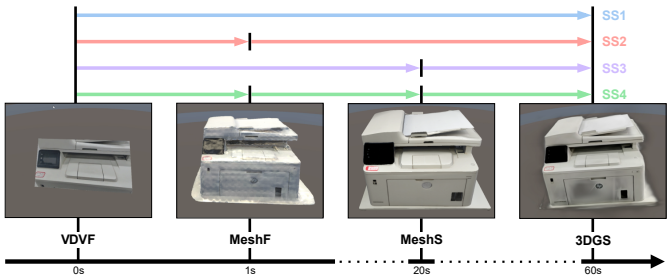

Fig. 5: Switching schemes in Study 2 (SS1, SS2, SS3, SS4). The images illustrate the visualizations of the 4 switching schemes corresponding to their modalities and switching times. Participants will perform 4 distal observation tasks (0.78m) for each switching scheme. More details of the task and visualization results can be found in the supplemental file and video.

relatively low in Real-time performance). Together with the high-real-time-performance VDVF modality, we need to further examine user prioritization in the R dimension and the resulting temporal adaptation scheme.

To flatten the curve in the R dimension and avoid keeping users waiting too long for a 3D modality, we introduce a new modality based on a generative approach [4], which can generate a visualization in mesh from a single view in less than 1s. To distinguish it from the previous Mesh modality, we name the new faster modality "Fast Mesh (**MeshF**)", and name the previous Mesh "Slow Mesh" (**MeshS**). The S-F-R characteristics of these 4 modalities are annotated in Fig. 1. Based on them, we design and explore four temporal modality switching schemes for distal contexts, as the preferred modality in proximal contexts, i.e., VDVF, is immediately available and thus needs no temporal adaptation.

Switching Scheme 1 (SS1): (VDVF - 3DGS) Users engage in distal tasks using VDVF until 3DGS generation is complete and displayed.

Switching Scheme 2 (SS2): (VDVF - MeshF - 3DGS) Users engage in distal tasks using VDVF, then using MeshF after a short time, and finally using 3DGS.

Switching Scheme 3 (SS3): (VDVF - MeshS - 3DGS) Users engage in distal tasks using VDVF, then using MeshS after a relatively long time, and finally using 3DGS.

Switching Scheme 4 (SS4): (VDVF - MeshF - MeshS - 3DGS) Users complete distal tasks using all 4 modalities switched in sequence, regarding their respective generation time.

We design the second study to explore the preferred temporal switching scheme in prolonged multitasking distal contexts. We show their dynamic effects in the supplemental video.

4.2 Study Scenario, Experiment Setup and Procedure

We recruited 16 participants (12 males and 4 females; average age: 25.1 (SD = 1.1)). Among them, 3 had little prior AR/VR experience, 10 had experienced AR/VR several times, and 3 were AR/VR developers. Most (15) of them were from Study 1 and joined this study as a follow-up. We believe there is no significant bias since the independent variables of these two studies do not overlap, and the design considerations focus on clearly distinct dimensions.

To determine the unified and exact generation time of MeshF, MeshS, and 3DGS modalities, which are reported in corresponding papers but on different GPUs, we measure the training times for them ten times using a single NVIDIA RTX 4090 GPU, and take the 95th percentile for each. The results(/s) are as follows: MeshF (M = 0.8, SD = 0.1), MeshS (M = 18.2, SD = 2.3), and 3DGS (M = 56.7, SD = 5.6). We thus set the following generation times: MeshF = 1s, MeshS = 20s, and 3DGS = 60s. These settings ensure that training for all three modalities is completed within the 95% confidence interval.

Study 2 also adopts a within-subject design, and the setup and tasks are the same as in Study 1, with only changes in the scenario. To simulate extended multitasking observation, each switching scheme includes 4 distal observation tasks using the optimal distances for 3DGS (M = 0.787m) and Mesh (M = 0.796m) from Study 1. As the values are close, we use a consistent distance of 0.78m for practicality. Participants press the VR controller trigger upon completing each task to pause time calculation. To differentiate from Study 1, we set 16 new distal observation tasks (detailed in Sec. 3.2 in the supplemental file).

For objective metrics, we record the total observation time and overall error rate for each switching scheme. Observation time is the sum of time from task start to trigger press on the VR controller, while the error rate is the ratio of correctly answered tasks. For subjective metrics, we use the NASA-TLX [15] task load scale (six factors) and collect user preferences among the switching schemes, reflecting their overall subjective experience.

The preparatory procedure and the post-study interview are similar to those in Study 1. The experiment has 16 observation tasks (4 switching schemes $\times$ 4 observation tasks). The orders of the schemes and tasks are counterbalanced hierarchically using Balanced Latin Squares. We make the following hypotheses:

H5: SS1, SS3, SS4 will lead to significant faster task completion than SS2. This is based on observed fidelity limitations of MeshF for certain tasks. **SS3** and **SS4** present users with higher-fidelity MeshS earlier to support task completion. In **SS1**, although having low spatiality and not prioritized for distal tasks, VDVF could remain useful for observation and completing the task earlier due to its sufficient fidelity, while users would be stuck with MeshF until 3DGS is ready in **SS2**.

H6: Participants will prefer SS4 the most due to the seamless update with progressively refined visualization fidelity.

4.3 Results

Observation Time We show statistical comparisons of the total observation times for the 4 switching schemes in Fig. 6(a). Friedman Tests find a significant difference among these schemes ($\chi^2 = 14.55, p = 0.0022$). Post-hoc Nemenyi Test results show that SS2 has significantly higher observation times than SS1, SS3, and SS4 ($p = 0.014, p = 0.013, p = 0.005$), while no significant differences are found among the other combinations.

Error Rate The error rates for the four switching schemes are as follows: SS1 (M = 0.0625, SD = 0.0125), SS2 (M = 0.0469, SD = 0.0102), SS3 (M = 0.1094, SD = 0.0164), and SS4 (M = 0.1250, SD = 0.0167). Friedman Tests indicate that there is no significant difference in error rates across them ($\chi^2 = 4.64, p = 0.2$).

NASA-TLX In Fig. 6(b), we show the NASA-TLX metrics of the 4 switching schemes. Detailed values of the Friedman Test and Post-hoc Nemenyi Test can be found in Sec. 3.3 in the supplemental file. Friedman Tests find significant differences among the 4 schemes in all 6 metrics. Post-hoc Nemenyi Test reveals that SS1 results in significantly higher loads than SS3 and SS4 on almost all metrics (except for the Fr metric, where there is no significant difference between SS1 and SS3). SS2 results in significantly higher loads than SS4 on half of the metrics (TD, Pe, Ef) and than SS3 on TD. There are no significant differences between SS3 and SS4 across all 6 metrics.

Overall Preference Out of all 16 participants, 8 chose SS3 as the optimal representation switching scheme, and another 8 opted for SS4. None of the participants selected SS1 or SS2.

4.4 Discussion

H5 is supported. The observation time for SS2 is significantly higher than that of other switching schemes, indicating that MeshF could only handle a small portion of distal tasks. This limitation stems from MeshF's ability to generate only a 3D front view of an object, lacking surface details on other sides (Fig. 5). User feedback highlights these issues: "MeshF is too coarse, so I have to wait," and "there's too little information, causing long waiting times". Participants manage to collect sufficient information for the task through VDVF in SS1, but with higher task loads. This agrees with Study 1 results and is reflected by users: "I can finish the task using VDVF just with more movements". The observation time range for SS3 and SS4 is broader, as some users complete tasks using VDVF or MeshS. However, task error rates for SS3 and SS4 (M = 10.94%, M = 12.5%) show a slight

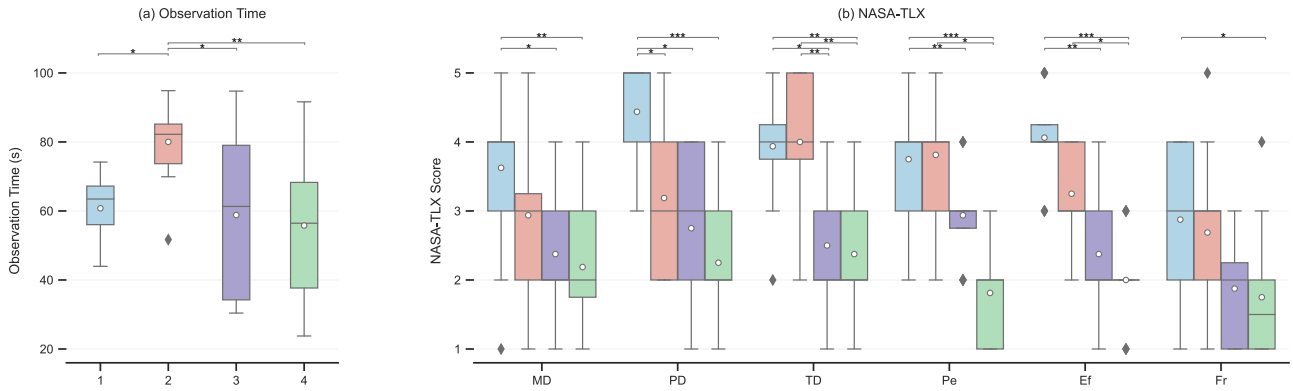

Fig. 6: The comparison of **SS1**, **SS2**, **SS3** and **SS4** task performance regarding (a) Observation time, (b) NASA-TLX scores in Study 2. **SS2** had a significantly longer observation time at long distance compared to **SS1**, **SS3** and **SS4**, NASA-TLX indicates that **SS3** and **SS4** have significantly lower task load than **SS1** and **SS2**. We discuss the results more in Sec. 4.4.

increase compared to SS1 and SS2 (M = 6.25%, M = 4.69%), aligning with findings from Study 1, where surface details in MeshS slightly differed from those in 3DGS. User feedback for SS3 and SS4 includes: "I can quickly complete tasks with MeshS," while others "prefer the longer observation time with 3DGS for more detailed views".

H6 is partially supported. Users prefer SS3 and SS4, each accounting for half of the selections. NASA-TLX results show no significant difference across the 6 metrics between these two schemes. SS4 results in a slightly lower Physical Demand, which is consistent with findings from Study 1 that VDVF in SS3 increases user movement during distal tasks. SS4 also shows slight improvements in Performance and reduced Effort, suggesting that the addition of MeshF helps users complete tasks more smoothly. This is reflected in users' feedback: "I prefer SS4. It transitions from rough to fine details, and the time of waiting feels shorter", and "SS3 and SS4 feel similar. MeshF in SS4 is too coarse to complete tasks and still requires waiting".

Although users do not prefer SS1 and SS2, their significant differences with SS3 and SS4 uncover additional insights. There is a notable difference between SS1 and SS3/SS4 across nearly all NASA-TLX metrics, suggesting that the MeshS used in SS3 and SS4 is crucial for improving performance and reducing task load. Although MeshS takes longer to generate than MeshF, it uses four photos for training, offering more detailed information. This results in MeshS's accuracy being closer to 3DGS, which can handle distal observation tasks in Study 1. The underlying logic of switching schemes is that the modalities' variation in the fidelity dimension should be minimal. Large fluctuations in fidelity (as in SS1 and SS2) tend to increase waiting times, as more perspectives are needed for higher-fidelity modalities, leading to longer training durations and reduced overall user experience.

In conclusion, SS4 (VDVF->MeshF->MeshS->3DGS) is the optimal temporal modality switching scheme. It balances the F and R dimensions, allowing for gradual refinement in the S and F dimensions. It makes the overall workflow more efficient, smoother, and more reasonable.

5 Implementation

In Study 1 (Sec. 3), we determine the preferred modality in different spatial contexts and the corresponding spatial switching threshold. In Study 2 (Sec. 4), we identify the optimal temporal switching scheme with the inevitable generation delay of 3D modalities. Building on these, we implement *CollabVisAdapt*, a proof-of-concept telecollaboration system that adaptively adjusts object visualization modalities based on spatio-temporal contexts. *CollabVisAdapt* comprises three core components: the telecollaboration component, the hybrid visualization component, and the modality adaptation component. We implement the system in Unity 2022.3 on Microsoft HoloLens 2 (OST AR HMD) and HTC VIVE Pro (VR HMD). It supports desktop-level collaboration for a local AR user (Intel i7-11800H, RTX 3060 Laptop GPU) and a remote VR user (Intel i7-13700KF, RTX 4080 GPU). A webcam is mounted on the AR user's HMD to capture the real-time first-person view. The VR user views the adaptive visualization of the object to assist the AR user, offering an intuitive and seamless collaboration workflow. We introduce the three components below.

5.1 Telecollaboration Component

The telecollaboration component supports basic user avatar interactions, collaborative object position framing, and local AR user hand collision detection. In the avatar interaction module, both AR and VR platforms feature 3D models of the other user's head and hands. They exchange headset position and orientation data and hand model information via UDP. The collaborative object framing module uses a 3D wireframe model (bounding box) to define object position, rotation, and scale. This data is transmitted through UDP to the corresponding model in the remote VR user's view. For hand collision detection, we utilize the NearInteractionModeDetector from MRTK to detect hand collisions with the BoundingBox, with a time threshold where only collisions lasting longer than 2 seconds are considered valid to prevent frequent detections from accidental touches. The peer-to-peer networking framework via UDP connections enables the system's scalability to scenarios with multiple AR and VR users.

5.2 Hybrid Visualization Component

The hybrid visualization component implements the VDVF, MeshF, MeshS, and 3DGS modalities:

VDVF: Synchronized with the AR user's camera feed, the system calculates θ_1 ($0^\circ \sim 360^\circ$, the angle between the AR HMD and the bounding box on the XZ plane) and θ_2 ($0^\circ \sim 90^\circ$, the angle on the Y-axis), with θ_2 discretized into three segments. The images and angles (θ_1, θ_2) are sent from the AR end to the VR end via UDP and stored in a 360×3 array on the VR end. The VR end presents the corresponding video frame with the camera angle closest to the VR user's current viewing angle. When the AR user returns to a previous capturing angle (θ_1, θ_2), the cached frame is updated with the current video frame. The continuous refreshing frames when the VR user's viewing angle matches the AR user's present a real-time dynamic view. After the AR user finishes interaction in a certain area, the VR end clears the frames from the corresponding hemispherical area to ensure the latest status of the visualization, as the area captured previously has been modified. This spatio-temporal cache-and-update mechanism of VDVF ensures the system's real-time usability in dynamic telecollaboration.

MeshF, MeshS, and 3DGS: We use the models from Studies 1 and 2 to generate MeshF [4], MeshS [64], and 3DGS [9] modalities. They are all trained on a server with an NVIDIA RTX 4090 GPU. The VR system uploads VDVF and downloads mesh (.glb) and 3DGS (.ply) files over a 500 Mbps HTTP connection, with transfer times of $\sim 0.08s$ for a 5MB .glb and $\sim 3.2s$ for a 200MB .ply model file. MeshF training begins when a usable image with $\theta_2 = 45^\circ$ and θ_1 in $(-10^\circ \sim 10^\circ)$ is captured during a new user action. MeshS supports both single- and

multi-view training: the single-view setup shares MeshF's conditions, while the multi-view setup requires four distinct images. 3DGS training starts with at least three usable images with θ_2 in $(-20^\circ \sim 20^\circ)$, and up to 12 images are selected based on angular diversity.

5.3 Modality Adaptation Component

The adaptation component switches the visualization modality based on the spatial and temporal contexts. Study 1 indicates that the switching threshold between VDVF and Mesh or 3DGS is 0.56m, measured from the VR headset to the bounding box surface. Based on Study 2, switching scheme 4 (VDVF-MeshF-MeshS-3DGS) is selected for distal contexts. To ensure real-time modality generation, any ongoing training is halted when the user moves to the next action, and a new training session begins. This allows users in action-intensive tasks to quickly obtain a rough outline using MeshF (generation time $M = 0.8s$) or MeshS (generation time $M = 18.2s$), while those in observation-intensive tasks can use 3DGS (generation time $M = 56.7s$) for detailed observation.

6 Evaluation

Based on the *CollabVisAdapt* prototype, we conduct a within-subject empirical user study to evaluate the effectiveness and usability of our adaptive visualization mechanism by comparing it with the workflow using manual switching, representing the condition with the ablation of the modality adaptation component. The comparison of various modalities and the comparison between adaptive visualization and the single fixed-modality method have been implicitly evaluated in Studies 1 and 2, and thus will not be reiterated in this section.

6.1 Participants

We recruited 12 graduate students (10 males and 2 females; average age: 25.1 (SD = 1.0)) from the local campus as participants. Among them, 3 had no prior AR/VR experience, 7 had experienced AR/VR several times, and 2 were AR/VR developers. We invited 4 participants of Study 1 to join the evaluation. We believe this involves no significant bias since the number of reused users is small and the conditions in Study 1 are pre-recorded, while the evaluation in this section is based on a real-time running system.

6.2 Study Conditions

For the evaluation, we have two conditions. Adaptive visualization is the full system implementation and includes spatio-temporal context-aware modality adaptation. Manual switching is the condition with the ablation of the modality adaptation component. Users can manually switch the modality as they wish, while the other components remain unchanged.

6.3 Study Scenario and Experiment Setup

We focus on evaluating how the adaptive object visualization for the remote VR user facilitates telecollaboration by setting a common expert-novice remote support scenario following previous research [47]. In this scenario, the VR user is an expert leading the task, while the local AR user follows the VR user's instructions to complete physical operations. We thus assign participants the role of the VR expert for its dominance in the task and direct engagement with the adaptive visualization, and let one researcher play the role of the local AR user, following previous practice [47]. They move and communicate naturally to complete tasks as in real-world scenarios without scripted routes. To maintain consistency in 3D representations and minimize errors, we select the printer from Studies 1 and 2 as the task object. The remote VR expert (participant) guides the local AR user to repair the printer from various malfunctions. We set four common printer malfunctions simulating varying task intensities (details are illustrated in Tab. 8 in the supplemental file), and teach participants how to fix them before the experiment. A third party sets up malfunctions to prevent prior knowledge interference. The AR user (researcher) first calibrates the AR bounding box of the object (as shown in the supplemental video) as a preparation stage similar to real-world collaboration. A preliminary visualization of the printer in its initial state in each modality is also generated from an approximate 360° orbit shot of it acquired in-situ by the AR user in this preparation stage. The modalities are continuously maintained and updated during the collaboration process (as described in Sec. 5.3). A trial ends when the printer is fixed and outputs the expected content. The orders of conditions and malfunctions are balanced using a Balanced Latin Square.

For objective metrics, we record the completion time for each condition, from the time the AR user finishes calibrating the bounding box to the time the VR (participant) sees the final printed content. For subjective metrics, we use 1) NASA-TLX task load metrics and 2) three metrics related to the experience of AR content adaptation from previous work [63]: Easy-to-observe, Task communication load, and Task smoothness. Easy-to-observe measures how quickly participants can view desired parts. Task communication load assesses the communication effort with the AR user. Task smoothness assesses the experience flow regarding continuity and interruptions. These 9 metrics are rated on a five-point Likert scale (1 to 5), with lower scores generally indicating better performance, except for Easy-to-observe and Task smoothness, where higher scores are better.

In the experiment, participants complete conditions 1 and 2 separately, rate task load and collaboration experience, and provide their preferences (1 for adaptive visualization, 5 for manual switching) at the end.

6.4 Results

Following previous approaches [26], we first conducted Shapiro-Wilk Normality Tests and found non-normal distributions. We then used non-parametric Wilcoxon Signed-Rank Tests to evaluate the significance of the two conditions across various metrics.

Completion Time The data indicate that the completion time (/s) for adaptive visualization (M=289.2, SD=20.25) is significantly lower than that for manual switching (M=327.83, SD=44.14), with a p-value of 0.0068.

NASA-TLX From the results shown in Fig. 7(a), we can see that adaptive visualization has significantly lower scores than manual switching in all metrics.

Telecollaboration Experience From the results shown in Fig. 7(b), we can see that there are significant differences between adaptive visualization and manual switching across three metrics. Adaptive visualization has significantly higher scores than manual switching in the Easy-to-observe and Task smoothness metrics, and significantly lower scores in the Task communication load metric.

6.5 Discussion

The results show that adaptive switching outperforms manual switching in remote collaboration tasks. Adaptive switching significantly reduces completion time and aligns with users' observation behavior and intentions. This is in line with user reflections: "Adaptive visualization adapts to my viewing distance, which is reasonable and intuitive." and "Manual switching is distracting, requiring frequent adjustments." This is further supported by NASA-TLX MD, TD, and Easy-to-observe metrics.

Adaptive switching significantly reduces users' PD compared to manual switching, allowing easier object viewing with fewer operations. Users reflect "Adaptive switching reduces button clicks, simplifying operations." It also improves task smoothness, as users comment, "I can focus on the task without distractions from switching modalities myself." Users find Adaptive switching intuitive and helpful: "When I supervise the AR user operating in front of the printer through VDVF, I can simultaneously go to the other side to inspect labels. When I walk to change my view, it switches to Mesh/3DGS modality, which is very helpful as a spatial reference for me to find the angle, and the automatic switching back to VDVF for detailed examination was well-timed and appropriate. The whole process is very spatial-faithful and intuitive for me." Additionally, Adaptive visualization lowers communication load, with users remarking "Manual switching distracts me, so I subconsciously just stay in the video mode (VDVF) and talk more to the AR user to make adjustments." and "Manual switching requires frequent communication with the AR user, making the task harder." In

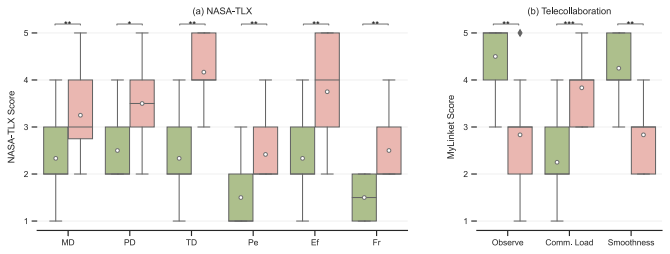

Fig. 7: The comparison of **condition 1 (adaptive visualization)**, and **condition 2 (manual switching)** task performance regarding (a) NASA-TLX scores, (b) Telecollaboration experience scores in Evaluation. NASA-TLX indicates that **adaptive visualization** has significantly lower task load than **manual switching**. The telecollaboration experience scores indicate that **adaptive visualization** provides a significantly better user experience compared to **manual switching**. We discuss the results more in Sec. 6.5.

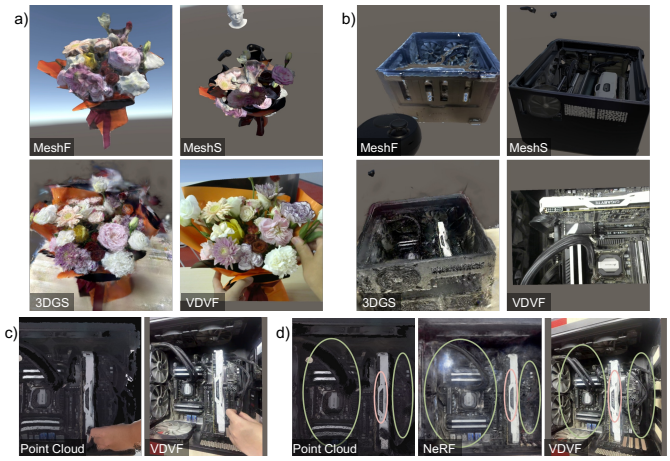

Fig. 8: Demonstrations of the system in (a) collaborative flower arranging and (b) computer assembly scenarios, and preliminary comparisons with the effect of existing method [47] using point cloud and NeRF in (c) dynamic and (d) static scenarios. In (a) and (b), 3D modalities provide decent spatial interactivity. For areas with deep recesses, such as inside the computer case, the effects of 3D modalities are unsatisfactory, but VDVF effectively compensates information from desired views. For dynamic parts, VDVF enables real-time observation of flower moving and computer component assembling. (c) When wiring inside the computer case, point cloud exhibits grainy and blurry textures on the board and hands, hindering the supervision of dynamic wiring details, while real-time VDVF presents the process clearly. (d) When inspecting the case, both point cloud and NeRF are not high-fidelity enough for showing intricate structures, such as wires and connectors (green circles) and texts (pink circles), while cached VDVF clearly displays them.

most cases, participants could directly obtain information from desired views, facilitated by the combination of spatially cached VDVF and the real-time updated VDVF bound to the AR user, as illustrated above. In a few situations where cached VDVF is insufficient for observing occluded details, asking the AR user to adjust to desired capturing angle for better real-time VDVF would efficiently solve the problem. This also indicates the flexibility and scalability of VDVF to more complex areas with detailed textures that needs viewing from challenging angles, e.g., the computer assembly scenario in the subsequent demonstration.

In conclusion, compared to Manual switching, Adaptive visualization enables smoother completion of telecollaboration tasks with less effort, and leads to reduced completion time and communication costs. It syncs well with users' viewing intentions and motions, intuitively providing the most contextually appropriate modality.

To show the practical applicability to broader everyday telecollaboration scenarios with various physical tasks, we demonstrate the system in a collaborative flower arranging scenario and a computer assembly scenario, as shown in Fig. 8 (a) and (b) and the supplemental video. The adaptive visualization provides intuitive on-demand and implicit switching among S-F-R priorities. VDVF effectively presents dynamic flower moving and component assembling, and supports internal inspection of intricate wiring patterns within an opened computer case. We also present preliminary comparisons with the effect of existing approach [47] using point cloud (implemented based on a Kinect depth camera) and NeRF [41] in computer assembly in Fig. 8 (c) and (d). Point cloud and NeRF both exhibit insufficient fidelity in dynamic and static proximal interactions relying on observing complex details.

7 Conclusion, Limitations and Future Work

In conclusion, we have explored the adaptation of shared object visualization based on spatio-temporal contexts in MR telecollaboration with a mobile and easy-to-access setup. We have identified the spatial and temporal contexts and studied their respective influence on the choice of object visualization modality in real-time telecollaboration. In terms of the spatial context, we have explored the user-preferred modality regarding the viewing distance and the type of task, and obtained the corresponding distance thresholds for switching between them through the first user study. In terms of the temporal context, we have conducted the second user study and derived the optimal modality switching strategy when there exists a delay before the preferred modality is generated. Based on the study results, we developed *CollabVisAdapt*, a proof-of-concept prototype that implements the spatio-temporal context-aware adaptation of shared object visualization to facilitate object-centered MR telecollaboration workflow. It empowers an efficient telecollaboration experience with an MR HMD for each user via a telecollaboration component, a hybrid representation component, and a modality adaptation component. The evaluation has validated the effectiveness of the proposed adaptation compared to manual switching and the usability of the system. Next, we address the limitations of our work and discuss the corresponding potential future work.

Generation time and automatic calibration. We employ current state-of-the-art models for 3D reconstruction in this work, but the generation speed remains limited. In future work, we could explore accelerating reconstruction model training and improving computational efficiency for telecollaboration scenarios. Moreover, the current system requires the AR user to manually calibrate the visualization and bounding box. Future work could explore automatic calibration through image corresponding or directly aligning generated 3D models.

Large-scale and multiple objects. We focus on desktop-scale collaboration on an individual object in this work. We plan to explore tasks involving larger objects and space, such as room-scale or outdoor collaborations in large equipment maintenance and scene arrangement scenarios. We will also explore multi-object scenarios with workflows switching among objects and transitions in the space [58], and further investigate the effect of spatio-temporal contexts in these scenarios.

Multi-user scenarios and dynamic tasks. We currently only conduct studies in single-user or minimal two-user scenarios. We are interested in exploring the dynamics of multi-user scenarios with more diverse interactions and collaboration contexts. Furthermore, studies with participants playing both roles of local AR and remote VR users with varying levels of expertise in the task and diverse backgrounds could enhance the understanding of such telecollaboration systems.

Acknowledgments

We thank the anonymous reviewers for their constructive comments and the interview and user study participants for their time. This work was supported by grants from the National Natural Science Foundation of China (No. 62502375, 62192781, 62172326, 62137002, and 62302384), in part by the Key Research and Development Project in Shaanxi Province under Grant 2022GXLH-01-03, and by the Project of China Knowledge Centre for Engineering Science and Technology.

REFERENCES

[1] D. Anton, G. Kurillo, and R. Bajcsy. User experience and interaction performance in 2d/3d telecollaboration. *Future Generation Computer Systems*, 82:77–88, 2018. 2

[2] Y. Bai, X. Wang, Y.-p. Cao, Y. Ge, C. Yuan, and Y. Shan. DreamDiffusion: Generating high-quality images from brain EEG signals. doi: 10.48550/arXiv.2306.16934 2

[3] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022. 3

[4] M. Boss, Z. Huang, A. Vasishta, and V. Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint arXiv:2408.00653*, 2024. 2, 3, 6, 7

[5] K. Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 981–992, 2016. 3

[6] G. Daskalogrigorakis, A. McNamara, and K. Mania. Holo-box: Level-of-detail glanceable interfaces for augmented reality. In *ACM SIGGRAPH 2021 Posters*, pp. 1–2. Association for Computing Machinery, 2021. 2

[7] G. Daskalogrigorakis, A. McNamara, A. Marinakis, A. Antoniadis, and K. Mania. Glance-box: Multi-lod glanceable interfaces for machine shop guidance in augmented reality using blink and hand interaction. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 315–321. IEEE, 2022. 2, 3

[8] S. Davari, F. Lu, and D. A. Bowman. Validating the benefits of glanceable and context-aware augmented reality for everyday information access tasks. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 436–444. IEEE, 2022. 2, 3

[9] Z. Fan, K. Wen, W. Cong, K. Wang, J. Zhang, X. Ding, D. Xu, B. Ivanovic, M. Pavone, G. Pavlakos, et al. Instantsplat: Sparse-view sfm-free gaussian splatting in seconds. *arXiv preprint arXiv:2403.20309*, 2024. 2, 4, 5, 7

[10] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5501–5510, 2022. 2, 3

[11] S. R. Fussell, R. E. Kraut, and J. Siegel. Coordination of communication: effects of shared visual context on collaborative work. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, CSCW '00, pp. 21–30. Association for Computing Machinery, 2000. doi: 10.1145/358916.358947 2

[12] S. R. Fussell, L. D. Setlock, E. M. Parker, and J. Yang. Assessing the value of a cursor pointing device for remote collaboration on physical tasks. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, pp. 788–789, 2003. 2

[13] C. Gao, A. Saraf, J. Kopf, and J.-B. Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5712–5721, 2021. 2

[14] P. Gurevich, J. Lanir, and B. Cohen. Design and implementation of teleadvisor: a projection-based augmented reality system for remote collaboration. *Computer Supported Cooperative Work (CSCW)*, 24(6):527–562, 2015. 2

[15] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988. 6

[16] E. Hu, J. E. S. Grønbæk, W. Ying, R. Du, and S. Heo. 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*, pp. 1–22, 2023. 4

[17] E. Hu, M. Li, J. Hong, X. Qian, A. Olwal, D. Kim, S. Heo, and R. Du. Thing2reality: Transforming 2d content into conditioned multiviews and 3d gaussian objects for XR communication. doi: 10.48550/arXiv.2410.07119 2, 3, 4

[18] G. Huang, X. Qian, T. Wang, F. Patel, M. Sreeram, Y. Cao, K. Ramani, and A. J. Quinn. Adaptutar: An adaptive tutoring system for machine tasks in augmented reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021. 2, 3

[19] X. Huang, M. Yin, Z. Xia, and R. Xiao. 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*, UIST '24, article no. 55, 12 pages. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3654777.3676377 2, 3

[20] L. Höllein, A. Božič, M. Zollhöfer, and M. Nießner. 3dgs-LM: Faster gaussian-splatting optimization with levenberg-marquardt. doi: 10.48550/arXiv.2409.12892 2

[21] A. Irlitti, M. Latifoglu, T. Hoang, B. V. Syiem, and F. Vetere. Volumetric hybrid workspaces: Interactions with objects in remote and co-located telepresence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2024. 2

[22] A. Irlitti, M. Latifoglu, Q. Zhou, M. N. Reinoso, T. Hoang, E. Velloso, and F. Vetere. Volumetric mixed reality telepresence for real-time cross modality collaboration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2023. 2

[23] R. J. Jacob. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 11–18, 1990. 3

[24] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3

[25] D. Kim, J.-e. Shin, J. Lee, and W. Woo. Adjusting relative translation gains according to space size in redirected walking for mixed reality mutual space generation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 653–660. IEEE, 2021. 4

[26] H. Kim and I.-K. Lee. Is 3dgs useful?: Comparing the effectiveness of recent reconstruction methods in vr. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 71–80. IEEE, 2024. 2, 3, 8

[27] J. Kim and J. Lim. Integrating meshes and 3d gaussians for indoor scene reconstruction with SAM mask guidance. doi: 10.48550/arXiv.2407.16173 3

[28] S. J. Kim, D. D. Cao, F. Spinola, S. J. Lee, and K. S. Cho. Roomrecon: High-quality textured room layout reconstruction on mobile devices. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 544–553. IEEE, 2024. 3

[29] R. E. Kraut, S. R. Fussell, and J. Siegel. Visual information as a conversational resource in collaborative physical tasks. *Human–computer interaction*, 18(1-2):13–49, 2003. 2

[30] G. Kurillo and R. Bajcsy. 3d teleimmersion for collaboration and interaction of geographically distributed users. *Virtual Reality*, 17(1):29–43, 2013. 2

[31] J. Lawrence, D. B. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes, E. M. Gomez, S. Häberling, H. Hoppe, A. Huibers, C. Knaus, B. Kuschak, R. Martin-Brualla, H. Nover, A. I. Russell, S. M. Seitz, and K. Tong. Project starline: A high-fidelity telepresence system. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40(6), 2021. 1

[32] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20775–20785, 2024. 2

[33] C. Licoppe, P. K. Luff, C. Heath, H. Kuzuoka, N. Yamashita, and S. Tuncer. Showing objects: Holding and manipulating artefacts in video-mediated collaborative settings. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 5295–5306, 2017. 2

[34] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 300–309, 2023. 2

[35] D. Lindlbauer, A. M. Feit, and O. Hilliges. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*, pp. 147–160, 2019. 2, 3

[36] F. Lu and D. A. Bowman. Evaluating the potential of glanceable ar interfaces for authentic everyday uses. In *2021 IEEE virtual reality and 3D user interfaces (VR)*, pp. 768–777. IEEE, 2021. 2, 3

[37] F. Lu, S. Davari, L. Lisle, Y. Li, and D. A. Bowman. Glanceable ar: Evaluating information access methods for head-worn augmented reality. In *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 930–939. IEEE, 2020. 2, 3

[38] F. Lu and Y. Xu. Exploring spatial ui transition mechanisms with head-worn augmented reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2022. 3

[39] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[40] T. Mok and L. Oehlberg. Critiquing physical prototypes for a remote

audience. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pp. 1295–1307, 2017. 2

[41] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2, 3, 9

[42] Ninjamode. Unity vr gaussian splatting, 2024. 4

[43] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*, pp. 741–754, 2016. 1, 2

[44] S. Park, M. Son, S. Jang, Y. C. Ahn, J.-Y. Kim, and N. Kang. Temporal interpolation is all you need for dynamic neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4212–4221, 2023. 2

[45] T. Pejsa, J. Kantor, H. Benko, E. Ofek, and A. Wilson. Room2room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pp. 1716–1725, 2016. 2

[46] K. Pfeuffer, Y. Abdrabou, A. Esteves, R. Rivu, Y. Abdelrahman, S. Meitner, A. Saadi, and F. Alt. Artention: A design space for gaze-adaptive user interfaces in augmented reality. *Computers & Graphics*, 95:1–12, 2021. 3

[47] M. Sakashita, B. Thoravi Kumaravel, N. Marquardt, and A. D. Wilson. 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*, pp. 1–14, 2024. 2, 3, 4, 8, 9

[48] Y. Seo, Y. S. Choi, H. S. Son, and Y. Uh. Flod: Integrating flexible level of detail into 3d gaussian splatting for customizable rendering. *arXiv preprint arXiv:2408.12894*, 2024. 3

[49] L. Song, A. Chen, Z. Li, Z. Chen, L. Chen, J. Yuan, Y. Xu, and A. Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 2

[50] L. Song, A. Chen, Z. Li, Z. Chen, L. Chen, J. Yuan, Y. Xu, and A. Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 2

[51] T. Teo, A. F. Hayati, G. A. Lee, M. Billinghurst, and M. Adcock. A technique for mixed reality remote collaboration using 360 panoramas in 3d reconstructed scenes. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–11, 2019. 3

[52] T. Teo, L. Lawrence, G. A. Lee, M. Billinghurst, and M. Adcock. Mixed reality remote collaboration combining 360 video and 3d reconstruction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–14, 2019. 2

[53] T. Teo, M. Norman, G. A. Lee, M. Billinghurst, and M. Adcock. Exploring interaction techniques for 360 panoramas inside a 3d reconstructed scene for mixed reality remote collaboration. *Journal on Multimodal User Interfaces*, 14(4):373–385, 2020. 3

[54] B. Thoravi Kumaravel, F. Anderson, G. Fitzmaurice, B. Hartmann, and T. Grossman. 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*, pp. 161–174, 2019. 2

[55] D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao. TripoSR: Fast 3d object reconstruction from a single image. doi: 10.48550/arXiv.2403.02151 2

[56] P. Wang, S. Zhang, X. Bai, M. Billinghurst, L. Zhang, S. Wang, D. Han, H. Lv, and Y. Yan. A gesture-and head-based multimodal interaction platform for mr remote collaboration. *The International Journal of Advanced Manufacturing Technology*, 105(7):3031–3043, 2019. 2

[57] X. Wang, P. E. Love, M. J. Kim, and W. Wang. Mutual awareness in collaborative design: An augmented reality integrated telepresence system. *Computers in industry*, 65(2):314–324, 2014. 2

[58] X. Wang, H. Ye, C. Sandor, W. Zhang, and H. Fu. Predict-and-drive: Avatar motion adaption in room-scale augmented reality telepresence with heterogeneous spaces. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3705–3714, 2022. doi: 10.1109/TVCG.2022.3203109 9

[59] X. Wang, W. Zhang, and H. Fu. A3rt: Attention-aware ar teleconferencing with life-size 2.5d video avatars. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 211–221, 2024. doi: 10.1109/VR58804.2024.00044 3

[60] X. Wang, W. Zhang, C. Sandor, and H. Fu. Real-and-present: Investigating the use of life-size 2d video avatars in hmd-based ar teleconferencing. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):5626–5641, 2025. doi: 10.1109/TVCG.2024.3466554 3

[61] Z. Wang, C. Nguyen, P. Asente, and J. Dorsey. Pointshopar: Supporting environmental design prototyping using point cloud in augmented reality. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–15, 2023. 2

[62] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20310–20320, 2024. 2

[63] A. Wysopal, V. Ross, J. Passananti, K. Yu, B. Huynh, and T. Höllerer. Level-of-detail ar: Dynamically adjusting augmented reality level of detail based on visual angle. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 63–71. IEEE, 2023. 2, 3, 8

[64] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 2, 3, 4, 7

[65] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20331–20341, 2024. 2

[66] Z.-M. Ye, J.-L. Chen, M. Wang, and Y.-L. Yang. Paval: Position-aware virtual agent locomotion for assisted virtual reality navigation. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 239–247. IEEE, 2021. 4

[67] X. Zhang, X. Bai, S. Zhang, W. He, P. Wang, Z. Wang, Y. Yan, and Q. Yu. Real-time 3d video-based mr remote collaboration using gesture cues and virtual replicas. *The International Journal of Advanced Manufacturing Technology*, 121(11):7697–7719, 2022. 2