Immersive Telecommunication with Sharing Perspective and Spatial Awareness

Kai Shishido[1,*], Chun Xie[2] and Itaru Kitahara[2,*]

[1] University of Tsukuba, Doctoral Program in Empowerment Informatics, 305-8573 Tsukuba, Ibaraki, Japan
[2] University of Tsukuba, Center for Computational Sciences, 305-8577 Tsukuba, Ibaraki, Japan

Abstract

Conventional video communication constrains understanding of a partner's intentions and context because it relies on a single, narrow field of view. This paper introduces an immersive, perspective-taking telecommunication method that integrates omnidirectional scene sharing with the partner's first-person perspective video, gaze visualization, IMU-based stabilization, and seamless viewpoint switching. A wide view stream conveys global situational context, while the perspective stream reveals task-level detail; gaze is overlaid in image space to externalize attentional focus. Apparent camera rotations in the perspective stream are compensated through an IMU-driven homography. The system consists of four modules, acquisition, transmission, integration, and presentation, which capture and transmit multimodal streams, align them on reception, overlay gaze on stabilized video, and smooth transition between the wide view and first-person perspectives through an intermediate viewpoint that maintains geometric and motion continuity. The design supports symmetrical and bidirectional use. Effectiveness is evaluated in three user studies addressing different aspects of perspective-taking communication.

Keywords

Telecommunication, Perspective Taking, Remote Collaboration, AR, VR.

1. Introduction

With the development and widespread adoption of information and communication technology, the demand for telecommunication has further intensified [1]. In particular, video communication platforms such as Zoom, Microsoft Teams, and Google Meet have become vital means of communication in fields like remote collaboration, education, and support services [2][3].

On the other hand, in scenarios involving remote collaboration or support, it is highly effective for smooth communication to not only listen to the partner's speech but also to share contextual information based on their situation, such as what kind of environment they are in, what they are looking at, and what they are paying attention to. In telecommunication, there are moments when one needs to grasp the surrounding circumstances of the partner, and moments when one wants to stand in the partner's shoes to understand their object of interest or behavioral intentions. Thus, we naturally strive for mutual understanding by shifting back and forth between subjective and objective viewpoints [4]. Even in current video communication systems, it is possible to switch perspectives through camera switching or screen sharing. However, most of these transitions are instantaneous cuts that do not explicitly depict the spatial continuity between the different viewpoints. As a result, users are forced to mentally reconstruct "where and from what angle the partner is looking now," which can compromise their understanding of the partner's situation, intention, and focus, while also diminishing their sense of presence.

To address these challenges inherent in conventional video communication, this research aims to simultaneously satisfy the following three requirements, C1, C2 and C3.

- **C1: Context Awareness via Wide-View Presentation**— Continuous sharing of the surrounding situation without field-of-view limitations, capturing the partner's gestures and body language to support the understanding of their intentions and emotions.

APMAR'26: The 18th Asia-Pacific Workshop on Mixed and Augmented Reality, Aug. 3-5, 2026, Taipei, Taiwan"
[1*] Corresponding author.
shishido.kai@image.iit.tsukuba.ac.jp (K.Shishido); xiechun@ccs.tsukuba.ac.jp (C.Xie); kitahara@ccs.tsukuba.ac.jp (I.Kitahara)
0000-0003-4936-7404 (C. Xie); 0000-0002-5186-789X (I. Kitahara)

- **C2: Focus Sharing via Partner's Perspective Presentation**— Displaying the partner's first-person perspective to make it easier to indicate and share the specific object of attention.
- **C3: Spatial Continuity in Viewpoint Transitions**— Providing an intermediate viewpoint that allows for smooth, seamless transitions between the wide view and the partner's first-person view, thereby maintaining and enhancing situational comprehension and presence.

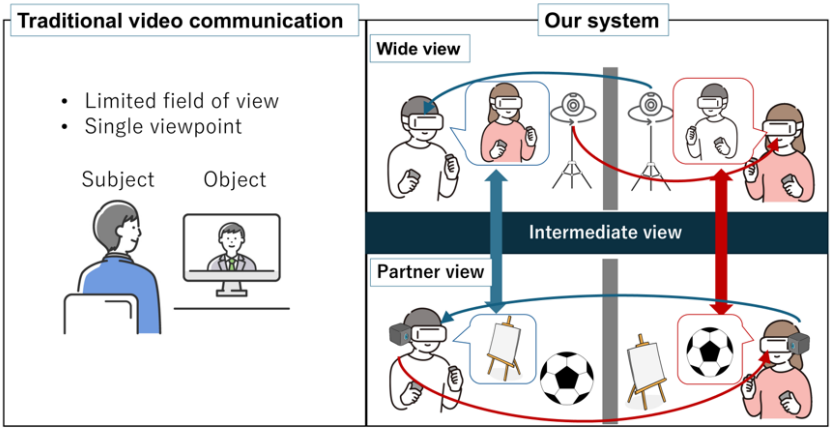


Figure 1: Conventional video communication (left) typically relies on a single video stream that captures only the partner's upper body. In contrast, the telecommunication realized in this study (right) enables the observer to access both an omnidirectional (wide view) stream captured by a camera placed in the partner's space and the partner's first-person perspective. This combination is expected to facilitate understanding of the partner's gestures as well as the surrounding situational context. In addition, presenting an intermediate view during viewpoint switching aims to maintain and enhance presence. The same mechanism applies symmetrically when the roles of observer and partner are interchanged.

Conventional video communication generally transmits video and audio via a single-perspective, narrow-field-of-view camera connected to a PC. While this setup is useful for basic remote conversation, it is insufficient to meet requirements C1 to C3 simultaneously. Specifically, the narrow field of view fails to share enough surrounding context, leading to a loss of information necessary for understanding intentions and emotions (lack of C1). Furthermore, without the partner's first-person perspective, indications of their actual focus remain ambiguous (lack of C2), hindering effective communication. These limitations manifest symmetrically even when the roles of the subject user (sender) and object user (recipient) are interchanged.

As a related work, JackIn Head [5] shares the physical environment by using a wearable omnidirectional camera to project surrounding video, allowing an observer using a Head-Mounted Display (HMD) to look around freely. While this satisfies requirement C1 (wide-view presentation), it does not incorporate the partner's specific first-person view (C2) or consider continuous transitions between viewpoints (C3) as design objectives. JackIn Space [6] combines a head-mounted partner-perspective camera with multiple depth sensors deployed in the environment to achieve a smooth switch to a third-person view. While this partially addresses requirement C3 (continuous transitions via intermediate views), the local and remote roles are asymmetrical, with high-performance equipment concentrated on one side. Operating both sides on equal footing would incur high initial setup costs, significant communication bandwidth, and strict location constraints. While remote robotics, telepresence, and collaborative virtual spaces offer high degrees of visualization freedom, there are few reports that comprehensively integrate live-action wide-view presentation (C1) and partner-perspective presentation (C2) with continuous transitions (C3) to mitigate the visual discomfort during switching [7]-[11]. Consequently, it can be argued that a live-action-based bidirectional telecommunication system that satisfies C1 (wide-view), C2 (partner-perspective), and C3 (continuous transition) simultaneously has not yet been fully established.

The purpose of our research is to propose a implementation method that comprehensively fulfills requirements C1–C3 in live-action-based telecommunication. As shown in the right side of Figure 1, this paper proposes a telecommunication method that enables switching between wide-view video and partner-perspective video, with an intermediate viewpoint bridging the two to ensure a continuous transition. Wide-view presentation (C1) is achieved using an omnidirectional camera and an HMD to enhance the sense of shared space and immersion. Partner-perspective presentation (C2) is realized using a head-mounted camera. This approach is inspired by the psychological concept of "perspective-taking," which refers to the act or capacity to understand the thoughts, emotions, and viewpoints of others [12]. We hypothesize that by simulation-testing another person's experience, one can deepen their understanding of that person. Mutually sharing experiences is expected to promote reciprocal understanding, thereby enabling genuine mutual comprehension [13]-[15]. Furthermore, by introducing continuous viewpoint transitions via an intermediate view that links the wide view and the partner's perspective (C3), the system suppresses visual discomfort during switching, aiming to

maintain and enhance situational understanding and presence [16][17]. This method supports bidirectional operation by applying the identical process symmetrically when interchanging the roles of the subject and object users, offering high scalability for video communication.

2. Proposed Telecommunication System

An immersive telecommunication system capable of sharing viewpoints and spaces is shown in Figure 2. The system consists of four primary modules: “Acquisition Module”, “Transmission Module”, “Integration Module”, and “Presentation Module”. Each module is defined by its role in terms of “what information it receives as input”, “how it processes that information”, and “what it outputs to the next stage”.

The Acquisition Module captures the wide-view video that provides global environmental context, the partner-perspective video that provides localized task details, eye-tracking data, head posture, and control inputs. The Transmission Module conveys the streams of wide-view video, partner-perspective video, eye-tracking data, head posture, and control inputs. The Integration Module utilizes a reception queue to overlay eye-tracking data onto the partner-perspective video and stabilizes the video by compensating for head-movement jitter, producing video assets suitable for presentation to the user. Gaze information is visualized as a fixation point on the video. The Presentation Module determines the timeline of when, what, and in which posture items are displayed, ensuring switching continuity through an intermediate viewpoint. It governs the display states of the wide-view video, partner-perspective video, gaze visualization, and the intermediate transitions during switching.

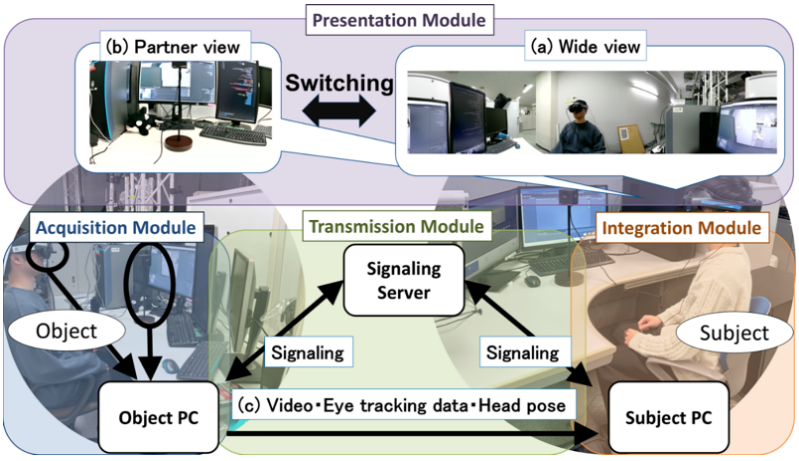


Figure 2: The communication process between the subject and the object: (a) wide view video from the object’s environment, (b) the partner view video, as well as the object’s eye-tracking data and head pose data are transmitted from object PC to subject PC, where the subject can view them.

2.1. Acquisition Module: Signal Model and Assumptions

2.1.1. Signal Model

As shown in Figure 3, the Acquisition Module extracts wide-view video providing global environmental context, partner-perspective video providing localized task details, and sensing signals (gaze, head posture, control inputs) as synchronized time-series. Video streams are captured as frame sequences, while sensing streams are established as sampling sequences.

2.1.2. Presentation and Observation of Wide-View Video

The presentation of wide-view video is achieved by projecting the captured omnidirectional video onto the inner surface of a virtual sphere. Geometric parameters such as projection format, horizontal alignment, color space, pixel layout, and mirroring flags are captured on the acquisition side and matched identically on the presentation side to establish symmetrical execution. Resolution and frame rate are aligned with the subsequent transmission and presentation layers.

2.1.3. Presentation and Observation of Partner-Perspective Video

Partner-perspective video is presented by projecting the captured video onto a virtual flat plane placed within the environment. Image-capturing parameters (minimum resolution, exposure stability, maximum allowable motion blur) are designed to guarantee legibility required for remote tasks.

2.1.4. *Sensing Signal Specifications (Gaze and Head Posture)*

Gaze information is calculated as the intersection point between the user's gaze ray originating from the HMD and the planar mesh rendering the partner-perspective video. The head posture and angular velocity measured by the internal IMU and tracking loops are retrieved as a posture sequence relative to a fixed reference frame. Control inputs represent button actions performed on handheld input controllers captured as discrete event series.

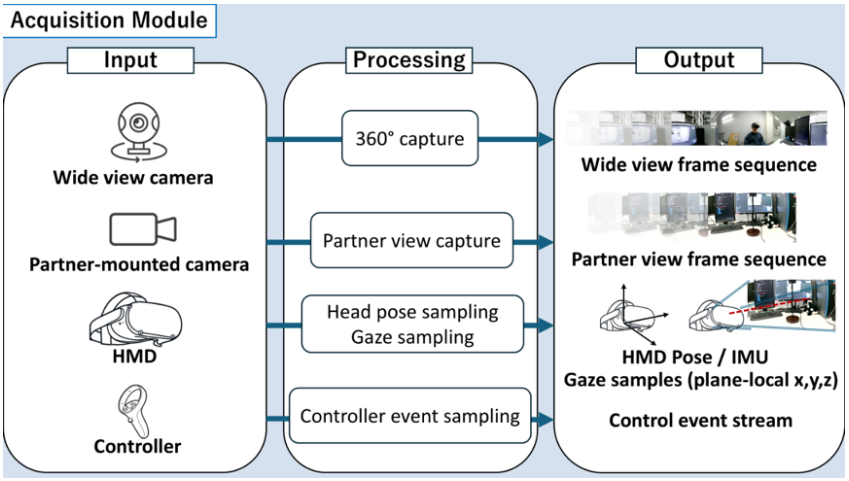


Figure 3: Acquisition module: The module captures omnidirectional (wide view) video for global context, partner view video for local task details, and sensing signals (gaze, head pose, controller input). These are converted into synchronized time series (frame sequences and sampled signals) and passed to the transmission module.

2.2. Transmission Module: Low-Latency Transportation and Sequential Refreshing

2.2.1. *Stream Separation and Transmission Schemes*

As illustrated in Figure 4, high-throughput video streams (wide-view and partner-perspective) are handled via continuous media streaming, whereas lightweight telemetry streams (gaze, head posture, control event streams) are delivered via packet-based channels. Designing with architectural symmetry allows identical execution when the roles of subject and object users are swapped.

2.2.2. *Reception Processing and Synthesis*

The transmission module keeps separate lanes for the video sequences and telemetry sequences, making them available at the receiver side. Telemetry coordinates keep updating with the latest packets to be polled synchronously with the display refresh loop, whereas the video pipeline extracts decoded frames as raw texture updates for the presentation layer.

2.2.3. *Coordinate Representation*

Gaze is sent as a local plane coordinate *(x,y,z)* computed from the intersection of the gaze ray and projection plane at the transmitter. The receiver maps this local coordinate into the texture plane to draw the red fixation pointer. Head posture tracking telemetry serves as an input layer to neutralize video jitter caused by head movements in the partner-perspective stream.

2.2.4. *Bidirectional Operation*

The system features a symmetrical framework capable of operating under identical pipelines when swapping user roles. Both lanes share a unified global coordinate definition and projection plane geometries, applying matching video projection, gaze mapping, and viewpoint switching control mechanics.

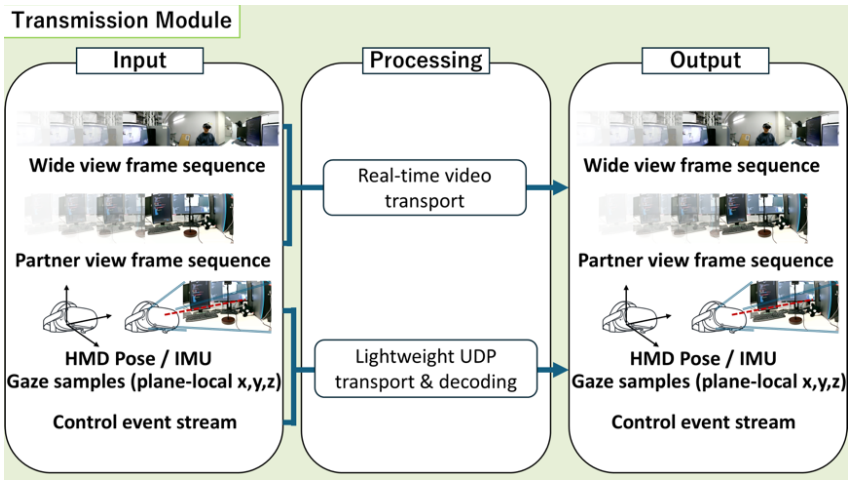


Figure 4: Transmission module: Video streams (wide view and partner view) are sent via real-time media transport, while lightweight streams (gaze, head-pose reference, control flags) are sent via packet-based channels.

2.3. Integration Module: Gaze Overlay, Jitter Compensation, and Spatial Alignment

2.3.1. Gaze Overlay

As detailed in Figure 5, the Integration Module converts incoming plane-local gaze coordinates into normalized texture space coordinates *(u,v)* based on the target geometry, which are subsequently translated into pixel-space image coordinates to isolate the target fixation location.

2.3.2. Video Stabilization via Homography-Based Jitter Compensation

The partner-perspective stream exhibits erratic video movements corresponding to the physical head movements of the object user. This method calculates a 2D projective transformation (homography) derived from the real-time HMD tracking telemetry to counter-balance head-induced visual shifts. By applying frame-by-frame 2D projective mapping combining lens parameters with structural orientation matrix offsets, rotational camera shifts are neutralized. The compensation factor adapts to angular velocities to prevent over-compensation or jumping artifacts during rapid transitions. Identical coordinate mappings are applied to the gaze data points to maintain spatial cohesion between the stabilized frames and the overlaid target point.

2.3.3. Real Space to Presentation Space Calibration

To transition seamlessly, a 3D point-cloud proxy representation of the physical workspace is integrated during viewpoint switching. Here, the active HMD position tracking values of the object user serve as the goal pose for the intermediate viewpoint. Any geometric mismatch between the point cloud and physical space causes position/orientation jumps. To eliminate this issue, a mapping transformation (handling scale, translation, and rotation components) manages the link between the HMD tracking loop and the point cloud space. During viewpoint changes, the system maps the destination of the intermediate viewpoint to the estimated pose of the object user's HMD to hold a unified frame of reference. This turns point-cloud space navigation into a natural representation of movement in the physical world, sustaining spatial continuity. As shown in Figure 6, the system handles three coordinate spaces:

- Stage Coordinate System: The local physical space tracking frame of the HMD.
- Cloud-Local Coordinate System: The local coordinate space of the point-cloud proxy data.
- World Coordinate System: The global virtual environment coordinate system where all elements are arranged.

To harmonize the raw tracking inputs with the point cloud, the transformation from Stage to Cloud-Local coordinates is defined as:

$$\boldsymbol{P}_C = \boldsymbol{T}_{local} + s \cdot (\boldsymbol{R}_{local}\boldsymbol{P}_S) + \boldsymbol{d}_{offset} \quad (1)$$

$$\boldsymbol{q}_C = \boldsymbol{R}_{local} \otimes \boldsymbol{q}_S \quad (2)$$

Where $\boldsymbol{P}_S$ and $\boldsymbol{q}_S$ are position/posture in Stage space, $\boldsymbol{P}_C$ and $\boldsymbol{q}_C$ represent targets in Cloud-Local space, $\boldsymbol{R}_{local}$ handles the rotation matrix between spaces, $\boldsymbol{T}_{local}$ tracks translational offsets, s represents user-specific scale adjustments, and $\boldsymbol{d}_{offset}$ allows manual adjustments.

To bridge Cloud-Local positions with the shared global coordinate scheme, the transformation to World coordinates is defined as:

$$\boldsymbol{P}_W = \boldsymbol{p}_{root} + \boldsymbol{R}_{root}\boldsymbol{P}_C \tag{3}$$

$$\boldsymbol{q}_W = \boldsymbol{q}_{yaw} \otimes (\boldsymbol{R}_{root} \otimes \boldsymbol{q}_C) \tag{4}$$

Where $\boldsymbol{P}_W$ and $\boldsymbol{q}_W$ define position/posture in World coordinates, $\boldsymbol{p}_{root}$ and $\boldsymbol{R}_{root}$ control the placement of the point cloud array in the global layer, and $\boldsymbol{q}_{yaw}$ handles orientation adjustments around the vertical axis.

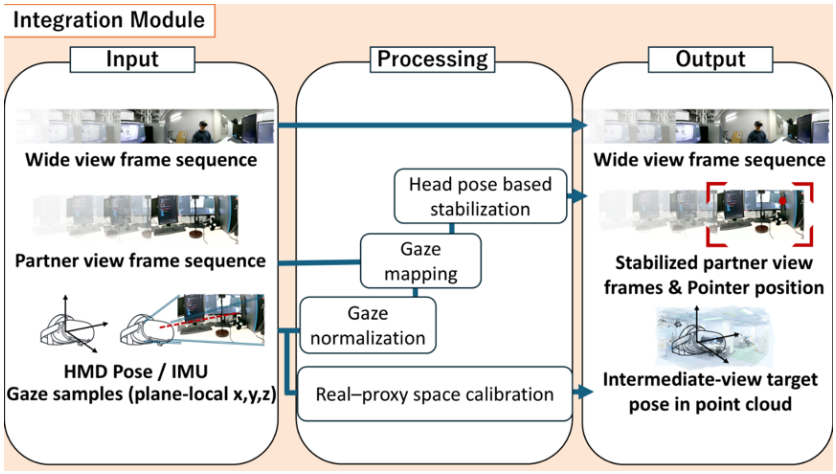


Figure 5: module: The module maps received gaze samples from plane-local coordinates to texture and image coordinates, overlays a pointer on the partner view image, compensates head-induced rotation using a homography derived from head pose, and aligns the reconstructed point-cloud space with the tracked HMD pose to maintain consistency between real and virtual spaces.

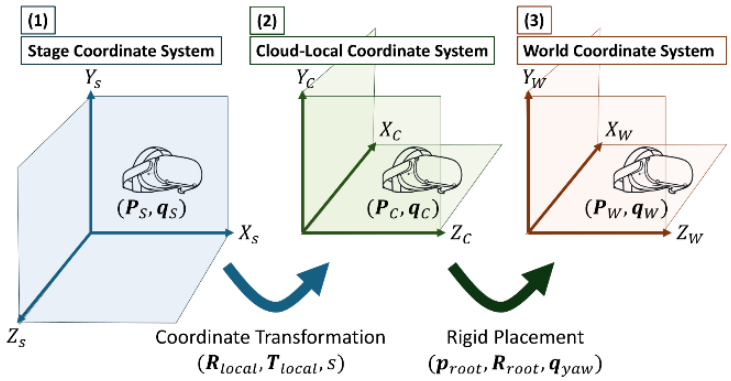


Figure 6: Pose mapping across three coordinate systems (Stage, point-cloud local, and World) that enables continuous viewpoint transitions between wide-view, intermediate, and partner-view presentations.

2.4. Presentation Module: View States, Intermediate Viewpoints, and Synchronization

2.4.1. Presentation States and Transition Control

As outlined in Figure 7, the presentation layer executes three viewing profiles: wide-view, intermediate viewpoint, and partner-perspective view. View transitions follow a symmetrical sequence (Wide-view ->

Intermediate -> Partner-perspective and vice versa) initiated by controller event signals.

2.4.2. Intermediate Viewpoint

The intermediate view is rendered within the point-cloud proxy space to provide structural continuity between the wide-view sphere and the partner's planar viewpoint. Because its target path lands directly on the tracked path of the remote user's HMD, it serves as a visual navigation cue rather than a cosmetic effect, helping users anticipate the target position during the transition. Opacity parameters of the spherical wide-view shell, planar window, and point-cloud array are adjusted dynamically to produce a seamless cross-fade.

2.4.3. HMD Synchronization

Physical HMD tracking inputs map directly into the point-cloud coordinates utilizing the alignment settings from section 2.3.3. This enables cohesive real-time video blending without breaking geometric tracking relationships during head movements.

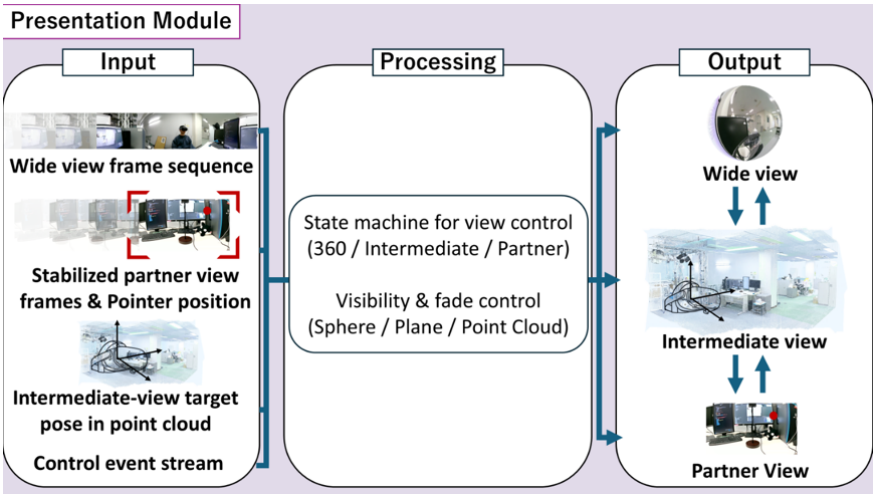


Figure 7: Presentation module: The module manages three presentation states—wide view, intermediate viewpoint in the point-cloud space, and partner view—and switches between them based on controller input. During switching, visibility of the spherical (wide view), point-cloud, and planar (partner view) elements is smoothly faded, while the camera is interpolated toward the calibrated target pose so that motion and spatial continuity are preserved.

3. Implementation of Immersive Telecommunication Framework

We implemented the comprehensive architecture described in Section 2 on the Unity. High-definition environmental video ingestion runs via WebCamTexture loops. Real-time multi-stream transportation relies on the Unity Render Streaming library backed by optimized WebRTC protocols [18]. Structural geometries, homographies, and image space mapping conversion operations are handled efficiently via custom compute pipelines linking OpenCVForUnity native components [19]. Device instrumentation tracking pipelines retrieve physical parameters and sensor readings via Unity XR tracking frameworks. Point-cloud environment generation utilizes optimized Panoramic Neural Radiance Field from a Single Panorama (PeRF) algorithms [20], allowing our pipelines to construct high-fidelity spatial data sets from single panoramas and parse them seamlessly into the Unity execution layers as interactive proxy spaces. Individual module implementation parameters are detailed below:

3.1. Ingestion and Acquisition Layer Execution

3.1.1. Omnidirectional Context Streaming Ingestion

An industrial USB Video Class (UVC) panoramic lens system acts as our physical wide-view collection point, executed programmatically via custom WebCamTexture polling threads. For every update frame, pixel vectors copy directly onto hardware-bound GPU RenderTexture slots to undergo streaming delivery. Structural formats, horizontal baseline calibrations, texture color profiles, sampling layouts, and viewport flip attributes are matched to keep structural harmony with the downstream spherical environment display loops.

3.1.2. *Task-Level Partner-Perspective Video Ingestion*

The head-mounted partner camera runs on an identical independent WebCamTexture driver, dumping synchronized update sequences onto dedicated RenderTexture buffers. Because this feed serves to indicate fine workspace actions, exposure properties, shutter parameters, and grid formats are locked to preserve edge sharpness and character legibility for remote operators. Hardware-bound geometric calibration or pixel filtering are excluded from this tier, feeding clean raw arrays to preserve low-latency transportation.

3.1.3. *Multimodal Sensing and Event Telemetry Ingestion*

Gaze telemetry relies on hardware-integrated tracking frameworks embedded inside the Meta Quest Pro headset to gather left/right eye gaze rays. Our processing threads compute real-time spatial intersections against the virtual plane mesh rendering the partner-perspective window. The resulting intersection points match plane-local coordinate parameters, compiled as a continuous series of (x,y,z) coordinate structures. Concurrently, head posture metrics and orientation quaternion streams are pulled from the XR Stage coordinate systems and stored alongside instantaneous angular velocity data arrays. User actions performed on hardware controllers are monitored through asynchronous event loops to map button strokes or structural flag toggles directly into logic control structures.

3.2. Transportation Infrastructure Layer Execution

3.2.1. *High-Throughput Media Transportation Handling*

Wide-view environment blocks and task-level perspective streams register as independent input elements inside the Unity Render Streaming controller, executing low-latency transportation via optimized WebRTC protocols. Active view state switching is achieved by updating the active source reference pointers linking the streaming loops. Video blocks undergo acceleration via hardware-bound H.264 video encoders, allowing wide-view environment data sets to scale dynamically to higher bitrate properties compared to local text windows.

3.2.2. *Asynchronous Telemetry Transport Pipelines*

Plane-local tracking data structures are transmitted from the source using three-component floating-point structures. The receiver framework handles incoming packets asynchronously via UDP (User Datagram Protocol) to minimize network overhead. Logic command flags (such as active state changes) pass via dedicated 1-byte control indicators. Structural head posture orientation values translate directly into quaternion packages to feed the video stabilization pipeline on the destination host.

3.2.3. *Temporal Matrix and Queue Alignment Handling*

The destination host extracts telemetry elements asynchronously via UDP, overriding internal status parameters to synchronize with the standard engine update loop. Concurrently, incoming media files update via active WebRTC callbacks. This approach guarantees that gaze overlays and view state adjustments reflect telemetry states immediately upon video frame ingestion.

3.3. Integration Processing Layer Execution

3.3.1. *Multimodal Gaze Coordinate Linking Chains*

Plane-local coordinates enter a normalization pass based on the layout metrics of the presentation view window to extract target texture parameters *(u,v)*. These coordinates translate into pixel coordinates inside the image space. To maintain spatial alignment during head movements, inverse transformation operations translate the gaze pointer to match the homography transformation applied to the stabilized video frame, restoring consistent coordinate mappings.

3.3.2. *Homography-Driven Video Jitter Stabilization*

Erratic video shifts in the partner-perspective feed are suppressed via OpenCVForUnity projective transformations. Camera lens parameters, field-of-view matrices, focal attributes, and principal points are resolved dynamically, while tracking data sets compute real-time structural transformation matrices. To prevent over-correction during rapid head turns, compensation coefficients scale adaptively based on real-time angular velocity telemetry before executing frame updates using warpPerspective native calls.

3.3.3. *Multi-Space Spatial Matching & Alignment Calibration*

To enable smooth transitions through the 3D proxy environment, HMD position tracking values from the source are mapped directly onto coordinate systems inside the point cloud. Scale parameters, translational offset variables, and structural rotation matrices are calculated during initialization and stored to enable persistent configuration restoration. The presentation layer uses these configuration states to drive the destination parameters of the intermediate viewpoint camera, aligning them with the remote operator's physical viewpoint. Any persistent alignment offsets can be manually adjusted via an administrative overlay.

3.4. Presentation and Visualization Layer Execution

3.4.1. View State Lifecycle and Transition Management

The interface rendering pipelines manage three states: wide-view environment rendering (Figure 8), partner-perspective overlay view (Figure 9), and intermediate viewpoint visualization. State modifications are driven by controller inputs or received remote command flags. Transitions execute symmetrical interpolation trajectories (Wide-view -> Intermediate -> Partner-perspective and vice versa). While a transition is active, shader opacity values modify the visibility of the wide-view sphere shell, the planar window mesh, and the point-cloud proxy array to create smooth visual cross-fades.

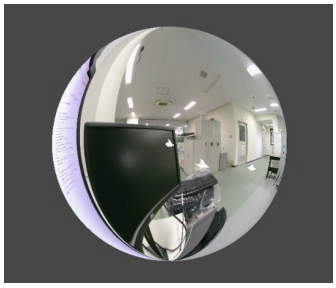


Figure 8: Wide view projected onto the inner sphere representing the partner's surrounding environment.

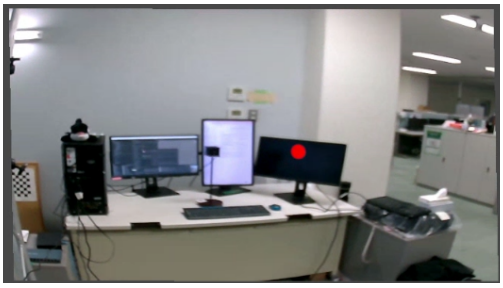


Figure 9: Partner view mapped onto a front-facing plane, with gaze position overlaid in red.

3.4.2. Intermediate Spatial Interpolation & Rendering

Point-cloud structures derived from single panoramas via PeRF are integrated directly within the Unity hierarchy to act as a 3D proxy space during viewpoint transitions (Figure 10). Camera spatial position updates via linear interpolation paths, while camera orientation values use spherical linear interpolation (Slerp) steps to eliminate angular velocity spikes or tracking discontinuities. Opacity properties are cross-faded smoothly across the layers to preserve spatial context.

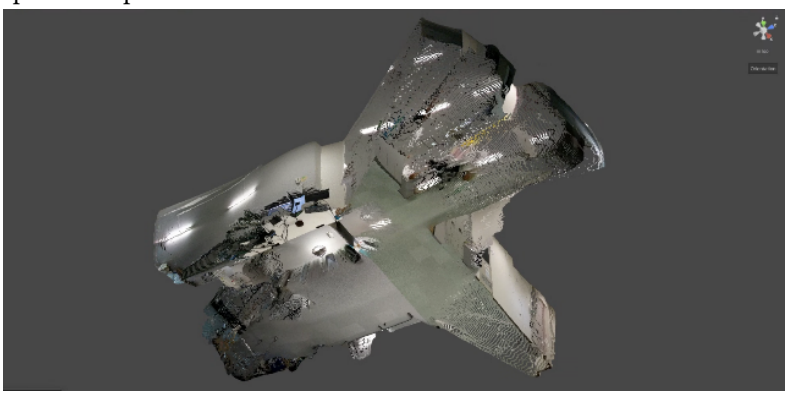


Figure 10: Proxy 3D environment as a point cloud reconstructed from a panorama using PeRF, serving as an intermediate space for viewpoint transitions.

3.4.3. *Device State Retention and Calibration Continuity*

Hardware calibration profiles and matching scale parameters are persistently saved to disk. This architecture guarantees identical replication profiles across application lifecycles, maintaining structural continuity without requiring recalibration after system reboots.

4. Experimental Evaluations

In this evaluation, we investigate the effectiveness of the proposed method by addressing the following Research Questions (RQs):

- **RQ1 (Corresponding to C1):** Does video presentation without field-of-view restrictions improve the subject user's comprehension of the object user's intentions and emotions?
- **RQ2 (Corresponding to C2):** Does the addition of the partner's first-person perspective improve the subject user's understanding of the object user's focus of interest?
- **RQ3 (Corresponding to C1–C3 integration):** Does the integrated setup combining wide-view video, partner-perspective video, and intermediate viewpoints enhance scenario understanding (context, intent, task execution) and presence?

To address these research questions, this paper conducts Experiment 1 through Experiment 3. Experiment 1 evaluates the impact of omnidirectional viewing on intention and emotion recognition (RQ1), while Experiment 2 measures focus sharing performance via the partner's view and gaze visualization overlays (RQ2). Finally, Experiment 3 comprehensively verifies the influence of the proposed configuration—uniting wide-view video, partner-perspective data, and intermediate viewpoint interpolation—on overall scenario comprehension and user presence (RQ3).

4.1. Experiment 1: Comparative Evaluation on Understanding the Partner's Intentions and Emotions

The objective of Experiment 1 is to verify whether lifting the field-of-view restrictions is effective in helping the subject user interpret the object user's intentions or emotions, thereby answering RQ1.

4.1.1. *Procedure*

We recorded baseline videos representing traditional video communication, capturing the object user from the shoulders up while performing five types of body language [21]-[23] known to correlate with specific emotional states listed in Table 1. By using the proposed system, we recorded the object user executing the identical five body language sets. Participants acting as subject users observed both video groups and inferred the intended emotional state. Additionally, subject users evaluated the perceived readability of the object user's body language across both conditions.

4.1.2. *Evaluation Method*

For each body language sequence, the subject user rated the perceived intensity of each emotional category on a 7-point Likert scale (ranging from -3 = "Strongly Disagree" to +3 = "Strongly Agree"). Furthermore, participants provided a 7-point scale rating to assess whether the proposed system improved the visibility and tracking of body language compared to the baseline traditional video communication setup.

4.1.3. *Results*

The results obtained from 15 participants are summarized in Figure 11. Comparing traditional video communication with our proposed system, the baseline condition failed to convey the object user's intended emotions accurately to the subject user, resulting in a wider distribution of responses. Conversely, when using our system, the object user's target emotions were identified with significantly higher accuracy. Furthermore, subjective ratings regarding body language visibility confirmed that the proposed framework provided a clear and easily observable viewing experience compared to the baseline.

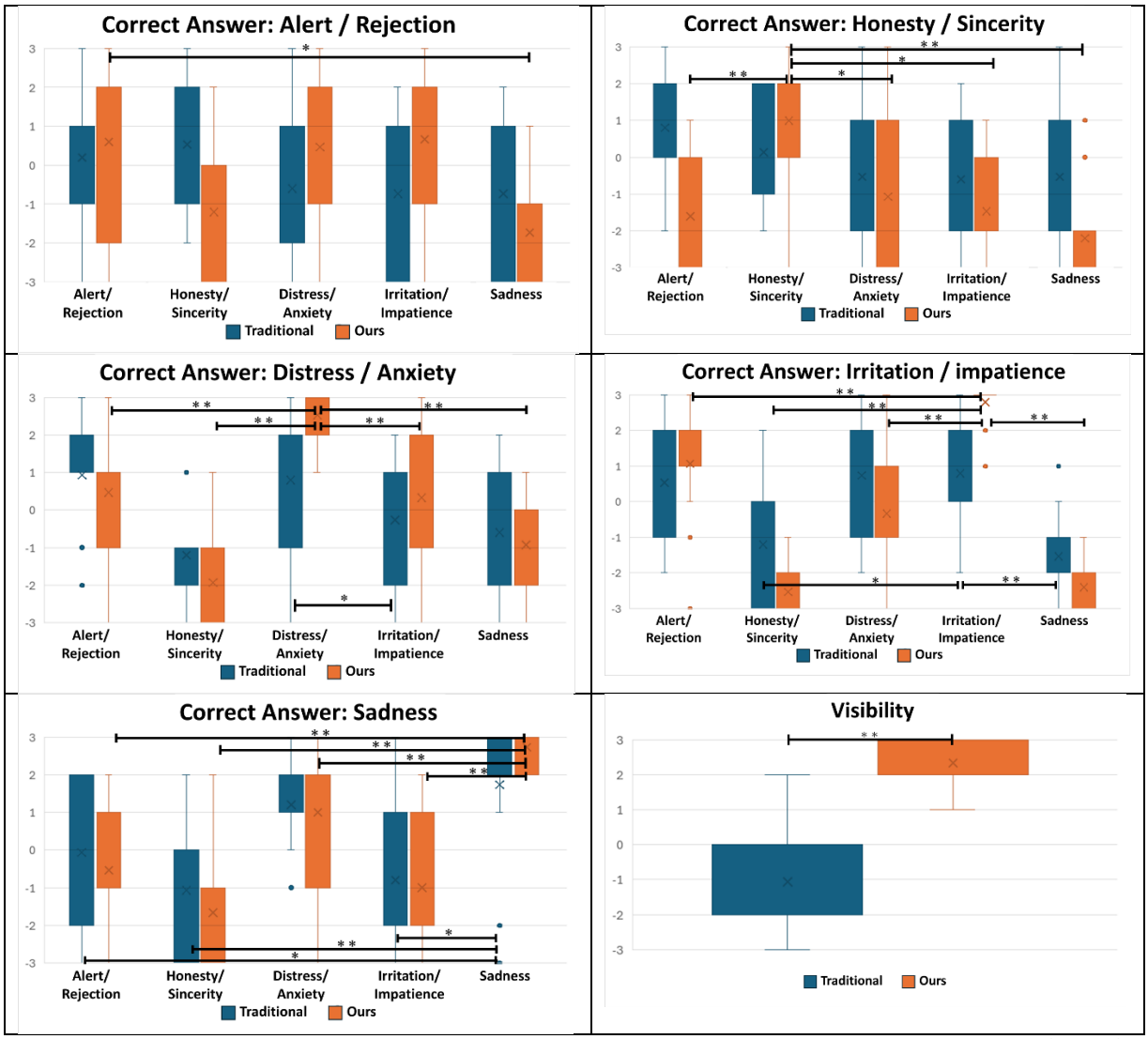


*Figure 11: Box-and-whisker plots of intended and rated emotions. The intended emotions are wariness/rejection (top-left), honesty/sincerity (top-right), anxiety/agitation (middle-left), impatience/irritation (middle-right), and sadness (bottom-left). The bottom-right plot compares the perceived readability of body language between the traditional and proposed systems. * and ** indicate statistically significant differences after multiple-comparison correction (*: $p < 0.05$, **: $p < 0.01$).*

Discussion

To confirm whether the intended emotional categories were evaluated significantly higher than non-intended emotions, a repeated-measures one-way ANOVA was conducted for each target emotion across both traditional and proposed methods using the five emotional ratings (-3 to +3) as levels. When the ANOVA indicated a significant main effect, paired-sample comparisons were performed between the target emotion and the remaining four categories, utilizing the Holm method for multiple-comparison correction. In Figure 12, * and ** denote corrected statistical significance at $p < 0.05$ and $p < 0.01$, respectively. Overall, the statistical analysis confirms that our system supports emotional state communication more effectively than traditional platforms. However, certain adjacent categories, such as "Wariness/Rejection" versus "Honesty/Sincerity," showed a lower mean rating trend for the sincere profile, though it did not reach statistical significance. Similarly, no significant differences were observed between "Wariness/Rejection" and "Distress/Anxiety," or "Wariness/Rejection" and "Impatience/Irritation." A potential contributing factor to these overlaps is emotional category similarity, combined with the structural occlusion of the eye region caused by the HMD, which removed critical facial expression cues.

4.2. Experiment 2: Evaluation on Identifying the Partner's Object of Interest

The purpose of Experiment 2 is to evaluate whether perspective-taking and real-time gaze visualization support the subject user's tracking of the object user's interest focus, addressing RQ2.

4.2.1. *Procedure*

The object user sequentially focused their attention on five target objects arranged on a desk as illustrated in Figure 12. Concurrently, the subject user entered the perspective-taking state and, under the condition where the object user's gaze path was actively visualized, identified which object the partner was observing. Each subject participant performed this test 5 times while the target objects changed randomly.

Figure 12: Five objects of interest: top left is a folding fan, top right is a shark plush toy, bottom left is a smartphone, center is a controller, bottom right is a mouse.

4.2.2. *Evaluation Method*

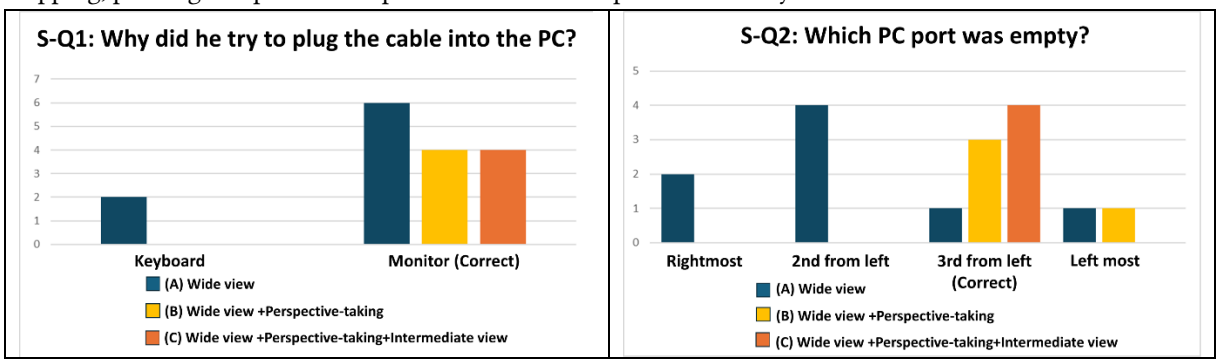
Performance was evaluated based on the absolute correctness of the object identified by the subject user, utilizing the total number of correct trials as the primary evaluation metric.

4.2.3. *Results*

Out of a total of 75 trials executed across all participants, a perfect accuracy score of 100% was recorded (75/75 correct responses). This result demonstrates that every participant successfully localized the object user's focus of interest without errors.

4.2.4. *Discussion*

The 100% success rate across 75 trials demonstrates that perspective-taking allows for high-precision identification of a partner's point of focus. Specifically, real-time gaze visualization allowed subject users to interpret spatial relationships intuitively. For comparison, a preceding pilot study evaluating a perspective-taking condition without the red fixation dot recorded 46 correct responses out of 48 trials (approximately 95.8% accuracy). This comparison suggests that explicit gaze overlay minimizes the cognitive load of spatial mapping, pushing comprehension performance to near-perfect accuracy.



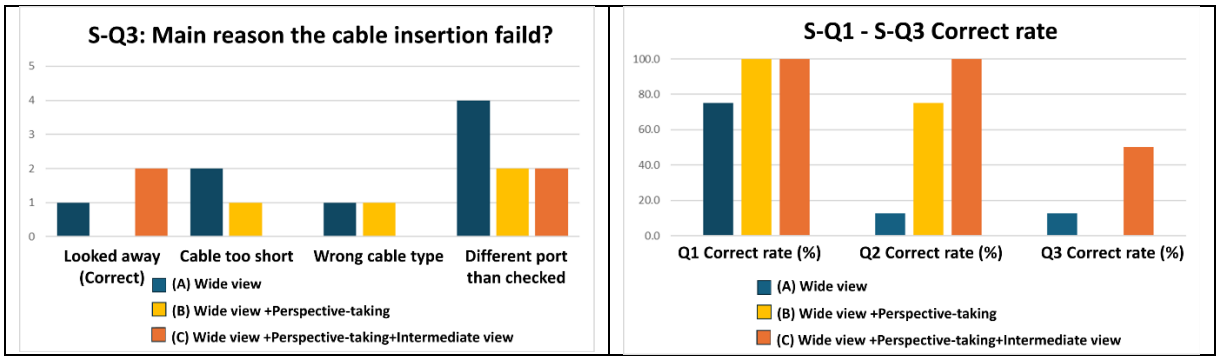


Figure 13: Response distributions and correct rates for Experiment 3-1 (scenario understanding) under three presentation conditions (wide view-only, wide view+perspective-taking, and wide view+perspective-taking+intermediate view). The upper-left to lower-left panels show participants' selected answers for S-Q1 (goal recognition), S-Q2 (situation recognition: identifying the empty port), and S-Q3 (understanding the reason for failure). The lower-right panel shows the correct rate (%) for S-Q1–S-Q3 in each condition.

4.3. Experiment 3: Evaluation on Scenario Comprehension and Intermediate Viewpoint Transitions

Experiment 3 comprehensively investigates the performance of the proposed method from two main perspectives to answer RQ3.

Experiment 3-1 evaluates scenario comprehension (intent, contextual empty-port localization, and failure-cause mapping) by comparing three viewing conditions using recorded sequences of a physical cable-routing task: (A) Wide-view only, (B) Wide-view + Partner's perspective, and (C) Wide-view + Partner's perspective + Intermediate viewpoint. Real-time gaze overlays were enabled for conditions featuring the partner's view.

Experiment 3-2 isolates the viewpoint switching mechanic across the identical three conditions to evaluate the impact of intermediate view interpolation on user presence, switching continuity, perceived self-motion, predictability, post-switch spatial orientation, and subjective preference.

4.3.1. Procedure

For Experiment 3-1, participants watched a series of cable-routing tasks across three counterbalanced conditions (A, B, and C). For Experiment 3-2, a separate set of video assets was provided where participants could freely trigger viewpoint changes while the presentation order was systematically counterbalanced to neutralize bias.

4.3.2. Evaluation Metrics

In Experiment 3-1, the accuracy rates for S-Q1 ("Goal Recognition"), S-Q2 ("Situation Recognition: Empty Port Identification"), and S-Q3 ("Understanding Failure Causes") were measured alongside Simulator Sickness Questionnaire (SSQ) tracking. For Experiment 3-2, 7-point Likert scale tracking was compiled for: T-Q1 (Co-presence), T-Q2 (Virtual Environment Presence), T-Q3 (Spatial Layout Comprehension), T-Q4 (Perceived Transition Continuity), T-Q5 (Sensation of Self-Motion), T-Q6 (Target Path Predictability), and T-Q7 (Post-Switch Orientation Speed). Finally, a forced-choice question captured overall transition style preference.

4.3.3. Results

The results compiled from 16 participants for Experiment 3-1 are illustrated in Figure 13. Conditions incorporating the partner's view yielded superior scenario understanding compared to wide-view viewing alone, particularly for S-Q2. While S-Q3 performance improved slightly with perspective-taking, some errors remained. SSQ scores stayed low within safe operational levels across all profiles without statistical divergence. The subjective benchmarks for Experiment 3-2 are shown in Figure 14. Condition C outperformed instantaneous switching across T-Q1, T-Q3, T-Q4, T-Q5, and T-Q7 benchmarks. Forced-choice metrics confirmed that a substantial majority of users preferred intermediate view interpolation.

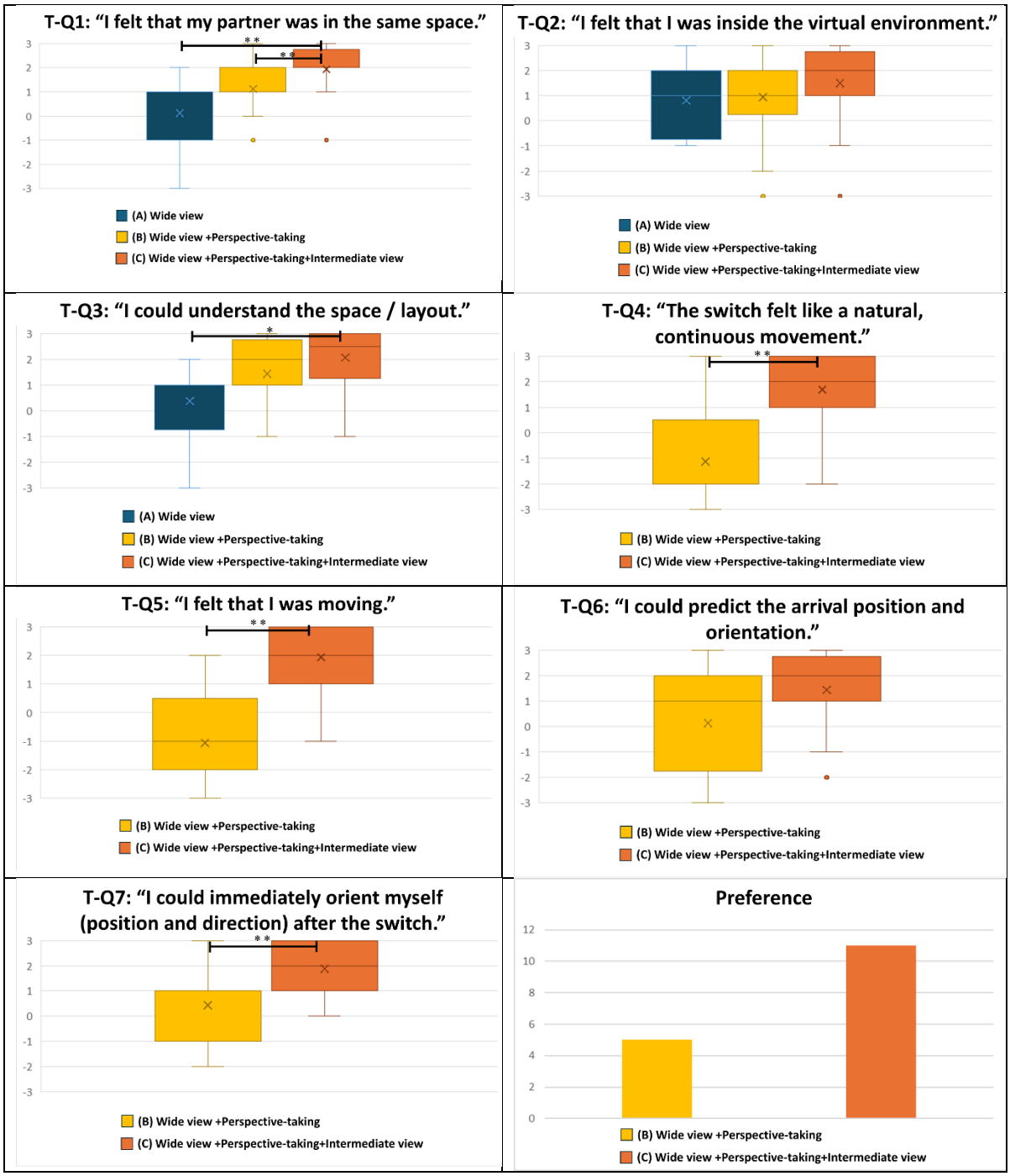


*Figure 14: Subjective ratings of switching experience (T-Q1–T-Q7) and preference between the two switching methods. Box-and-whisker plots summarize 7-point ratings for each item, comparing the instantaneous switch (B) and the intermediate-view switch (C) (and A where applicable). The bar chart shows the number of participants preferring each method. * and ** indicate statistically significant differences after Holm correction (*: $p < 0.05$, **: $p < 0.01$).*

4.3.4. Discussion

Experiment 3-1 demonstrates that adding the partner's perspective provides substantial support for situation recognition (empty-port identification) and overall task comprehension. Reviewing log files for S-Q3 errors revealed that a tendency to stare at the center of the viewport or inappropriate manual switching to the wide-view layer at critical moments caused context loss. This can be resolved by scaling the gaze pointer dynamically or using blinking effects to emphasize focus changes. Statistical analysis for Experiment 3-2 applied a repeated-measures ANOVA across the three conditions (A/B/C) for T-Q1–T-Q3 with Holm post-hoc corrections, and paired t-tests with Holm corrections for T-Q4–T-Q7 between profiles B and C. Significant differences were confirmed for T-Q1, where Condition C scored significantly higher than A and B. T-Q3 metrics showed Condition C was significantly superior to A. For transition metrics, Condition C scored significantly higher than B across T-Q4, T-Q5, and T-Q7, while T-Q6 did not reach statistical divergence. These findings

demonstrate that bridging the spatial gap with intermediate views alleviates spatial discontinuity and orientation confusion, directly contributing to elevated co-presence and overall switching comfort.

4.4. Summary of Experiments

Through the evaluation of RQ1 to RQ3, we investigated how the three core pillars, 'omnidirectional space sharing without field-of-view limits,' 'partner perspective and gaze overlay sharing,' and 'continuous viewpoint switching via intermediate views', contribute to communication comprehension and user experience quality.

In Experiment 1 (RQ1), lifting the field-of-view restrictions through wide-view video proved highly effective for observing body language and interpreting user intent compared to narrow-FOV traditional communication. However, identifying closely related emotional profiles remains a challenge due to the lack of eye-region details caused by the HMD form factor. This limitation can be resolved by integrating emerging HMD technologies that feature external facial expression replication displays. Combining our structural framework with such systems will minimize information loss and improve recognition accuracy for subtle facial and emotional variations.

In Experiment 2 (RQ2), combining the partner's first-person view with a gaze overlay allowed for near-perfect identification of target objects. This indicates that task-level details, which are difficult to extract from a wide-view perspective alone, can be explicitly externalized. Comparative analysis against gaze-free baselines confirmed that explicit visualization minimizes cognitive load, making identification rapid and robust.

In Experiment 3 (RQ3), the integrated presentation mode demonstrated a significant impact on scenario comprehension and presence. While intermediate viewpoints did not directly alter factual text-based score metrics in Experiment 3-1, Experiment 3-2 proved that interpolation dramatically enhances the transition experience, perceived continuity, self-motion tracking, and post-switch orientation. This proves that the three components fulfill distinct and complementary roles: wide-view feeds global context, perspective-gaze tracking isolates task-level details, and intermediate views geometrically link both worlds to preserve presence.

5. Conclusions

This paper introduces an immersive telecommunication framework designed to share perspectives and spatial awareness. By organizing the system into four modular components—Acquisition, Transmission, Integration, and Presentation—we successfully synthesized omnidirectional context sharing, partner-perspective detail tracking, homography-driven IMU video stabilization, and continuous intermediate view transitions. This framework resolves the data fragmentation challenges of traditional single-perspective setups and supports symmetrical, bidirectional scalability.

User studies validated system efficacy across three main dimensions matching RQ1–RQ3. Wide-view streams improved body language visibility and intent interpretation (RQ1). First-person view tracking with real-time gaze replication delivered high-precision focus localization (RQ2). Finally, intermediate viewpoint interpolation minimized spatial discontinuity during switching, significantly improving co-presence, orientation speed, and tracking predictability (RQ3). In conclusion, our system successfully integrates global contextual awareness with close-range perspective-taking, making it a viable solution for remote collaboration, education, and technical support applications.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] DeFilippis, E., Impink, S. M., Singell, M., Polzer, J. T., and Sadun, R., The impact of COVID-19 on digital communication patterns, Humanities and Social Sciences Communications, vol. 9, Article 180, 2022.

[2] Bergmann, R., Rintel, S., Baym, N., Sarkar, A., Borowiec, D., Wong, P., and Sellen, A., Meeting the pandemic: Videoconferencing fatigue and evolving tensions of sociality in enterprise video meetings during COVID-19, Computer Supported Cooperative Work, vol. 32, no. 2, pp. 347–383, 2022.

[3] Meng, W., Yu, L., Liu, C., Pan, N., Pang, X., and Zhu, Y., A systematic review of the effectiveness of online learning in higher education during the COVID-19 pandemic period, Frontiers in Education, vol. 8, Article 1334153, 2024.

[4] Zaman, F., Anslow, C., and Rhee, T. J., Vicarious: Context-aware Viewpoints Selection for Mixed Reality Collaboration, Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology (VRST '23), Article no. 23, pp. 1–11, 2023.

[5] Kasahara, S. and Rekimoto, J., JackIn head: Immersive visual telepresence system with omnidirectional wearable camera for remote collaboration, Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology, pp. 217–225, 2015.

[6] Komiyama, R., Miyaki, T., and Rekimoto, J., JackIn space: designing a seamless transition between first- and third-person view for effective telepresence collaborations, Proceedings of the 8th Augmented Human International Conference, Article no. 14, pp. 1–9, 2017.

[7] Misawa, K. and Rekimoto, J., ChameleonMask: Embodied physical and social telepresence using human surrogates, Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, pp. 401-411, 2015.

[8] Cai, M. and Tanaka, J., Mixed-reality communication system providing shoulder-to-shoulder collaboration, International Journal on Advances in Software, vol. 12, no. 3–4, pp. 332-342. 2019.

[9] Kasahara, S. and Rekimoto, J., JackIn: integrating first-person view with out-of-body vision generation for human–human augmentation, Proceedings of the 5th Augmented Human International Conference, Article no. 46, pp. 1–8, 2014.

[10] VRChat, [Online]. Available: https://hello.vrchat.com/. Accessed: Dec. 10, 2025.

[11] Wei, X., Jin, X., and Fan, M., Communication in Immersive Social Virtual Reality: A Systematic Review of 10 Years' Studies, Proceedings of the Tenth International Symposium of Chinese CHI, pp. 27-37, 2024.

[12] Galinsky, A. D. and Moskowitz, G. B., Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism, Journal of Personality and Social Psychology, vol. 78, no. 4, pp. 708–724, 2000.

[13] Peck, T. C., Seinfeld, S., Aglioti, S. M., and Slater, M., Putting yourself in the skin of a black avatar reduces implicit racial bias, Consciousness and Cognition, vol. 22, no. 3, pp. 779–787, 2013.

[14] Nishida, J., Takatori, H., Sato, K., and Suzuki, K., CHILDHOOD: Wearable suit for augmented child experience, Proceedings of the 2015 Virtual Reality International Conference (VRIC '15), Article no. 22 pp. 1–4, 2015.

[15] Oliveira, E. C., Bertrand, P., Lesur, M. R., Palomo, P., Demarzo, M., Cebolla, A., Baños, R. M., and Tori, R., Virtual Body Swap: A Feasible Tool to Be Explored in Health and Education, Proceedings of the XVIII Symposium on Virtual and Augmented Reality, pp. 81–89, 2016.

[16] Wang, L., Wu, W., Zhou, Z., and Popescu, V., View Splicing for Effective VR Collaboration, Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality, pp. 509–519, 2020.

[17] Jung, S., Wisniewski, P. J., and Hughes, C. E., Limbo: The Effect of Gradual Visual Transition Between Real and Virtual on Virtual Body Ownership Illusion and Presence, Proceedings of the 2018 IEEE Conference on Virtual Reality and 3D User Interfaces, pp. 267–272, 2018.

[18] Unity Technologies, Unity Render Streaming (Version 3.1.0-exp.7) [Software]. Available: https://docs.unity3d.com/ja/Packages/com.unity.renderstreaming@3.1/manual/index.html.

[19] Enox Software, OpenCV for Unity (Version 3.0.1) [Software]. Available: https://assetstore.unity.com/packages/tools/integration/opencv-for-unity-21088.

[20] Wang, G., Wang, P., Chen, Z., Wang, W., Loy, C. C., and Liu, Z., PeRF: Panoramic Neural Radiance Field From a Single Panorama, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 10, pp. 6905–6918, 2024.

[21] 23 Essential Body Language Examples and Their Meanings, [Online]. Available: https://web.archive.org/web/20250125185533/https://www.scienceofpeople.com/body-language-examples/. Accessed: Dec. 10, 2025.

[22] Dael, N., Mortillaro, M., and Scherer, K. R., Emotion expression in body action and posture, Emotion, vol. 12, no. 5, pp. 1085–1101, 2012.

[23] Bindemann, M., Burton, A. M., and Langton, S. R. H., How do eye gaze and facial expression interact?, Visual Cognition, vol. 16, no. 6, pp. 708–733, 2008.