GIST: Group Interaction Sensing Toolkit for Mixed Reality

Diana Romero*
dgromer1@uci.edu
University of California, Irvine
USA

Yasra Chandio*
ychandio@umass.edu
University of Massachusetts Amherst
USA

Fatima M. Anwar
fanwar@umass.edu
University of Massachusetts Amherst
USA

Salma Elmalaki
salma.elmalaki@uci.edu
University of California, Irvine
USA

Abstract

Understanding how teams coordinate, share work, and negotiate roles in immersive environments is critical for designing effective mixed-reality (MR) applications that support real-time collaboration. However, existing methods either rely on external cameras and offline annotation or focus narrowly on single modalities, limiting their validity and applicability. To address this, we present a novel ***group interaction sensing toolkit*** (**GIST**), a deployable system that passively captures multi-modal interaction data, such as speech, gaze, and spatial proximity from commodity MR headset's sensors and automatically derives both *overall static interaction networks* and *dynamic moment-by-moment behavior patterns*. We evaluate **GIST** with a human subject study with 48 participants across 12 four-person groups performing an open-ended image-sorting task in MR. Our analysis shows strong alignment between the identified behavior modes and shifts in interaction network structure, confirming that momentary changes in speech, gaze, and proximity data are observable through the sensor data.

Keywords

Mixed Reality, Sensing, Group Collaboration

1 Introduction

Collaborative Mixed Reality (MR) applications are transforming fields such as surgical planning and training [17], shared 3D architectural walkthroughs [11], remote industrial equipment maintenance guidance [18], and co-creative product prototyping [7], each relying on seamless coordination among multiple users (in real time). However, despite these rich, group-based scenarios, we still lack scalable, real-time methods to observe and analyze how groups interact in immersive technologies such as Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR)[1]. Effective collaboration depends on participants' communication patterns, decision-making, turn-taking, social norm setting, and authority negotiation [16, 28, 33, 34], all of which vary with group size, composition and task complexity [20, 21, 25] and ultimately shape productivity, creativity and performance [46, 48, 58]. While traditional research has examined group behavior in physical settings, the emergence of immersive technologies (AR/VR/MR) offers new frontiers for collaborative work.

MR, in particular, seamlessly integrates digital objects into the physical environment, enabling interaction as if they were physically present. This seamless integration creates collaborative scenarios distinct from VR (fully immersive) or AR (digital overlay), introducing novel social and spatial variables such as spatial awareness [35, 39], presence [27, 61], and multimodal communication [37, 59]. Moreover, inter-, intra-, and multi-user variability [12, 71] demands methods that can capture and interpret these complex, evolving interactions. Previous research has significantly advanced our understanding of collaboration through various methods, including VR/AR studies [8, 45, 70], sociometric badges [29], and external techniques such as motion capture cameras, intrusive wearable sensors, or manual video analysis [69]. However, these approaches cannot be directly applied to the on-device environment of Mixed Reality (MR) headsets [13]. MR uniquely blends physical and virtual worlds, leading to challenges such as occlusions and spatial constraints. In this context, the aforementioned methods are unable to capture crucial in situ cues such as gaze, speech, and motion, which are essential to understand group behavior [1].

In this paper, we address the need for a deeper understanding of group behavior in MR by answering the following research questions (**RQ**): *How can the sensory systems in MR headsets effectively capture group behavior during collaborative tasks? **(RQ1)*** and *What algorithms can process and interpret the data to infer group behavior? **(RQ2)***. To answer these, we introduce ***Group Interaction Sensing Toolkit (GIST)***, the first end-to-end framework that passively senses, aggregates, and clusters multimodal signals from commodity MR headsets to infer group behavior. Using embedded headset sensors, we present **GIST** and make four main contributions:

1. *Passive Multimodal Sensing Pipeline.* We capture and preprocess raw gaze, binaural audio, and motion data using embedded sensors in commodity MR headsets, preserving natural interactions without extra hardware.
2. *Sociogram Aggregation Module.* We construct instantaneous interaction graphs (sociograms) to capture graph-based aggregation of group behavior using domain-informed thresholds to compute structural metrics of group behavior.
3. *Temporal Clustering Module.* We implement unsupervised time-series clustering of dyadic interactions to reveal evolving behavioral phases, without manual annotation.
4. *Empirical Validation.* We deployed **GIST** prototype in a 48-participant (12-group) unconstrained image-sorting study.

*Both authors contributed equally to this research.

[1] VR immerses users in a digital world, AR adds digital information onto the real world, and MR blends the two, allowing digital and physical objects to interact in real-time.

Our evaluation demonstrates the system's stability, interpretability, and ability to uncover meaningful group-level insights through structural and temporal analysis.

2 Related Work

2.1 Group Behavior Sensing in MR/VR

Decades of VR research have revealed how virtual contexts shape social presence, interpersonal dynamics, and collaboration [3, 15, 24], with proteus effects and other frameworks [60], explaining the emergence of group norms in digital spaces. MR adds further complexity by merging real and virtual worlds, introducing novel factors such as altered spatial awareness [35], shifts in presence [27], multimodal communication channels [59], and even changing basic behaviors such as eye contact compared to face-to-face settings [50]. But most MR studies focus on individual-user tasks, designing overlays for walking [32], biking [30, 38] or driving [52, 57] rather than multi-user collaboration. Prior work in immersive environments has utilized multi-modal sensors to instrument group interactions, but often for offline or lab-based analysis. TeamSense, for example, used badges to log speaking time and proximity for VR team cohesion [70], gaze and controller motion have characterized turn-taking in small VR groups [8, 45], and MR prototypes have recorded head pose and hand gestures under lab cameras for task coordination [13]. These approaches yield static, offline snapshots and require extra hardware or manual annotation, leaving a gap in fully automated, in-situ inference of evolving group behavior using only the sensors built into MR headsets.

2.2 Social Signal Processing & Sociometric tools

Ubiquitous computing has shown that simple audio, location, and motion cues can reveal rich social dynamics in co-located groups. Sociometric Badges tracked speaking patterns, turn-taking, and interpersonal distance to predict team cohesion and performance [29, 47], and ambient microphones or Bluetooth beacons have been used to detect synchrony and leadership emergence [8]. Social Network Analysis (SNA) and sociometry then provide systematic methods via *sociograms* to visualize and quantify relationships, roles and subgroup structures based on interaction frequency and strength [44, 65]. These techniques have been applied across domains, from nursing teams [10] to classroom groups [36] and virtual environments (VE), where gaze-based networks reveal leaders and cohesion patterns [2, 67]. While previous systems rely on dedicated hardware or fixed installations, our framework **GIST** brings on-device social-signal processing to MR headsets to compute sociometric indicators in without extra instrumentation.

2.3 Mixed Reality Collaboration Frameworks

Several MR platforms log extensive user activity such as gaze heatmaps, object interactions, and spatial trajectories to support interface evaluation or task coordination [37, 59]. For instance, collaborative design tools stream gaze and annotation events for post-hoc replay [19]. However, these logs remain siloed; they record low-level events but lack integrated models to translate them into social metrics such as participation balance or subgroup cohesion. We fill this gap by using sociometry, generating higher-order representations of group behavior in live MR sessions.

2.4 Beyond Related Work

While prior HCI and ubiquitous-computing studies have demonstrated how sensor-derived features correlate with self-reported engagement or presence, they stop short of delivering a turnkey, on-device analysis framework for MR. In contrast, **GIST** fills this gap with a headset-only platform that extracts objective metrics of group behavior without any external infrastructure. By moving from offline analytics and dedicated hardware to an integrated headset-only framework, our work provides the engineering and MR research communities with the first end-to-end solution for in situ group behavior interpretation.

3 System Design and Implementation

In this section, we present **GIST** to design an in situ, real-time passive group behavior sensing in MR environments. Our system addresses three key challenges: (1) synchronizing data across multiple headsets without any external infrastructure by *multi-modal sensor integration (§3.1)*, (2) extracting rich, meaningful interaction features from commodity headset sensors using *sociogram-based structural analysis (§3.2)*, and (3) unifying both structural and temporal views of group behavior into a cohesive analysis pipeline via *temporal clustering for dynamic pattern discovery (§3.3)*.

Design Motivations and Goals. Analyzing group behavior in MR demands methods that preserve natural interaction while still producing timely, reliable insights. However, traditional setups break immersion and limit real-world deployment. **GIST** therefore relies solely on the headset's built-in sensors to eliminate setup overhead and preserve interaction authenticity, but to maintain the consistency across sensing modalities, we perform *temporal synchronization* for coordinate behaviors such as joint attention and turn-taking with sub-100ms precision using network time protocol (NTP)-based clock alignment [42], essential since gaze convergence events occur within 50-100ms windows. Through lightweight feature extraction and concurrent sensor stream processing, we enable real-time behavioral insights without disrupting the MR experience.

We implement these goals through a modular pipeline as shown in Figure 1. **GIST** capture and synchronize sensor input from multiple headsets with global timestamp alignment; clean and calibrate data to reduce noise; extract dyadic primitives (gaze, speech, proximity); build session-level interaction networks (sociograms) to capture roles and group cohesion; and segment interactions into 8-32 second windows to identify recurring behavioral patterns. This dual-scale architecture provides both a holistic overview via network metrics that reveal leadership roles and participation equity, and a dynamic view through short-window clusters that track attention shifts and conversational transitions, making **GIST** suitable for real-world collaborative MR deployments.

3.1 Passive Sensing Module

GIST exploits a full suite of built-in MR headset sensors. In particular, we are interested in observing information related to three types of interaction, namely conversation via audio, shared attention via gaze, and proximity via position data, as they have shown relevance to unraveling group behavior in various works in the literature. During each MR session, we collect the following data for downstream behavioral modeling.

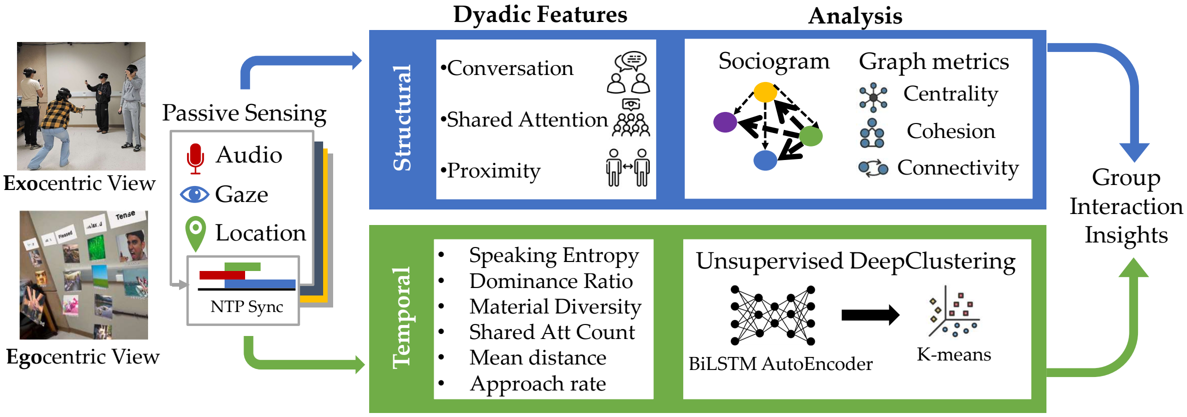

Figure 1: GIST provides behavioral insights at two complementary temporal scales. At the session level, aggregated interaction features construct sociograms capturing gaze, speech, and proximity relationships, analyzed using network metrics for group dynamics assessment. At the temporal level, short-window interaction features are clustered using unsupervised deep learning to identify recurring behavioral patterns, enabling fine-grained temporal segmentation of collaborative activities.

Audio. Ubiquitous sensor studies confirm that speech activities have long been shown to be critical for understanding small group interaction [4, 47, 70], and their patterns shift across platforms from desktop to VR [67]. We track conversation via headset microphones and a lightweight voice-activity detector that timestamps each user's speech segments to capture the verbal participation.

Gaze. Building on extensive work in joint attention and gaze awareness for collaboration [43, 63, 68], we capture shared attention from synchronized gaze rays whenever two users' gaze vectors intersect the same virtual object or region.

Position. We capture physical proximity from six-degree of freedom (6DoF) headset poses (a proxy for social interaction [9, 26]) with spatial coordination recorded whenever participants remain within a close distance of one another.

Together, these synchronized, timestamped streams of audio, gaze, and position data form the basis for both session-level network analysis and fine-grained temporal clustering of dynamic group behavior. Passive data collection occurs entirely during the MR session. Processing and analysis are performed offline, allowing for scalable deployment without affecting the in situ user experience.

3.2 Structural Analysis Module

Sociograms, first popularized in social psychology to visualize group dynamics [44], are network representations that capture social relationships and interaction patterns in groups, providing a concise snapshot of who interacts with whom and how strongly. See Figure 1 for an example of a sociogram. In **GIST**, we automatically generate a sociogram, translating the raw sensor streams into a series of weighted interaction graphs, where each node is a *participant*, and each edge's weight reflects the cumulative interaction strength across audio, gaze, and position modalities. This automated, annotation-free process produces static snapshots of group structure, revealing emergent leaders, subgroup clusters, and overall cohesion.

3.2.1 Feature Selection. We focus on interaction channels from our three main modalities.

First, *conversation patterns* serve as a window into turn-taking dynamics and conversational balance. Beyond simply summing speaking time, we compute measures such as speaking-turn entropy (how evenly the floor is shared) and overlap rate (frequency of interruptions or back-channels). These metrics map onto constructs of dominance and engagement established in small-group studies [47, 54]. Speech segments shorter than 0.5 s are discarded to filter out non-substantive utterances, while longer turns are attributed directionally to capture the speaker–listener asymmetry intrinsic to group dialogue.

Second, *shared attention* reflects the coordination of visual focus, which captures and underpins joint problem solving and mutual awareness. We detect joint fixations by intersecting each user's 3D gaze vector with virtual objects and counting only overlaps that exceed fleeting glances. From these events, we derive edge weights by total duration and fixation frequency, and mean inter-fixation interval, allowing us to distinguish sustained collaboration on a single artifact from rapid visual shifts across multiple items.

Third, *physical proximity* captures embodied aspects of collaboration that foster informal communication and trust [26, 64]. Using six-DoF head pose data, we record both the cumulative time dyads spend within a distance threshold and their approach-and-withdraw dynamics, quantified as the rate of change in inter-headset distance. These features differentiate static closeness, such as huddling around a shared task, from dynamic movement patterns that signal pacing, turn-taking movement, or opportunistic side conversations.

We omit gesture and facial-expression cues because capturing them with sufficient fidelity would require external cameras, which are incompatible with our untethered MR deployment.

3.2.2 Modality-Specific and Fused Sociogram Construction. To capture the multi-dimensional nature of MR collaboration coherently, we construct a separate sociogram for each modality rather than scattering metrics across them. This is to both isolate behaviors that could otherwise be conflated (for example, participants who remain physically close yet do not speak) and to keep the analysis operational when a sensor stream is temporarily unavailable. For each session, we generate a sociogram per interaction channel (conversation, shared attention, and proximity). We then merge the three modality-specific sociograms into a *fused multimodal sociogram*, providing a comprehensive view of group interactions.

- *Temporal scope.* We produce sociograms at two temporal scales, one covering the entire session *and* another using sliding 32 s windows with a 16 s stride. This 32 s window is enough observations for stable edge-weight estimates, and the 16 s overlap captures sub-minute transitions in group dynamics, consistent with prior conversational windows and aligning with our temporal clustering module.
- *Edge definition.* In each window, we assign edge weights based on the total *duration* of interaction. For conversation, the sum of spoken time; for shared attention, the overlap of gaze fixations; and for proximity, the accumulated intervals of co-presence within the defined distance threshold. Longer, sustained interactions, therefore, contribute more heavily than brief encounters. To capture directional dynamics, conversation graphs are treated as directed, reflecting the asymmetry between speaker and listener observed in our pilot study. Attention and proximity networks remain undirected, since those forms of engagement are inherently mutual.
- *Pre-processing and Thresholds* Prior to graph construction, we apply modality-specific thresholds grounded in our empirical studies. Between two users, gaze overlaps must last at least 13 ms, matching the lower bound of visual processing latency [49] to consider them jointly attentive. Proximity events require headsets of users to remain within 1.5 ft, corresponding to Hall's intimate-distance zone [22]; and conversation edges aggregate only speech segments of 0.5 s or longer to filter out breaths and back-channels while preserving monosyllabic utterances [56].
- *Fused graph.* After building three separate sociograms, we normalize each adjacency matrix and fuse them into a single multimodal network using PCA-derived weights so that each channel contributes proportionally while retaining the directed nature of conversational ties (see §3.2.4).

3.2.3 Network Metrics for Behavioral Insights. To translate raw sociograms into interpretable group-level insights, we convert each network metric into a three-tiered scale, based on either fixed thresholds or session-relative percentiles, ***high****: top 20% of values (or* $z \geq +1$*),* ***medium****: middle 40%, and* ***low****: bottom 40% (or* $z \leq -1$*).* For metrics naturally bounded in $[0, 1]$, we define *low* ≤ 0.29, *medium* $0.30-0.59$, *high* ≥ 0.60; adjusting the *high* to ≥ 0.50 when $n \leq 4$). We apply this classification to the three categories of metrics across each modality-specific and fused sociogram.

Centrality measures highlight participants who act as leaders or brokers within conversation and shared-attention networks. *Cohesion* metrics capture how tightly connected subgroups are, whether through physical proximity or mutual gaze. Finally, *connectivity* in the conversation graph is quantified via reciprocity, indicating the balance of two-way exchanges. Together, these metrics provide both a static snapshot of the group structure and a means to track how roles, subgrouping, and engagement evolve over time. In Table 1, we summarize how each metric is mapped onto interpretable aspects of group behavior. We compute these measures on the directed conversation graphs and the undirected attention and proximity sociograms, and then we apply them to the fused multimodal sociogram to capture combined interaction effects.

3.2.4 Implementation Details. For each sliding window $[t_0, t_1]$, we maintain three $N \times N$ adjacency matrices $W^{(\text{conv})}, W^{(\text{att})}, W^{(\text{prox})}$, indexed by participant pairs. Upon each update:

In the **conversation** graph, we initialize $W^{(\text{conv})} \leftarrow 0$ at the start of each window. For every speech segment $(s_{\text{start}}, s_{\text{end}}, p)$ overlapping $[t_0, t_1]$, we compute $\Delta = \min(s_{\text{end}}, t_1) - \max(s_{\text{start}}, t_0)$. If $\Delta \geq 0.5$, we capture the total time participant p spoke to every other member q as: $W_{pq}^{(\text{conv})} \mathrel{+}= \Delta$

For **attention**, we clip each user's gaze intervals to the window $[t_0, t_1]$ and, for each unordered pair (i, j), sum all overlapping gaze durations $\delta \geq 13$ ms, so that repeated joint fixations accumulate into a stronger undirected tie. We set: $W_{ij}^{(\text{att})} = W_{ji}^{(\text{att})} = \sum \delta$

For **proximity**, we align headset poses at common timestamps in $[t_0, t_1]$ and treat each inter-sample interval Δt as a unit of time. Whenever the pairwise distance $d_{ij} \leq 1.5$ ft, we increment: $W_{ij}^{(\text{prox})} = W_{ji}^{(\text{prox})} + \Delta t$

Finally, we fuse these three modality-specific matrices into a single multimodal *fused* adjacency matrix, where $\{\alpha_m\}$ are principal component analysis (PCA)-derived weights summing to 1 so that each channel contributes proportionally as:

$$W^{(\text{fused})} = \sum_{m \in \{\text{conv}, \text{att}, \text{prox}\}} \alpha_m \, W^{(m)}$$

All thresholds (0.5s for speech, 13 ms for gaze overlap, 1.5 ft for proximity) are configurable, as are window length and stride. Each execution yields an instantaneous sociogram ready for metric computation or logging.

3.3 Temporal Clustering Module

Sociograms offer a static snapshot of group interactions over full sessions or windowed aggregates, but they can not detect when particular interaction patterns emerge or dissolve. Our **Temporal Clustering Module** addresses this by segmenting each session's dyadic behavioral features (such as balanced engagement, leader-driven dialogue or disengagement) into temporal phases via unsupervised clustering of time-series features.

3.3.1 Feature Selection and Construction. To focus on the most informative signals (*at the moment*), we first standardize and filter our initial pool of over 20 dyadic features of moment-to-moment interaction between participants extracted on a 1*s* window, removing those with low variance, high pairwise correlation ($r \geq 0.95$) or minimal impact on clustering quality as determined by silhouette-based importance ranking. This pruning yields a concise feature representation spanning the following core behavior dimensions:

Verbal dynamics are captured via *entropy_speaking*, which quantifies unpredictability in turn-taking, and *dominance_ratio*, the imbalance in total speaking time per dyad.

Interaction diversity is measured with *material_diversity*, which measures how many distinct object pairs jointly attend to, reflecting the richness of shared focus.

Proximity is captured by features that characterize static closeness and dynamic movement patterns. *dist_mean* is average dyadic

Table 1: Interpretation of network metrics by interaction mode. Conv = conversation, Att = shared attention, Prox = proximity.

Modality	Metric (Ref.)	High Value Interpretation	Low Value Interpretation
Centrality metrics identify potential leaders and information brokers in conversation and attention networks.			
Conv, Att, Prox	Eigenvector [14]	Connected to other highly central participants	Linked mainly to peripheral participants
Cohesion metrics quantify bonding and tightness in proximity and attention networks.			
Att, Prox	Avg. Clustering Coef. [66]	High values indicate that a node's neighbors are densely interconnected, reflecting tight local subgroup cohesion	Low values indicate sparse neighbor connections, reflecting weak local cohesion
Conv, Att, Prox	Density [55]	High values signify a well-connected network with active group engagement	Low values signify a fragmented or minimally interacting group
Connectivity metrics assess the balance of two-way exchanges in the conversation network.			
Conv	Reciprocity [23]	Balanced two-way exchanges (dialogue)	Predominantly one-way communication

distance, *prox_binary* measures co-presence within 1.5ft, and *approach_rate* to capture speed of movement toward or away. **Shared attention** (*shared_att_cnt*) counts the number of joint gaze fixations on the same virtual object, indicating peaks of mutual engagement.

All features are computed in 1s windows, z-normalized per dyad, and aligned to a uniform temporal grid. This reduced feature set captures both transient and sustained collaboration signals, providing high-resolution temporal segmentation while remaining computationally minimal.

3.3.2 Sequence Encoding and Model Architecture. We treat each dyad's interaction over a segment as a $T \times F$ matrix, where T is the number of 1s windows per segment and $F = 7$ is the number of retained features. We determine both T and the stride S via grid search, optimizing for cluster coherence on held-out data; in practice, we use partial overlap ($S < T$) to balance temporal resolution and embedding stability. Each sequence is processed by a deep clustering architecture: a convolutional-recurrent autoencoder implemented in PyTorch. Two 1-D convolutional layers (with kernel sizes and filter counts selected via grid search) followed by ReLU activations and max-pooling extract local temporal motifs. A bidirectional LSTM then ingests the pooled features, producing a fixed-length latent vector. The decoder mirrors this architecture, upsampling and LSTM layers reconstruct the original sequence. Hyperparameters such as convolutional kernel dimensions, LSTM hidden state size, and dropout rate are tuned to maximize silhouette scores while maintaining low reconstruction error.

3.3.3 Clustering and Loss Function. The trained encoder maps each segment to a latent embedding, which we cluster using K-Means. To jointly optimize embeddings, both reconstruct their inputs accurately and form tight, well-separated clusters, we minimize the composite loss as: $\mathcal{L} = (1-\lambda)\ \mathcal{L}_{\text{rec}} + \lambda\ \mathcal{L}_{\text{clu}}$, where $\mathcal{L}_{\text{rec}}$ is the mean squared reconstruction loss and $\mathcal{L}_{\text{clu}}$ is the squared Euclidean distance to the assigned cluster centroid. We sweep cluster weight (λ) in $[0.3, 0.7]$, choosing the value that yields the best trade-off between silhouette score and reconstruction fidelity. When the number of clusters k is unspecified, we apply a stability-informed elbow criterion, combining within-cluster inertia with cross-run adjusted Rand index (ARI) consistency to select k.

3.3.4 Window Length and Stride Selection. To trade off reconstruction accuracy against cluster coherence, we evaluate three $< window : stride >$ combinations ($< 8s : 4s >$, $< 16s : 8s >$ and $< 32s : 16s >$). Longer windows yielded slightly higher reconstruction error but notably better silhouette scores, which plateaued beyond 32s. Increasing the stride to 16 s further improved cluster separation by reducing overlap, with only a marginal impact on loss. We therefore use a 32s window and 16s stride throughout, which also aligns our temporal clusters with the sociogram snapshots and guarantees at least 30s of data per network for stable metric estimation.

3.3.5 Implementation and Output. All heavy computation, such as feature extraction, encoding, and clustering, is performed offline after data collection. To scale across many dyads or lengthy sessions, these tasks run in parallel, and a *fast evaluation mode* subsamples up to 5000 windows, cutting runtime from $\sim 30min$ to under $5min$ with negligible quality loss. The module outputs a cluster label for each dyadic window, which can be rendered as phase-aligned timelines or heatmaps, offering a time-resolved map of how group behavioral patterns evolve.

4 Deployment and Study Setup

4.1 Participants

We recruited 48 participants (12 groups of 4; 36 male, 8 female; age mean (μ) = 24.2, standard deviation (SD) = 4.7). Pre-study demographic questionnaires measured prior immersive-tech experience on 7-point Likert scales: MR ($\mu = 1.8, SD = 1.2$), AR ($\mu = 3.1, SD = 1.8$), VR ($\mu = 3.4, SD = 2.1$). We fixed the group size at four to maximize the number of dyadic interactions (six per group) while keeping computation tractable [62].

4.2 Materials

We conducted the study in `10ft` × `5ft` space where participants navigated and collaborated in close quarters (cleared of materials to minimize distractions). Each participant used a Meta Quest Pro headset [41], which captured and streamed eye gaze, binaural audio, and 6DoF pose data over our local Wi-Fi network. We synchronized all devices with NTP ($< 50ms$ offset) and built the collaborative MR app in Unity with the Meta XR SDK [40]. By aligning every virtual object in each user's coordinate frame, we guaranteed a shared reference without any additional prompts, cues, or enforced turn-taking.

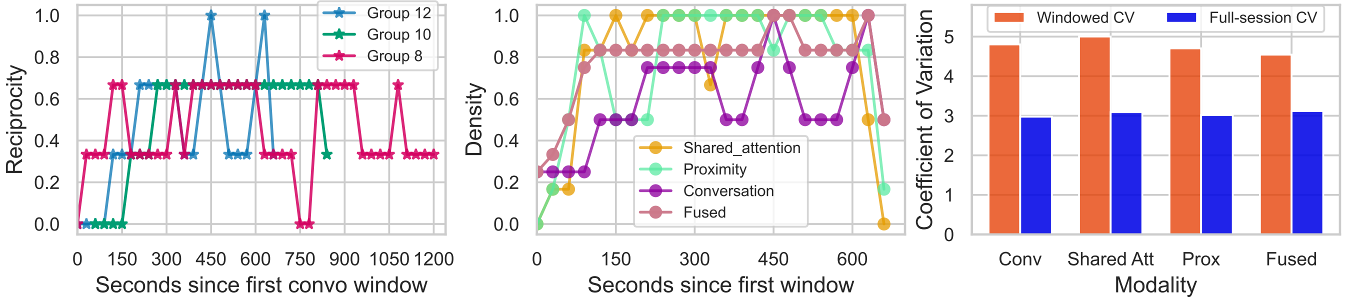

Figure 2: Windowed-session analysis to assess behavioral patterns obscured by session-level aggregation. Conversation reciprocity for Groups 8, 10, 12 **(left), Group 12's multimodal density trajectory (center), and density variation (right).**

4.3 Collaborative Task

We asked each group to sort 28 OASIS images [31] (prevalidated for pleasantness and arousal and free of graphic content) into six affective labels (angry, bored, relaxed, tense, pleased, frustrated) based on Russell's circumplex model [53]. The images lay scattered throughout the virtual room, with floating *plates* labeled for each emotion hovering nearby. Without any time limit or scripted turns, participants freely approached and grabbed images, moved them to their chosen plates, and negotiated assignments through open-ended discussion. To sort an image, participants used a natural point-and-drag motion with their Meta Quest Pro controllers. They aimed at an image, held the grip button to *pick it up*, guided it toward the desired emotion plate, and released the button to lock it in place. Images only attach when positioned sufficiently close to a label, providing immediate visual confirmation. Only one person can manipulate a given image at a time, but different participants may simultaneously move other images within reach, mirroring the physical act of picking up and placing objects. This unstructured setting, where teams self-direct by clustering around images of interest, encourages natural decision-making, communication, and alignment as group members iteratively build consensus on each label [5, 6, 51]. On average, groups completed the task in 32.4 minutes (SD = 8.4).

4.4 Procedure

Upon arrival, participants reviewed an IRB-approved information sheet and provided verbal consent, then completed a brief demographics survey. We handed out Meta Quest Pro headsets, guided each person through focus and fit calibration, and ran a short tutorial using two practice images and categories to teach the grab–drag–release interaction and category placement in MR. Next, groups tackled the main task, sorting 28 images into 6 emotion categories. We instructed them to *work together to categorize these images by emotion, discuss and reach agreement on each label*, with no time limit or performance feedback. Participants self-directed their collaboration, moving freely around the space and negotiating assignments until everyone confirmed consensus. After they finished sorting, the session concluded, and participants returned their headsets. The entire session, including setup, training, task, and wrap-up, took under 35 – 45 minutes.

4.5 Data Collection and Ground Truth

We captured synchronized multimodal data passively from each headset and ran it through **GIST**'s end-to-end pipeline to infer group behavior. To validate these automated inferences, we first assessed clustering stability and checked that sociogram metrics remained coherent across time and groups. *We deliberately omitted subjective collaboration surveys; such ratings cannot reliably capture the dynamic, moment-to-moment shifts our system targets and would not align with its end-to-end, automated design.* Therefore, we manually validated our results against time-aligned passively collected egocentric video from each headset, confirming the system corresponded to actual behaviors on camera; this served as our ground truth. This multi-layered validation demonstrates **GIST** 's robustness and scalability for passive, real-world MR group settings.

5 Results

5.1 Structural Analysis of Group Behavior

5.1.1 Windowed vs. Session-Level Sociograms. To see how temporal granularity shapes our structural insights, we first asked whether slicing full sessions into short, overlapping windows would uncover interaction patterns that a single aggregated sociogram misses. In Figure 2 we compared 32-minutes full-session sociograms against sliding-window sociograms (32s windows, 16s stride; $N = 12$). On the left, *conversation reciprocity* traces for Groups 8, 10, and 12 show Group 8 moves from one-sided turns (0) to balanced exchange (~ 0.67), Group 10 oscillates widely from 0.11 to 0.85 (reciprocity $\mu = 0.45, SD = 0.27$) revealing alternating episodes of symmetry and dominance rather than a single persistent leader, and Group 12 gradually rebalances its dialogue. In contrast, a session-level reciprocity of 0.97 (SD 0.10) barely hints at any variation.

In the center of Figure 2, we show Group 12's PCA-fused *multimodal density over time* and density over time for each modality. We applied PCA to the z-scored edge-weight for each modality across all 12 dyads. The first principal component captured 54% of the total variance, drawing almost equally from proximity (loading = 0.708) and attention (0.706) but hardly at all from conversation (0.025). For example, in one pair the PCA-fused edge weight peaked at 504 (in normalized units), while the original conversation tie retained a directional weight of just 3.1 from B⟶A, demonstrating that although PCA fusion summarizes overall interaction volume

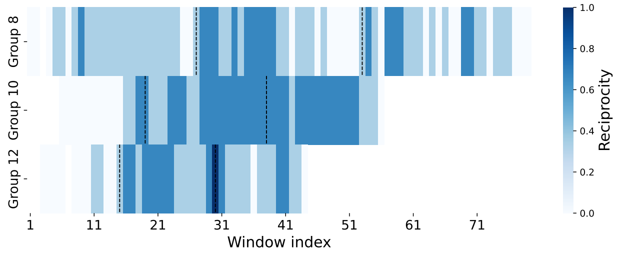

Figure 3: Windowed conversation reciprocity heat maps for Groups 8, 10, and 12. Each row represents a 32 s window; darker shades denote higher reciprocity on a $[0.00, 1.00]$ scale. Dashed lines mark early/middle/late thirds.

effectively, it can wash out the very speech-based asymmetries needed to detect leadership or dominance.

Next, to isolate each modality's influence on the fused sociogram, we performed leave-one-out ablation experiments. We compare the coefficient of variation in density between full-session and windowed analyses by calculating each condition's coefficient of variation across all groups and modalities as shown in Figure 2 (right). Removing conversation edges left overall density unchanged but wholly inverted the ranking of the strongest ties (Spearman $\rho = -0.20$, $p < 0.05$), indicating that directional speech cues drive relational importance. Excluding proximity produced a moderate reshuffling ($\rho = 0.60$, $p < 0.01$), while dropping shared attention had almost no effect ($\rho = 1.00$, n.s.). These results confirm that **GIST** PCA-based fusion effectively summarizes total interaction volume but under-weights the critical speech-based asymmetries needed to detect leadership and dominance. Accordingly, we recommend fused graphs for lightweight, real-time coordination monitoring and conversation-specific or fused graphs augmented with directional features, for post-hoc role analysis.

5.1.2 Conversation Reciprocity Over Time. To track how speaking turns balance or skew over the course of a session, we computed bidirectional conversation reciprocity (the proportion of mutual exchanges out of all directed exchanges) within 32s windows stepped every 16s (12 groups total). Heatmaps for Groups 8, 10, and 12 in Figure 3 show that despite near-maximal session-level density, reciprocity patterns diverge markedly. An ANOVA across early, middle, and late thirds confirms this shift ($F(2, 33) = 8.45, p < 0.001$). Group 8's reciprocity climbs from 0.00 to 0.67, showing a transition from one-sided monologues to balanced dialogue; Group 10 oscillates between 0.11 and 0.85 indicating alternating bouts of dominance rather than a single persistent leader.; and Group 12 spans the full $0 - 1$ range but drifts upward, ending with a third-period mean of 0.35 signaling partial rebalancing. These temporal trends demonstrate **GIST** 's capacity to pinpoint moments of leadership emergence and shifts in dialogue equity.

5.2 Temporal Clustering of Dyadic Interactions

To capture how interaction dynamics evolve, we evaluate the behavioral patterns uncovered by our temporal clustering pipeline.

5.2.1 Latent Embedding and Cluster Selection. We divided each session into 32s windows with an 16s stride, yielding 71404 dyadic segments. Each segment was encoded by a three-layer convolutional-recurrent autoencoder (latent dimension = 16). We then applied K-means ($k = 2-10$, 20 restarts) to the learned embeddings. An inertia elbow, a silhouette score of 0.87, and an $ARI > 0.8$ all pointed to $k = 4$ clusters. Figure 4 projects our 16-D embeddings into 2D using UMAP, revealing four clearly separated clusters. Their relative sizes, 44.7%, 34.0%, 15.4%, and just 5.8% for clusters 0, 1, 2, 3, respectively, show that high-energy co-manipulation is rare, whereas rhythmic leadership and monotone focus dominate.

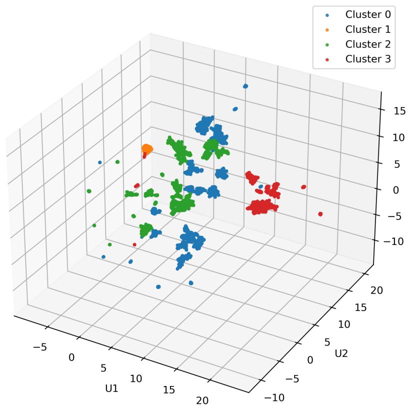

Figure 4: 3D UMAP of 71404 window embeddings. Colors denote clusters; distinct manifolds confirm high silhouette quality.

5.2.2 Cluster Characterization. Now we link each cluster to a distinct behavioral pattern by inspecting its z-scored feature profile (Figure 5). Cluster 0 shows high *dominance_ratio*, low *speaking_entropy*, and minimal proximity, indicative of structured, turn-based leadership with predictable dialogue pacing. Cluster 1 combines elevated *speaking_entropy*, frequent *shared_att_cnt* events, and close proximity to capture animated, synchronous co-manipulation marked by rapid speech and movement fluctuations. Cluster 2 features low *material_diversity* and muted speech dynamics, reflecting a narrow, repetitive task focus with balanced but monotone interaction. Finally, Cluster 3 pairs high *material_diversity* with low *dominance_ratio*, embodying instructor-style demonstrations in which one participant explores varied content while others observe.

5.2.3 Generalizability Across Dyads. We further validated our rule hierarchy using Shapley additive explanations (SHAP) values, which quantify each feature's contribution to individual cluster predictions. We distilled each behavioral motif into a clear decision hierarchy using a surrogate decision tree, with SHAP values confirming the relative importance of each feature. In practice, any 32 s window with $speaking\_entropy > 1.2$ is labeled as Cluster 1 (animated collaboration); if not, a $dominance\_ratio < -0.7$ assigns it to Cluster 3 (instructor-style behavior); failing that, $material\_diversity < -0.3$ indicates Cluster 2 (monotone focus); all remaining segments fall

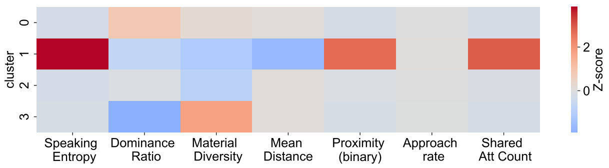

Figure 5: Heatmap of clusters (rows) vs. features (columns), where color intensity shows each feature's deviation from its mean.

Table 2: Entropy of cluster membership across groups, pairs, and actors (lower values = more context-specific).

Cluster	Group Ent.	Pair Ent.	Actor Ent.
0 (Rhythmic Leader–Follower)	**1.23**	2.91	2.42
1 (Animated Collaboration)	2.47	3.41	2.85
2 (Monotone Focus)	2.10	2.68	2.21
3 (Instructor Demonstration)	2.60	**3.73**	2.88

into Cluster 0 (rhythmic leadership). This sequence prioritizes conversational unpredictability first, then leadership asymmetry, and finally task variety.

To understand how these motifs distribute across teams and individuals, we computed the categorical entropy of cluster membership at the group, pair, and actor levels in Table 2. A low entropy score signifies a behavior that is tightly tied to specific contexts, while a high score points to a widely shared interaction style. Cluster 0 shows the lowest group entropy (1.23), indicating that rhythmic leader–follower patterns tend to be team-specific. By contrast, Cluster 1's high entropy across all three levels confirms that animated collaboration is a universally occurring motif. Clusters 2 and 3 occupy intermediate positions, revealing a blend of context-sensitive and broadly shared behaviors. These findings demonstrate that our temporal clustering identifies distinct interaction phases and captures their generalizability across diverse MR group settings.

Manual Validation: To assess our clustering labels against human judgment, we randomly sampled 100 windows (evenly across the four predicted clusters, and balanced by group) to match available coding resources. Overall, manual labels agreed with our automated assignments in 71 cases (71% accuracy). Performance was consistent for Clusters 0–2. Precision ranged from 0.75 to 0.84, recall from 0.64 to 0.73, and $F1 \approx 0.72$. In contrast, Cluster 3 (expert demonstration) achieved high recall (0.92) but lower precision (0.48), indicating that the model often over–predicted this state. Most misclassifications involved false positives for Cluster 3 or confusions between Clusters 0 and 2, whose feature profiles overlap partially. The macro-averaged metrics precision 0.71, recall 0.74, and $F1 \approx 0.70$ confirm broadly uniform performance across classes. These results suggest that the pipeline reliably identifies the three dominant collaboration modes, while further refinement is needed to reduce false positives before using expert-demonstration labels for real-time adaptation.

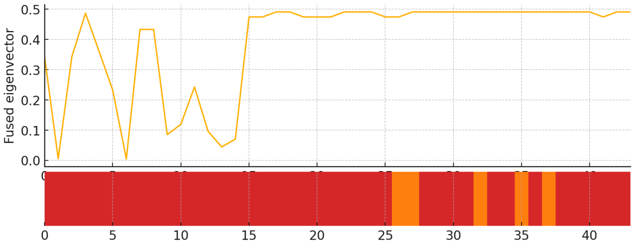

Figure 6: Fused eigenvector over time for one pair in Group 10, with the transition from Cluster 3 to Cluster 1 highlighted.

5.3 Structural vs. Temporal Alignment

5.3.1 Alignment of Clusters and Network Metrics. To verify that our temporal and structural analyses capture the same collaborative phenomena from different angles, we cross-tabulated the four behavioral clusters against tertile-binned network metrics (low/medium/high) over 131 windows. After confirming chi-square (χ^2) assumptions, five metrics showed significant associations (Table 3). Conversation reciprocity exhibited the strongest link ($\chi^2 = 12.73, p = 0.0475, V = 0.49$), with animated collaboration (Cluster 1) and expert demonstration (Cluster 3) over-represented in the highest-reciprocity tertile. Eigenvector centrality also aligned moderately with cluster membership across conversation ($V = 0.26$), fused ($V = 0.24$), proximity ($V = 0.24$), and attention ($V = 0.23$) graphs (all $p < 0.05$), indicating that bursts of centrality in the sociogram coincide with the high-energy and demonstration modes. In contrast, simpler measures like density and clustering coefficient did not vary by cluster, suggesting that higher-order network metrics are more sensitive to shifts in group behavior.

5.3.2 Illustrative Case Study: Lead–Lag Dynamics. In Figure 6 we overlay Group 10's fused eigen-vector trajectory with the cluster label for a single illustrative pair. For the first fifteen windows the pair stays in Cluster 3, which we interpret as an explanatory or demonstration mode; during the same span the group-level centrality meanders at modest values, dipping to its session minimum (≈ 0.01) and never rising above ≈ 0.34. A decisive switch occurs at window 16, where the pair enters Cluster 1, the animated-collaboration state, and the eigenvector score surges past 0.47. Centrality then plateaus near the session maximum (≈ 0.49) for more than six minutes while the dyad remains in Cluster 1. Brief returns to Cluster 3 later in the session (e.g., windows 42 and 56) are accompanied by proportional drops to the $0.43 - 0.45$ range, illustrating how even momentary shifts back to turn-taking discourse redistribute network influence. The tight temporal coupling between the dyad's cluster state and the fused centrality confirms that **GIST** flags changes in interaction style with window-level precision.

6 Discussion

Our evaluation demonstrates that **GIST** 's dual-approach of combining windowed sociograms with temporal clustering effectively uncovers group behavior that static or unimodal approaches would miss. By slicing sessions into overlapping 32s windows, we revealed clear shifts in speaking equity, spatial cohesion, and attention alignment that full-session aggregation flattens. For instance, Group 8's

Table 3: Structural metrics whose low/medium/high tertile distributions vary across clusters ($N = 131$ windows).

Metric (binned)	χ^2	p	Cramér's V
Conversation reciprocity	12.73	0.0475	0.49
Conversation eigenvector	16.82	0.0100	0.26
Fused eigenvector	14.14	0.0281	0.24
Proximity eigenvector	14.92	0.0209	0.24
Attention eigenvector	14.13	0.0282	0.23

transition from monologue to balanced dialogue, Group 10's alternating bouts of dominance, and Group 12's gradual rebalancing were all invisible in single, session-wide sociograms but became evident through sliding-window reciprocity (Figure 2) and multimodal density trajectories (Figure 3).

Our leave-one-out ablations showed that while PCA-fused sociograms summarize overall interaction volume, they risk washing out speech-based asymmetries crucial for detecting leadership; omitting conversation ties entirely inverted tie rankings (Spearman $\rho = -0.20$), whereas removing proximity or attention had far smaller effects. This finding suggests a two-track monitoring strategy: fused graphs for lightweight, real-time coordination signals, and directed conversation graphs (or fusion augmented with speech asymmetry) for post-hoc analyses of influence and dominance.

The temporal clustering module further characterizes collaboration at the process level, distilling 71404 dyadic windows into four interpretable motifs (rhythmic leadership, animated co-manipulation, monotone focus, and expert demonstration) that together cover the breadth of interaction styles. A compact decision hierarchy based on speaking entropy, dominance ratio, and material diversity recovers these motifs with 71% accuracy against manual coding. Entropy analyses reveal that while some behaviors (animated collaboration) generalize widely across teams, others (rhythmic leader–follower) remain group-specific, highlighting where adaptive support should be tailored versus where generic cues suffice.

By aligning cluster labels with tertiled network metrics, we confirmed that high-reciprocity and elevated eigenvector-centrality states co-occur with animated and demonstration phases (Table 3), whereas simple density and clustering coefficients remain blind to these shifts. A case study of Group 10's fused eigenvector trace shows centrality spikes reliably precede bursts of high-energy collaboration (Figure 6), illustrating how structural flags can anticipate emergent interaction modes.

6.1 Implications & Practical Recommendations

By uniting structural sociograms with temporal clustering, **GIST** delivers actionable group analytics in MR, transforming raw sensor streams into insights that can drive adaptive collaboration support. First, our session-level sociograms reliably detect shifts in group roles. Conversation reciprocity ($V = 0.49$) and eigenvector centrality ($V \approx 0.24$ across modalities) track changes in influence distribution spikes in these metrics, flagging impending leadership hand-offs. In contrast, simpler measures like density and clustering remain constant, underscoring the importance of directional and centrality features for real-time role monitoring.

Second, the four 32-second behavioral *micro-states* uncovered by our clustering pipeline are both compact and interpretable ($silhouette = 0.87, ARI > 0.80$). A shallow decision tree using $speaking\_entropy$, $dominance\_ratio$, and $material\_diversity$ reproduces 91% of cluster assignments. These clear, human-readable cues directly map onto MR design knobs, such as prompting floor control when dominance spikes or suggesting task variety when monotony sets in, without human intervention.

Third, the complementary strengths of each scale enable proactive, rather than reactive, interventions. Structural metrics consistently rise about 32 s before a cluster transition (fused centrality climbs just before animated collaboration begins), offering an early warning system. A three-step workflow: (1) flag potential change via sociogram metrics, (2) confirm the new state through clustering, and (3) trigger an adaptive response (open a shared annotation panel), can support seamless, context-aware MR experiences.

Finally, our joint analysis yields four design-relevant insights: leadership behavior oscillates on the order of sub-minutes; balanced turn-taking predicts engaging collaboration; monotony can be detected in real time through low task diversity and speaking entropy; and while some interaction modes (like animated co-manipulation) generalize across teams, others (such as rhythmic leader–follower) are team-specific.

Design implications. In practice, fused sociograms from **GIST** offer a lightweight, live view of overall coordination intensity, ideal for immediate adaptation, whereas conversation-specific graphs and temporal cluster labels support richer post-hoc reflection, personalized feedback, and leadership coaching. This dual-purpose approach lets developers tailor MR collaboration tools to their application's latency, interpretability, and analytic depth requirements.

6.2 Limitations and Future Work

Our evaluation of **GIST** focused on 4-person image sorting in a controlled lab environment. Different tasks, larger groups, or more dynamic spatial settings may introduce new interaction modes (side conversations, head-gestures) or sensor challenges (occlusion). Moreover, while our temporal clustering reliably distinguishes three dominant motifs, it over-assigns the infrequent state (expert demonstration); boosting its precision through class-weighted training, data augmentation of under-represented patterns, incorporating gesture recognition, or the addition of task-specific features (tool-use gestures) and leveraging additional semantic cues will be essential before live deployment.

All analyses currently run offline; real-time clustering and metric computation would require further optimization and efficient buffering of sensor streams. Our PCA-based fusion effectively summarizes overall interaction volume but under-weights directional speech asymmetries; exploring alternative fusion techniques (multi-view graph learning) could better preserve both volume and directionality. Integrating lightweight speech-to-intent analysis (while respecting privacy) could help disambiguate instructional monologues from narrative commentary.

We deliberately omitted subjective collaboration surveys, as moment-to-moment behavioral shifts are poorly captured by end-of-session ratings. Future work should include user studies that integrate **GIST** with adaptive MR features (dynamic floor-control prompts) to measure their impact on team coordination, decision-making, and subjective experience in real-world remote and co-located settings. Finally, controlled A/B experiments comparing static versus adaptive MR conditions will be critical to demonstrate the **GIST** sensing platform's technical robustness and its human-centered benefits in authentic, open-ended collaboration scenarios.

7 Conclusion

We presented **GIST**, the first end-to-end platform for passive, in situ sensing of group behavior in MR. By combining lightweight, headset-only multimodal sensing with graph-based structural analysis and deep temporal clustering, our system uncovers both who drives collaboration and how interaction patterns evolve moment-to-moment. Through a 48-participant (12-group) user study, we demonstrated its ability to reveal leadership rhythms, dialogue balance, and distinct collaboration modes that static or offline methods miss. **GIST** bridges the gap between raw MR sensor logs and actionable social metrics, laying the groundwork for adaptive, socially aware MR applications that can support teamwork in real time.

Acknowledgments

This work is supported by the U.S. National Science Foundation (NSF) under grant number 2339266 and 2237485.

References

[1] Huidong Bai, Prasanth Sasikumar, Jing Yang, and Mark Billinghurst. 2020. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[2] Jeremy N Bailenson, Andrew C Beall, Jack Loomis, Jim Blascovich, and Matthew Turk. 2004. Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence: Teleoperators & Virtual Environments* 13, 4 (2004), 428–441.

[3] Jeremy N Bailenson, Nick Yee, Jim Blascovich, Andrew C Beall, Nicole Lundblad, and Michael Jin. 2008. The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *The journal of the learning sciences* 17, 1 (2008), 102–141.

[4] Robert F Bales. 1950. Interaction process analysis; a method for the study of small groups. (1950).

[5] Noemi Berlin, Mamadou Gueye, and Stéphanie Monjon. 2024. Feedback and Cooperation: An Experiment in Sorting Behaviour. doi:10.2139/ssrn.4718101

[6] Maria Friis Bjerre. 2015. Card Sorting as Collaborative Method for User-Driven Information Organizing on a Website: Recommendations for Running Collaborative Group Card Sorts in Practice. *Communication & Language at Work* 4, 4 (May 2015), 74–87. doi:10.7146/claw.v1i4.20773 Number: 4.

[7] Gaetano Cascini, Jamie O'Hare, Elies Dekoninck, Niccolo Becattini, Jean-François Boujut, Fatma Ben Guefrache, Iacopo Carli, Giandomenico Caruso, Lorenzo Giunta, and Federico Morosi. 2020. Exploring the use of AR technology for co-creative product and packaging design. *Computers in Industry* 123 (2020), 103308.

[8] Daniel Chaffin, Ralph Heidl, John R. Hollenbeck, Michael Howe, Andrew Yu, Clay Voorhees, and Roger Calantone. 2017. The Promise and Perils of Wearable Sensors in Organizational Research. *Organizational Research Methods* 20, 1 (Jan. 2017), 3–31. doi:10.1177/1094428115617004

[9] Marco Cristani, Giulia Paggetti, Alessandro Vinciarelli, Loris Bazzani, Gloria Menegaz, and Vittorio Murino. 2011. Towards Computational Proxemics: Inferring Social Relations from Interpersonal Distances. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 290–297. doi:10.1109/PASSAT/SocialCom.2011.32

[10] Amy Drahota and Ann Dewey. 2008. The sociogram: A useful tool in the analysis of focus groups. *Nursing research* 57, 4 (2008), 293–297.

[11] Jing Du, Yangming Shi, Chao Mei, John Quarles, and Wei Yan. 2016. Communication by interaction: A multiplayer VR environment for building walkthroughs. In *Construction Research Congress 2016*. 2281–2290.

[12] Salma Elmalaki. 2021. FaiR-IoT: Fairness-aware Human-in-the-Loop Reinforcement Learning for Harnessing Human Variability in Personalized IoT. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 119–132.

[13] Barrett Ens, Joel Lanir, Anthony Tang, Scott Bateman, Gun Lee, Thammathip Piumsomboon, and Mark Billinghurst. 2019. Revisiting collaboration through mixed reality: The evolution of groupware. *International Journal of Human-Computer Studies* 131 (Nov. 2019), 81–98. doi:10.1016/j.ijhcs.2019.05.011

[14] Linton C Freeman, Douglas Roeder, and Robert R Mulholland. 1979. Centrality in social networks: II. Experimental results. *Social networks* 2, 2 (1979), 119–141.

[15] Daniel Gatica-Perez, L McCowan, Dong Zhang, and Samy Bengio. 2005. Detecting group interest-level in meetings. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 1. IEEE, I–489.

[16] Hüseyin Gençer. 2019. Group dynamics and behaviour. *Universal Journal of Educational Research* (2019).

[17] Jaris Gerup, Camilla B. Soerensen, and Peter Dieckmann. 2020. Augmented reality and mixed reality for healthcare education beyond surgery: an integrative review. 11 (Jan. 2020), 1–18. doi:10.5116/ijme.5e01.eb1a

[18] Mar Gonzalez-Franco, Rodrigo Pizarro, Julio Cermeron, Katie Li, Jacob Thorn, Windo Hutabarat, Ashutosh Tiwari, and Pablo Bermell-Garcia. 2017. Immersive Mixed Reality for Manufacturing Training. *Frontiers in Robotics and AI* 4 (Feb. 2017). doi:10.3389/frobt.2017.00003 Publisher: Frontiers.

[19] Wenchen Guo, Wenbo Zhao, Guoyu Sun, Yanni Li, Yangyi Ye, and Su Wang. 2024. Enhancing the Digital Inheritance and Embodied Experience of Zen based on Multimodal Mixed Reality System. In *ACM SIGGRAPH 2024 Posters (SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, 1–2. doi:10.1145/3641234.3671076

[20] J Richard Hackman and Neil Vidmar. 1970. Effects of size and task type on group performance and member reactions. *Sociometry* (1970), 37–54.

[21] Terry Halfhill, Eric Sundstrom, Jessica Lahner, Wilma Calderone, and Tjai M Nielsen. 2005. Group personality composition and group effectiveness: An integrative review of empirical research. *Small group research* 36, 1 (2005), 83–105.

[22] Edward T Hall, Ray L Birdwhistell, Bernhard Bock, Paul Bohannan, A Richard Diebold Jr, Marshall Durbin, Munro S Edmonson, JL Fischer, Dell Hymes, Solon T Kimball, et al. 1968. Proxemics [and comments and replies]. *Current anthropology* 9, 2/3 (1968), 83–108.

[23] Robert A Hanneman and Mark Riddle. 2005. Introduction to social network methods.

[24] Jörg Hauber, Holger Regenbrecht, Mark Billinghurst, and Andy Cockburn. 2006. Spatiality in videoconferencing: trade-offs between efficiency and social presence. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. 413–422.

[25] Malcolm Higgs, Ulrich Plewnia, and Jorg Ploch. 2005. Influence of team composition and task complexity on team performance. *Team Performance Management: An International Journal* 11, 7/8 (2005), 227–250.

[26] Martin Hoegl and Luigi Proserpio. 2004. Team member proximity and teamwork in innovative projects. *Research Policy* 33, 8 (Oct. 2004), 1153–1165. doi:10.1016/j.respol.2004.06.005

[27] Andrew Irlitti, Mesut Latifoglu, Qiushi Zhou, Martin N Reinoso, Thuong Hoang, Eduardo Velloso, and Frank Vetere. 2023. Volumetric Mixed Reality Telepresence for Real-time Cross Modality Collaboration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3544548.3581277

[28] Norbert L Kerr and R Scott Tindale. 2004. Group performance and decision making. *Annu. Rev. Psychol.* 55, 1 (2004), 623–655.

[29] Taemie Kim, Erin McFee, Daniel Olguin Olguin, Ben Waber, and Alex "Sandy" Pentland. 2012. Sociometric badges: Using sensor technology to capture new forms of collaboration. *Journal of Organizational Behavior* 33, 3 (2012), 412–427. doi:10.1002/job.1776

[30] Thomas Kosch, Andrii Matviienko, Florian Müller, Jessica Bersch, Christopher Katins, Dominik Schön, and Max Mühlhäuser. 2022. NotiBike: Assessing Target Selection Techniques for Cyclist Notifications in Augmented Reality. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI (Sept. 2022), 197:1–197:24. doi:10.1145/3546732

[31] Benedek Kurdi, Shayn Lozano, and Mahzarin R. Banaji. 2017. Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods* 49, 2 (April 2017), 457–470. doi:10.3758/s13428-016-0715-3

[32] Wallace S. Lages and Doug A. Bowman. 2019. Walking with adaptive augmented reality workspaces: design and usage patterns. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 356–366. doi:10.1145/3301275.3302278

[33] Maria Knight Lapinski and Rajiv N Rimal. 2005. An explication of social norms. *Communication theory* 15, 2 (2005), 127–147.

[34] Harold J Leavitt. 1951. Some effects of certain communication patterns on group performance. *The journal of abnormal and social psychology* 46, 1 (1951), 38.

[35] Weizhou Luo, Anke Lehmann, Hjalmar Widengren, and Raimund Dachselt. 2022. Where Should We Put It? Layout and Placement Strategies of Documents in Augmented Reality for Collaborative Sensemaking. In *Proceedings of the 2022 CHI*

Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3491102.3501946

[36] A. Martinez, Y. Dimitriadis, B. Rubia, E. Gomez, and P. de la Fuente. 2003. Combining qualitative evaluation and social network analysis for the study of classroom social interactions. *Computers & Education* 41, 4 (2003), 353–368. doi:10.1016/j.compedu.2003.06.001

[37] Florian Mathis, Brad A. Myers, Ben Lafreniere, Michael Glueck, and David P. S. Marques. 2024. MR-Driven Near-Future Realities: Previewing Everyday Life Real-World Experiences Using Mixed Reality. In *Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 76–85. doi:10.1145/3678957.3685748

[38] Andrii Matviienko, Florian Müller, Dominik Schön, Paul Seesemann, Sebastian Günther, and Max Mühlhäuser. 2022. BikeAR: Understanding Cyclists' Crossing Decision-Making at Uncontrolled Intersections using Augmented Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3491102.3517560

[39] Fintan McGee, Roderick McCall, and Joan Baixauli. 2024. Comparison of Spatial Visualization Techniques for Radiation in Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3613904.3642646

[40] Meta. 2024. Import Meta XR Packages | Meta Horizon OS Developers. https://developers.meta.com/horizon/documentation/unity/unity-package-manager/

[41] Meta. 2024. Meta Quest Pro: Premium Mixed Reality | Meta Store. https://www.meta.com/quest/quest-pro/tech-specs/#tech-specs

[42] David L Mills. 2002. Internet time synchronization: the network time protocol. *IEEE Transactions on communications* 39, 10 (2002), 1482–1493.

[43] Chris Moore and Philip J. Dunham (Eds.). 1995. *Joint attention: Its origins and role in development.* Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US. Pages: vii, 286.

[44] Jacob Levy Moreno. 1941. Foundations of sociometry: An introduction. *Sociometry* (1941), 15–35.

[45] Daniel Olguin Olguin and Peter A Gloor. 2009. Capturing Individual and Group Behavior with Wearable Sensors. In *Proceedings of the 2009 aaai spring symposium on human behavior modeling, SSS*. Vol. 9. https://aaai.org/papers/0012-ss09-04-012-capturing-individual-and-group-behavior-with-wearable-sensors/

[46] Paul Paulus. 2000. Groups, teams, and creativity: The creative potential of idea-generating groups. *Applied psychology* 49, 2 (2000), 237–262.

[47] Alex Sandy Pentland. 2012. The new science of building great teams. *Harvard business review* 90, 4 (2012), 60–69.

[48] Philip M Podsakoff, Michael Ahearne, and Scott B MacKenzie. 1997. Organizational citizenship behavior and the quantity and quality of work group performance. *Journal of applied psychology* 82, 2 (1997), 262.

[49] Mary C. Potter, Brad Wyble, Carl Erick Hagmann, and Emily S. McCourt. 2014. Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception & Psychophysics* 76, 2 (Feb. 2014), 270–279. doi:10.3758/s13414-013-0605-z

[50] Erik Prytz, Susanna Nilsson, and Arne Jönsson. 2010. The importance of eye-contact for collaboration in AR systems. In *2010 IEEE International Symposium on Mixed and Augmented Reality*. 119–126. doi:10.1109/ISMAR.2010.5643559

[51] Abebe Rorissa and Samantha K. Hastings. 2004. Free sorting of images: Attributes used for categorization. *Proceedings of the American Society for Information Science and Technology* 41, 1 (2004), 360–366. doi:10.1002/meet.1450410142 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/meet.1450410142.

[52] Michelle L. Rusch, Mark C. Schall, Patrick Gavin, John D. Lee, Jeffrey D. Dawson, Shaun Vecera, and Matthew Rizzo. 2013. Directing driver attention with augmented reality cues. *Transportation Research Part F: Traffic Psychology and Behaviour* 16 (Jan. 2013), 127–137. doi:10.1016/j.trf.2012.08.007

[53] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (Dec. 1980), 1161–1178. doi:10.1037/h0077714

[54] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language* 50, 4 (1974), 696–735.

[55] John Scott. 2011. Social network analysis: developments, advances, and prospects. *Social network analysis and mining* 1 (2011), 21–26.

[56] Elizabeth Shriberg. 1999. Phonetic consequences of speech disfluency. In *Proceedings of the international congress of phonetic sciences*, Vol. 1. 2.

[57] Helen Stefanidi, Markus Tatzgern, and Alexander Meschtscherjakov. 2024. Augmented Reality on the Move: A Systematic Literature Review for Vulnerable Road Users. *Proc. ACM Hum.-Comput. Interact.* 8, MHCI (Sept. 2024), 245:1–245:30. doi:10.1145/3676490

[58] Ralph M Stogdill. 1972. Group productivity, drive, and cohesiveness. *Organizational behavior and human performance* 8, 1 (1972), 26–43.

[59] Krzysztof Adam Szczurek, Raul Marin Prades, Eloise Matheson, Jose Rodriguez-Nogueira, and Mario Di Castro. 2023. Multimodal Multi-User Mixed Reality Human–Robot Interface for Remote Operations in Hazardous Environments. *IEEE Access* 11 (2023), 17305–17333. doi:10.1109/ACCESS.2023.3245833

[60] Henri Tajfel and John C Turner. 2004. The social identity theory of intergroup behavior. In *Political psychology*. Psychology Press, 276–293.

[61] Tanh Quang Tran, Tobias Langlotz, and Holger Regenbrecht. 2024. A Survey On Measuring Presence in Mixed Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–38. doi:10.1145/3613904.3642383

[62] Indiana State University. [n. d.]. 8.2 Defining Small Groups and Teams. ([n. d.]). https://textbooks.whatcom.edu/comm101/chapter/8-2/ Book Title: Introduction to Public Communication Publisher: Originally published by Indiana State University.

[63] Roel Vertegaal. 1999. The GAZE groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '99)*. Association for Computing Machinery, New York, NY, USA, 294–301. doi:10.1145/302979.303065

[64] Geert Vissers and Ben Dankbaar. 2013. Knowledge and Proximity. *European Planning Studies* 21, 5 (May 2013), 700–721. doi:10.1080/09654313.2013.734459

[65] Stanley Wasserman and Katherine Faust. 1994. Social network analysis: Methods and applications. (1994).

[66] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *nature* 393, 6684 (1998), 440–442.

[67] Ying Yang, Tim Dwyer, Michael Wybrow, Benjamin Lee, Maxime Cordeil, Mark Billinghurst, and Bruce H. Thomas. 2022. Towards Immersive Collaborative Sensemaking. *Proceedings of the ACM on Human-Computer Interaction* 6 (Nov. 2022), 722–746.

[68] Nicola Yuill, Steve Hinske, Sophie E. Williams, and Georgia Leith. 2014. How getting noticed helps getting on: successful attention capture doubles children's cooperative play. *Frontiers in Psychology* 5 (May 2014). doi:10.3389/fpsyg.2014.00418

[69] Xiangyu Zhang, Xiaoliang Bai, Shusheng Zhang, Weiping He, Peng Wang, Zhuo Wang, Yuxiang Yan, and Quan Yu. 2022. Real-time 3D video-based MR remote collaboration using gesture cues and virtual replicas. *The International Journal of Advanced Manufacturing Technology* 121, 11 (2022), 7697–7719.

[70] Yanxia Zhang, Jeffrey Olenick, Chu-Hsiang Chang, Steve W. J. Kozlowski, and Hayley Hung. 2018. TeamSense: Assessing Personal Affect and Group Cohesion in Small Teams through Dyadic Interaction and Behavior Analysis with Wearable Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (Sept. 2018), 1–22. doi:10.1145/3264960

[71] Tianyu Zhao, Mojtaba Taherisadr, and Salma Elmalaki. 2024. Fairo: Fairness-aware sequential decision making for human-in-the-loop cps. In *2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 87–98.