MURMR: A Multimodal Sensing Pipeline for Automated Group Behavior Analysis in Mixed Reality

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

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

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

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

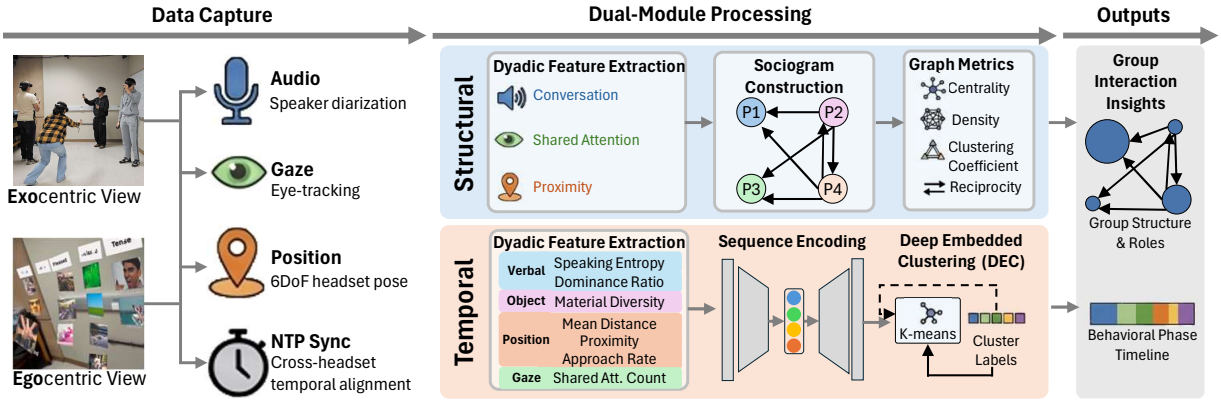

Figure 1: **MURMR** pipeline for sensing and analyzing collaborative group behavior in MR, starts by synchronizing multimodal sensor data to maintain consistency, then operates at two complementary modules. At a macro-level, session-long sensor data is aggregated to build sociograms that reveal overall group structure. At a micro-level, a temporal analysis of short interaction windows classifies fine-grained behavioral patterns.

Abstract

When teams coordinate in immersive environments, collaboration breakdowns can go undetected without automated analysis, directly affecting task performance. Yet existing methods rely on external observation and manual annotation, offering no annotation-free method for analyzing temporal collaboration dynamics from headset-native data. We introduce **MURMR**, a passive sensing pipeline that captures and analyzes multimodal interaction data from commodity MR headsets without external instrumentation. Two complementary modules address different levels of analysis: a structural module that generates automated multimodal sociograms and network metrics at both session and intra-session granularities, and a temporal module that applies unsupervised deep clustering to identify moment-to-moment dyadic behavioral phases without predefined taxonomies.

An exploratory deployment with 48 participants in a co-located object-sorting task reveals that intra-session structural analysis captures significant within-session variability lost in session-level aggregation, with gaze, audio, and position contributing non-redundantly. The temporal module identifies five behavioral phases with 83% correspondence to video observations. Cross-tabulation shows that behavioral transitions consistently occur within structurally stable states, demonstrating that the two modules capture complementary dynamics. These results establish that passive headset sensing provides meaningful signal for automated, multi-level collaboration analysis in immersive environments.

Index Terms: Collaborative Mixed Reality, Passive Sensing, Unsupervised Clustering, Multimodal Interaction

1 Introduction

Collaborative Mixed Reality (MR) applications are transforming fields such as surgical training [15], shared 3D architectural walkthroughs [11], remote industrial equipment maintenance guidance [16], and co-creative product prototyping [7], all of which rely on seamless coordination among multiple users in real time. When collaboration breaks down in these settings, whether through miscommunication during a surgical procedure or a lapse in coordination during remote maintenance, the consequences directly compromise safety, task outcomes, and team performance [41, 47]. Yet such breakdowns often go undetected without trained specialists or extensive post-hoc video review, neither of which scale to routine use. Commodity MR headsets, however, already capture gaze direction, spatial position, and audio as a byproduct of normal operation. Making this an underexploited data source for automated collaboration analysis.

Prior work in computer-supported cooperative work (CSCW) and human-computer interaction (HCI) has pursued two main avenues toward understanding collaboration in immersive environments. The first builds tools that help specialists observe and analyze group behavior, for example through replaying recorded sessions and manually coding interaction patterns [56]. While effective, these approaches depend on expert availability and post-hoc annotation. The second avenue pursues

*Diana Romero and Yasra Chandio contributed equally to this work as co-first authors.

automated in-situ analysis, including temporal methods such as process mining and sequence analysis, but these typically rely on external infrastructure such as motion capture systems [58], wearable sensors like sociometric badges [24, 59], or predefined behavioral categories [27] validated only under scripted task conditions. No existing approach passively derives both the structural organization and the temporal behavioral dynamics of group collaboration from commodity MR headset sensors alone, without relying on external hardware or annotation.

We investigate this gap through three research questions:

- **RQ1**: How do group collaboration patterns vary across temporal granularities, and what dynamics does session-level aggregation obscure?
- **RQ2**: What distinct behavioral phases characterize moment-to-moment dyadic interaction in co-located MR collaboration?
- **RQ3**: To what extent do structural and temporal analyses capture complementary aspects of collaboration?

We approach these questions through **MURMR** (Multimodal Unsupervised Relational MR), a passive sensing pipeline that combines structural and temporal analysis of multimodal data from commodity MR headsets. Our contributions are:

1. The first *passive multimodal sensing pipeline* that derives group collaboration analytics exclusively from commodity MR headset sensors, without external hardware or annotation.
2. A *structural analysis module* that translates multimodal sensor streams into weighted sociograms at both session and intra-session granularities, revealing within-session variability in group organization that session-level aggregation obscures.
3. An *unsupervised temporal clustering module* that identifies distinct behavioral phases from multimodal dyadic interaction streams without predefined categories, surfacing moment-to-moment dynamics inaccessible to session-level methods.
4. An *exploratory deployment* with 48 participants demonstrating that structural and temporal analyses capture complementary dynamics: behavioral transitions occur within structurally stable states, indicating that both levels of analysis are necessary for characterizing collaboration.

2 Related Work

We organize related work around three threads: behavioral sensing systems in MR (§2.1), analytical methods for collaboration pattern extraction (§ 2.2), and a comparative positioning of **MURMR** against existing approaches (§ 2.3).

2.1 Group Behavior Sensing in VR/MR

Research on sensing group behavior originated in physical settings, where wearable and mobile sensors, including sociometric badges, body-worn accelerometers, and microphones, were used to capture face-to-face interaction patterns and infer team cohesion [59, 38, 24, 42]. This body of work established that commodity sensors can surface rich group dynamics, but the reliance on dedicated external hardware limits scalability and deployability outside controlled lab environments.

Within immersive environments, a growing ecosystem of tools addresses different stages of the behavioral data lifecycle. In-situ and immersive replay tools such as MIRIA [6], ReLive [21], and ISA [26] enable researchers to re-enter recorded environments for observation and behavioral coding, with domain-specific extensions such as AutoVis [22] and Tesseract [30] demonstrating the value of immersive data review in automotive and design contexts. Infrastructure systems such as PLUME [23], PSI [4], and MRAT [36] provide capture, synchronization, and standardized formats for XR behavioral data, but do not infer group-level social dynamics from the data they collect.

Across both lines of work, collaboration pattern extraction remains manual and analyst-driven. **MURMR** is complementary to these tools: it operates on the same headset-native sensor streams they capture but shifts the focus from replay and manual coding to automated structural and temporal analysis of group behavior.

2.2 Automated Collaboration Analytics

A longstanding tradition in small group research uses Social Network Analysis (SNA) and sociometry, translating observed behavioral signals into sociograms that visualize relationships, roles, and subgroup structures [35, 50]. This approach has been applied across domains including nursing teams [10], classrooms [43], and VR settings where interaction and gaze patterns reflect meaningful differences in collaborative behavior [56, 1]. However, most applications construct sociograms as static, session-level aggregates.

Temporal analysis methods in CSCW and learning analytics capture how collaboration evolves over time rather than summarizing it in aggregate. Methods such as process mining [49], sequence analysis [31], and epistemic network analysis [60] extract dynamic patterns from interaction logs, but require predefined state labels or manually coded categories as input. Systems that automate collaboration assessment face analogous constraints: Echeverria et al. [12] fuse positioning, audio, and physiological data into visual proxies organized around predefined theoretical dimensions, while Léchappé et al. [27] pursue real-time detection of collaboration profiles from multimodal VR signals but validate against scripted dyadic scenarios with known behavioral categories. Across both temporal and assessment-oriented approaches, applicability is limited to settings where behavioral categories are established in advance and task conditions can be controlled.

Unsupervised representation learning offers an alternative by discovering latent behavioral patterns without pre-specified state definitions. Ma et al. [29] proposed a CNN-BiLSTM autoencoder combined with K-means clustering to discover human activity categories from unlabeled wearable sensor streams, demonstrating that deep representations can recover semantically meaningful behavioral groupings without annotation. Nguyen et al. [37] applied hierarchical Dirichlet processes to sociometric badge data to infer latent interaction patterns, showing that social contexts can similarly be extracted unsupervised. In a collaborative desktop setting, Huang et al. [20] applied unsupervised clustering to multimodal sensor data, discovering collaborative states correlated with task performance, though using external sensors without structural network analysis. Across these works, unsupervised behavioral discovery remains largely disconnected from sociometric methods: studies that cluster interaction patterns do not typically construct sociograms, and studies that build sociograms do not typically discover temporal states from the data.

In summary, SNA provides structural characterization but typically at session-level granularity. Temporal and assessment methods capture dynamics but require predefined categories. Unsupervised learning discovers patterns but has not been integrated with sociometric analysis. No prior work combines structural and temporal analysis to discover multi-level collaboration patterns from passive MR headset data without predefined behavioral taxonomies.

2.3 Positioning and Comparison

We compare the systems discussed in §2.1 and §2.2 across nine capability dimensions spanning behavioral inference, deployment model, and scope (Table 1). Two caveats are reflected in the comparison: first, **MURMR**'s sensing is headset-native but its analysis currently runs on an external server, which we mark as partial for the "No ext. hardware" dimension. Second, while the pipeline's modular design supports future real-time deployment, the current study evaluates offline processing only.

3 System Design and Implementation

MURMR's architecture is driven by a single constraint: derive both the structural organization and temporal behavioral dynamics of group collaboration entirely from commodity headset sensors, without external hardware or annotation. The pipeline achieves this in three stages (Figure 1). First, a passive sensing module synchronizes gaze, audio, and position data across headsets. The structural analysis module then converts these streams into weighted interaction graphs (automated

Table 1: Comparative analysis of related works across nine dimensions of collaborative MR analytics. Columns are organized around three capability tiers: behavioral inference (collaboration metrics, automated metrics, unsupervised inference, temporal analysis), deployment model (no external hardware, annotation-free), and scope (group-level, subgroup detection, multi-user*). Symbols: ● = present, ◖ = partially present or limited, ○ = absent.

Work	Collab. metrics	Auto. metrics	Unsupervised inference	Temporal analysis	No ext. hardware	Annotation -free	Group-level	Subgroup detection	Multi-user*
MURMR (ours)	●	●	●	●	◖	●	●	●	●
Automated collaboration analytics (§2.2)									
Yang (2022) [56]	●	○	○	○	●	○	●	○	◖
Echeverria (2019) [12]	●	◖	○	◖	○	○	●	○	●
Léchappé (2025) [27]	●	◖	○	◖	◖	◖	●	○	●
Nguyen (2013) [37]	◖	●	●	○	○	●	●	◖	●
Huang (2019) [20]	●	●	●	◖	○	◖	●	○	●
Ma (2021) [29]	○	●	●	◖	◖	●	○	○	○
Sensing, replay, and infrastructure (§2.1)									
ISA (2024) [26]	◖	◖	○	◖	○	○	●	◖	●
ReLive (2022) [21]	○	○	○	○	◖	◖	◖	○	●
MIRIA (2021) [6]	○	○	○	○	●	○	◖	○	●
AutoVis (2023) [22]	○	○	○	○	○	◖	◖	○	◖
Tesseract (2023) [30]	○	○	○	○	○	○	○	○	○
PLUME (2024) [23]	○	○	○	○	◖	◖	○	○	○
PSI (2021) [4]	○	○	○	○	○	○	○	○	●
MRAT (2020) [36]	○	○	○	○	◖	○	◖	○	●

Multi-user denotes native support for analyzing inter-user collaboration dynamics, not merely recording multiple simultaneous users.

sociograms) and derives network metrics that characterize group organization at both session and intra-session granularities. Finally, a temporal analysis module encodes short dyadic interaction segments into latent representations and clusters them to identify recurring behavioral phases without predefined categories. Together, the two analytical modules address complementary levels of collaboration that neither captures alone: stable relational structure and moment-to-moment behavioral variability.

3.1 Design Rationale

Sensor scope. We restrict input to three modalities available on commodity MR headsets with built-in eye tracking, spatial tracking, and microphone capabilities: gaze direction, spatial position, and audio. This eliminates the need for external infrastructure and allows deployment in any setting where participants wear standard headsets. These three signals form a minimum viable sensor set: prior work has established that conversational turn-taking [42], spatial co-presence [17], and shared visual attention [53] each carry meaningful and complementary information about group coordination. Richer modalities such as hand tracking, body pose, or egocentric video could extend the pipeline in future work, but are not required for the structural and temporal analyses we target here. Cross-modal dependency analysis in §5.1.2 examines whether these three modalities contribute redundant or complementary information to the fused interaction representation.

Output design. MURMR produces structured, machine-readable representations of group collaboration: sociogram edge weights, network metrics (centrality, density, reciprocity), cluster assignments, and temporal phase timelines. The present study evaluates the pipeline's analytical capabilities rather than its integration into specific end-user workflows, though the outputs are designed to support downstream applications such as real-time instructor dashboards, post-session collaborative debriefs, and adaptive task scaffolding in MR environments.

3.2 Multimodal Passive Sensing Module

MURMR is implemented as a lightweight Unity package that operates as a background service within the Unity runtime, interfacing with the headset's eye-tracking, audio, and spatial tracking APIs without modifying the host application's logic or rendering pipeline. During each MR session, the module captures three synchronized data streams from each headset. Integration requires only importing the package and initializing the sensing service at session start, with no scene modification required. To maintain temporal consistency across headsets, we perform clock alignment using the Network Time Protocol (NTP) [34], achieving sub-100 ms precision. This is sufficient because temporal overlap is accumulated across frames rather than requiring frame-exact coincidence. The finest-grained events in the pipeline, gaze convergence, operate at 50 ms–100 ms timescales [55].

Each sensor stream maps onto one of three dyadic interaction constructs used throughout the pipeline: audio produces directed *conversation* edges encoding speaking-listening duration, gaze produces *joint attention* edges capturing concurrent fixation on virtual objects, and position produces *proximity* edges encoding physical co-presence. These constructs serve as the building blocks for both the structural sociograms (§3.3) and the temporal feature vectors (§3.4).

Audio. We capture speech via the headset's onboard microphone, recording at 44.1 kHz and processing with Pyannote [5] for speaker diarization and voice activity detection. This produces timestamped speech segments with speaker identity for each participant. Segments shorter than 0.5 s are filtered, as prior work indicates this falls below the minimum duration at which listeners extract contextual meaning from speech [39] (formalized in §3.3.3).

Gaze. We capture gaze ray direction from each headset's built-in eye tracker, timestamped to the hardware clock. For each frame, gaze rays are cast against the bounding volumes of virtual objects in the scene, and a fixation on an object is registered when the ray intersects its hit box. Joint attention between two participants is detected when both users fixate on the same virtual object within overlapping time windows of at least 13 ms, a conservative lower bound adopted from vergence latency literature [55]. Overlapping gaze durations accumulate per dyad to form attention edge weights (formalized in §3.3.3).

Position. We record six-degree-of-freedom (6DoF) headset poses at 1 Hz intervals. All headsets are registered to a shared coordinate frame established at session start. In our implementation, this uses Photon Unity Networking, where one headset creates a spatial anchor at a known physical location and all other headsets align to it, though any networking stack that provides shared spatial anchoring could serve this role. The 1 Hz logging rate is consistent with the seconds-to-minutes timescale typical of co-presence events in collaborative tasks. Derived features such as approach rate are computed at this granularity, which limits sensitivity to rapid positional changes (§6). Proximity interactions are recorded whenever pairwise headset distance falls within 50 cm, corresponding to Hall's intimate distance zone [17], with accumulated co-presence duration serving as the edge weight (formalized in §3.3.3).

3.3 Structural Analysis Module

Dyadic interaction is inherently relational, existing between pairs of participants rather than within individuals, making weighted graphs a natural representation that preserves the strength and directionality of behavioral ties [35, 51]. In **MURMR**, we translate the synchronized sensor streams into a series of such graphs (automated sociograms), where each node is a *participant* and the weight of each edge reflects the cumulative interaction strength across audio, gaze, and position modalities.

3.3.1 Modality-Specific and Fused Sociogram Construction

We construct a separate sociogram for each modality to isolate behaviors that could otherwise be conflated (for example, participants who remain physically close yet do not speak) and to allow per-modality analysis even when a sensor stream is temporarily unavailable. The three modality-specific sociograms (conversation, joint attention, and proximity) are then merged into a *fused multimodal sociogram*.

Analysis granularity. We construct sociograms at two granularities: session-level aggregations capturing overall interaction patterns, and sliding 32 s windows with 16 s stride for micro-level dynamic intra-session analysis. The 32 s window duration provides sufficient interaction events to generate stable edge weights while remaining sensitive to behavioral transitions that typically occur on sub-minute timescales in collaborative tasks (parameter selection detailed in §3.4.3).

Edge definition. In each window, we assign edge weights based on the total *duration* of interaction: the sum of spoken time for conversation, the overlap of gaze fixations for joint attention, and the accumulated intervals of co-presence for proximity (thresholds defined in §3.2). Conversation graphs are directed, reflecting the asymmetry between speaker and listener. Joint attention and proximity networks are undirected, as these forms of engagement are inherently mutual.

Fused graph. After constructing three separate adjacency matrices, we z-score each and fuse them into a single multimodal network using PCA-derived weights, so that each channel contributes according to its shared variance across dyadic edges rather than equal or ad hoc contributions, while retaining the directed nature of conversational ties (formalized in §3.3.3). To verify that this linear projection does not miss meaningful nonlinear structure, we compared against kernel PCA variants. Results in §5.1.2 confirm that the fused sociogram is stable across fusion methods.

Cross-modal dependency analysis. To assess whether the three modalities carry redundant or complementary information, we compute conditional probabilities between modality-specific edges (e.g., $P(\text{joint attention} \mid \text{proximity})$) and conduct a leave-one-out ablation, reconstructing the fused sociogram from only two modalities and measuring rank correlation against the full three-modality fusion. This analysis provides empirical justification for the combined representation rather than assuming complementarity a priori. Results are reported in §5.1.2.

3.3.2 Network Measures for Behavioral Insights

We compute four network metrics across the three modality-specific and fused sociograms: (1) eigenvector centrality [13], (2) clustering coefficient [52], (3) density [46], and (4) reciprocity [19]. The first three apply to all four sociogram types, while reciprocity applies only to the directed conversation graph. A summary of how each metric maps onto interpretable aspects of group behavior is shown in Table 2.

To convert continuous metric values into interpretable categories, we apply a three-tiered scale using session-relative z-scores: **high** ($z \geq +1$, top 16%), **medium** ($-1 < z < +1$, middle 68%), and **low** ($z \leq -1$, bottom 16%), following the standard thresholds of the normal distribution. The ± 1 threshold produced occupied bins across all 12 groups in our deployment, confirming that it captures meaningful variation rather than collapsing to a single category.

3.3.3 Implementation Details

For each window $k \in \{0,\ldots,K\}$, we define an interval $[kS, kS+T]$ with window duration $T = 32$ s and stride $S = 16$ s. This 50%overlap provides continuous temporal coverage with redundancy at window boundaries. Within each window we construct three $N \times N$ adjacency matrices $W^{(\text{conv})}, W^{(\text{att})}, W^{(\text{prox})}$, where N is the number of group members. The matrices are populated as follows, with threshold values discussed in §3.2.

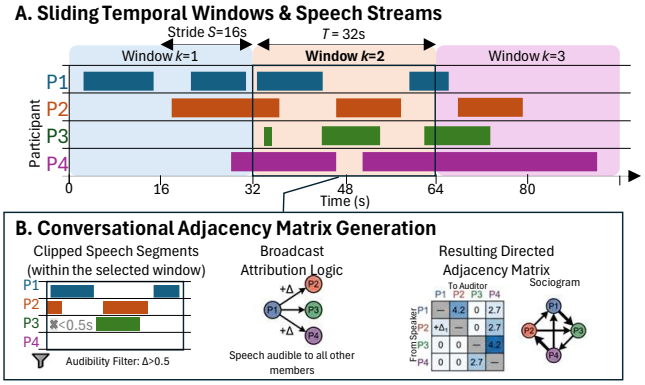

Figure 2: Directed conversation sociogram construction. (A) Speech segments are clipped to sliding windows ($T = 32$s, $S = 16$s) and filtered below 0.5 s. (B) Broadcast attribution assigns each speaker's effective duration within the window as directed edges to all other members, yielding the weighted adjacency matrix visualized as a directed sociogram.

Conversation. For every detected speech segment $(s_{\text{start}}, s_{\text{end}}, p)$ attributed to participant p that intersects the $\mathcal{I}_k$, we calculate the effective duration Δ as:

$$\Delta = |[s_{\text{start}}, s_{\text{end}}] \cap \mathcal{I}_k| \tag{1}$$

In our $3.0\,\text{m} \times 1.5\,\text{m}$ with $N = 4$, we assume speech is audible to all other group members. If $\Delta \geq 0.5$ s, we increment the directed edge weights from the speaker p to all auditors q:

$$W^{(\text{conv})}_{pq} \leftarrow W^{(\text{conv})}_{pq} + \Delta, \quad \forall q \in \{1,\ldots,N\} \setminus \{p\} \tag{2}$$

The resulting matrix $W^{(\text{conv})}$ represents the cumulative directed speaking-listening duration within the window. A visual example of constructing the conversation sociogram is shown in Figure 2.

Attention. W we clip each user's gaze intervals to $\mathcal{I}_k$ and aggregate overlapping fixation durations between pairs (i,j). Let $\mathcal{O}ij^k$ denote the set of overlapping fixation intervals between users i and j within $Ik\mathcal{I}k$:

$$Wij^{(\text{att})} = W^{(\text{att})}_{ji} = \sum_{\delta \in \mathcal{O}^k_{ij}} \delta, \quad \text{for } \delta \geq 13,\text{ms} \tag{3}$$

This results in an undirected adjacency matrix where edges represent the total duration of joint attentional focus.

Proximity. We align headset poses at common timestamps within $\mathcal{I}_k$ an indicator function to count instances where the pairwise distance d_{ij} falls within the proximity threshold:

$$W^{(\text{prox})}_{ij} = W^{(\text{prox})}_{ji} = \sum_{t \in \mathcal{I}k} \mathbf{1}[dij(t) \leq 50,\text{cm}] \tag{4}$$

This weight represents a sample count that reflects the frequency and duration of close-range physical interaction.

Fusion. To combine the three modality-specific matrices, we assemble an edge-weight matrix X where rows correspond to directed dyad pairs and columns to modalities. For undirected modalities (attention and proximity), both directions receive identical weights. Each column is z-scored to equalize scale across modalities. We then fit PCA with a single component to the standardized matrix, square the resulting loadings to obtain variance-contribution proportions, and normalize these to sum to 1, yielding fusion weights $\alpha_{\text{conv}}, \alpha_{\text{att}}, \alpha_{\text{prox}}$. The fused sociogram is computed as a weighted linear combination of the standardized edge weights:

$$W^{(\text{fused})} = \sum_{m \in \text{conv,att,prox}} \alpha_m, Z^{(m)} \tag{5}$$

where $Z^{(m)}$ denotes the z-scored edge weights for modality m. Squaring the loadings ensures all weights are non-negative. Because

Table 2: Interpretation of network metrics by interaction mode. 🎙: conversation, 👁: joint attention, 📍: proximity.

Modality	Metric (Ref.)	High Value Interpretation	Low Value Interpretation
Centrality measures identify potential leaders and information brokers in conversation, attention, and proximity networks.			
🎙, 👁, 📍	Eigenvector [13]	Connected to other highly central participants	Linked mainly to peripheral participants
Cohesion measures quantify bonding and tightness in proximity and attention networks.			
👁, 📍	Clustering Coef. [52]	Neighbors are densely interconnected, reflecting tight local subgroup cohesion	Sparse neighbor connections, reflecting weak local cohesion
🎙, 👁, 📍	Density [46]	Well-connected network with active group engagement	Fragmented or minimally interacting group
Connectivity measures assess the balance of two-way exchanges in the conversation network.			
🎙	Reciprocity [19]	Balanced two-way exchanges (dialogue)	Predominantly one-way communication

fusion operates on standardized values, the fused graph reflects relative interaction strength across modalities rather than raw duration counts.

Runtime. After filtering inactive windows, a typical group yields ∼50 structural windows (∼14 minutes of active interaction). Structural analysis of this volume completes in ∼14 s on commodity hardware (Intel i7-14th gen, 32 GB RAM).

3.4 Temporal Analysis Module

Sociograms characterize interaction strength within each window but do not classify which behavioral pattern a given window represents or track how patterns recur across a session. In collaborative tasks, dyads do not maintain a single interaction style throughout a session but shift between qualitatively different modes on timescales of seconds. Our temporal analysis addresses this by segmenting each session's dyadic behavioral features into temporal phases via unsupervised clustering of time-series features. The complete pipeline is shown in the temporal block of Figure 1.

3.4.1 Feature Selection and Construction.

To capture the behavioral richness of dyadic interaction, we extract features spanning four dimensions: how pairs converse (turn-taking dynamics, speaking balance), how they move relative to each other (distance, approach, co-presence), what objects they jointly attend to and how broadly they explore task materials. We begin with a pool of 23 such features computed from moment-to-moment participant interactions at one-second intervals[1]. We reduce this pool to a compact set through three successive filters that remove uninformative features, identify the most discriminative ones, and eliminate redundancy: (1) a variance filter ($t = 1 \times 10^{-5}$) removes near-constant features; (2) K-Means clustering is fit on the data using the remaining features, with k auto-selected via silhouette score, and a Random Forest classifier trained on the resulting cluster labels ranks features by importance, retaining the top 10; (3) pairwise absolute correlations are computed among the retained features, and any feature with $r \geq 0.95$ is dropped. We adopt this unsupervised-then-supervised surrogate ranking because it surfaces features that best discriminate naturally occurring groupings in the data, without requiring predefined behavioral labels. This process yields 7 features.

This surrogate procedure serves only as a dimensionality reduction step; its cluster labels are discarded after feature ranking. The downstream temporal clustering that produces all reported results uses a separate convolutional-recurrent autoencoder (§3.4.2) trained from scratch on only the 7 selected features, discovering its own latent space and cluster assignments. The 7 retained features span four behavioral dimensions:

Verbal dynamics: *entropy_speaking*, the evenness of turn-taking within the dyad (high = balanced exchange, low = silence or monopolization), and *dominance_ratio*, the imbalance in speaking time (0.5 = equal, deviations = asymmetric).

Object diversity: *material_diversity*, the count of distinct virtual objects both participants jointly attend to (high = broad surveying, low = sustained focus on few items).

Proximity: *dist_mean*, average dyadic distance (low = side-by-side, high = dispersed); *prox_binary*, co-presence within the defined proximity threshold (whether the pair is within arm's reach); and *approach_rate*, the rate of movement toward or away from one another (positive = convergence, negative = separation).

Joint attention: *joint_att_cnt*, the number of joint gaze fixations on the same virtual object, capturing moments when both participants inspect the same item simultaneously.

All features are computed at one-second intervals, z-normalized per dyad, and aligned to a uniform temporal grid.

3.4.2 Model Architecture.

We treat each dyad's interaction over a segment as a $T \times F$ matrix, where T is the number of 1 s windows per segment and $F = 7$ is the number of retained features. We select $T = 16$ and stride $S = 8$ (50% overlap) via grid search, jointly optimizing silhouette score and reconstruction loss on held-out data (selection detailed in §3.4.3). The structural module operates at a different granularity ($T = 32$, $S = 16$); the two modules are not temporally aligned by design, as each optimizes its own window parameters independently. Each sequence is encoded by a convolutional-recurrent autoencoder [29, 57] into a 32-dimensional latent vector, which is then clustered via K-means with a deep embedded clustering objective [54].

3.4.3 Window Length and Stride Selection.

For the temporal module (the structural module uses separate parameters, §3.3), we grid-searched window length $T \in 8,16,32$ and stride $S \in 4,8,16$ with the constraint $S \leq T/2$, yielding 6 valid combinations. Each configuration was evaluated in a coarse pass (1 epoch), which produced non-overlapping reconstruction loss (0.47–0.66) and silhouette score (0.17–0.28) ranges across configurations for stable ranking. The top 50% were then refined (5 epochs), and candidates were ranked by an equally weighted combination of silhouette score and reconstruction loss. Stride was searched systematically across all window sizes, not only at the final T. Among the top-ranked configurations, $\langle T{=}16, S{=}8 \rangle$ achieved the best composite rank score. We adopt this configuration throughout, yielding 5,334 dyadic windows across 12 groups (72 pairs).

3.4.4 Implementation and Output.

All computation is performed offline. To scale across many dyads or lengthy sessions, feature extraction, encoding, and clustering run in parallel. The full dataset (5,334 windows) was used without subsampling during final training; a fast evaluation mode that subsamples to 5,000 windows is available for rapid iteration but was not used for reported results. On a consumer CPU (Intel i7-14th gen, 32 GB RAM), the full pipeline processes 12 groups (6 dyads each, sessions averaging ∼32 minutes total) in ∼80 seconds total, with a per-group median of 3.25 s (SD = 0.33, range 2.89–4.00 s). The bottleneck is k-selection, which accounts for ∼99% of per-group runtime. The module outputs a cluster label for each dyadic window, from which feature heatmaps and phase-aligned timelines can be derived. Cluster interpretations are reported in §5.2.

4 Exploratory Deployment

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

[1] The full set of candidate features: *speak_overlap*, *speak_only_i*, *speak_only_j*, *speaker_switch*, *silence*, *floor_streak_i*, *floor_streak_j*, *resp_latency*, *burst_switch_rate*, *burst_overlap_rate*, *dominance_ratio*, *entropy_speaking*, *bigram_entropy*, *fano_switch*, *dist_mean*, *prox_binary*, *approach_rate*, *dist_accel*, *dist_jerk*, *joint_att_cnt*, *joint_att_dur*, *shared_att_ratio*, *material_diversity*. Symmetric features (_i/_j pairs) are retained separately because dyad members are ordered, not interchangeable.

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 [48].

4.2 Materials

We conducted the study in a $3.0\,\mathrm{m} \times 1.5\,\mathrm{m}$ space, cleared of materials to minimize distractions, where participants navigated and collaborated in close quarters. Each participant used a Meta Quest Pro headset [33], which captured and streamed eye gaze, audio, and 6DoF pose data over our local Wi-Fi network. We synchronized all devices with NTP (sub-100 ms precision, as detailed in §3.2) and built the collaborative MR app in Unity with the Meta XR SDK [32]. All virtual objects were aligned to a shared coordinate frame so that each user viewed a consistent scene without additional prompts, cues, or enforced turn-taking.

4.3 Collaborative Task

We asked each group to sort the 28 OASIS images [25] (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 [45]. This image sorting task has been adopted in various group dynamics studies [14, 9].

The virtual images were scattered throughout, overlaid on the see-through physical space, 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 ok 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 [44, 3, 2]. An egocentric view of the task is shown in Figure 1. The average task completion time was 32.4 minutes (SD=8.4SD = 8.4 SD=8.4), but substantial portions of each session contained no detectable dyadic interaction (e.g., individual sorting, idle periods between sub-tasks). The pipeline retains only windows in which at least one modality registers nonzero dyadic activity, yielding a mean of 53 active structural windows per group ($\sim$15 minutes of active interaction per session).

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. The entire session, including setup, training, task, and wrap-up, took approximately $35-45$ minutes.

4.5 Data Collection and Correspondence Checks

We captured multimodal data passively from each headset with post-hoc synchronization and ran it through **MURMR**'s end-to-end pipeline (structural and temporal modules described in §3.3 and §3.4, respectively) to analyze group behavior. We assessed the pipeline's outputs through two complementary checks: behavioral **correspondence check** with observed activity and **internal consistency** (clustering stability and sociogram metric consistency across time and groups).

For the **correspondence check**, one researcher reviewed time-aligned egocentric video from both members of each dyad. We sampled 100 windows (drawn from the middle portion of each session to avoid startup transients) and presented each alongside its predicted cluster label and characterization. Using synchronized video and audio from both participants' headsets, the coder judged whether the predicted behavioral pattern matched the observed interaction (match/mismatch/uncertain), and for mismatches recorded the observed cluster. This procedure yielded 83% agreement between automated labels and human judgment; per-cluster precision, recall, and interpretation are reported in §5.2. Coding was performed by a single researcher due to data access restrictions under our IRB protocol.

We did not collect post-hoc subjective collaboration surveys. Post-hoc ratings of moment-level behavioral states are unreliable indicators of observed interaction dynamics [8], and our pipeline targets observable behavioral patterns rather than participants' internal experience. The correspondence check instead verifies that the pipeline's outputs align with behaviors visible on camera, without claiming access to a definitive ground truth for collaboration states.

5 Results & Analysis

5.1 RQ1: Multi-Granular Temporal Structural Patterns

To answer **RQ1** presented in §1, we evaluate the structural module at two temporal granularities, examining sensing modality contributions to the fused sociogram through windowed metrics (§5.1.1), cross-modal dependencies (§5.1.2), and group-level network patterns (§5.1.3).

5.1.1 Session-Level vs. Windowed Analysis

We compared session-level and windowed interaction metrics presented in Table 2 across all groups. Session-level aggregation resulted in most graph metrics to have near-identical values across groups ($CV: 0.000-0.095$), as participant pairs eventually interact over full sessions. In contrast, 32 s sliding-window analysis (§3.3.3) revealed substantial within-session variability ($CV: 0.268-0.895$), confirming that temporal resolution is necessary to capture meaningful fluctuations in collaborative structure (Figure 3, right panel). Group 12's fused density trajectory over time illustrates this within-session variability (Figure 3, center panel) with per-modality density shifting significantly across windows that session-level aggregation obscures. Per-group reciprocity and fused density values are reported in Table 3

To assess conversation reciprocity variance, we analyzed three groups representing the observed spectrum: Group 5 (low; $\mu = 0.179, \sigma = 0.169$), Group 10 (moderate; $\mu = 0.451, \sigma = 0.273$), and Group 1 (high; $\mu = 0.538, \sigma = 0.292$). Across all groups, mean reciprocity ranged from 0.179 to 0.538 , offering granular differentiation that session-level aggregation would obscure. Their windowed reciprocity trajectories are illustrated in Figure 3 (left panel).

Hence, this analysis reveals that session-level metrics compress critical temporal dynamics that windowed analysis successfully surfaces.

5.1.2 Modality Contributions and Cross-Modal Dependencies

To evaluate the internal structure of the fused sociogram, we analyzed how individual sensing modalities contribute to the final representation. **Principle Component Analysis (PCA) loadings and Variance.** Across all groups, conversation and joint attention dominated the fusion with mean squared loadings of $=0.401$ and 0.409, respectively. Proximity contributed least at 0.190 The first principal component explained 57.2% of the variance on average, indicating that the fused interaction density reflects integrated signals rather than a single dominant channel. This pattern is evident in Group 12's data (Figure 3, center panel), where fused interaction density incorporates multi-modal contributions rather than being driven by a single dominant channel.

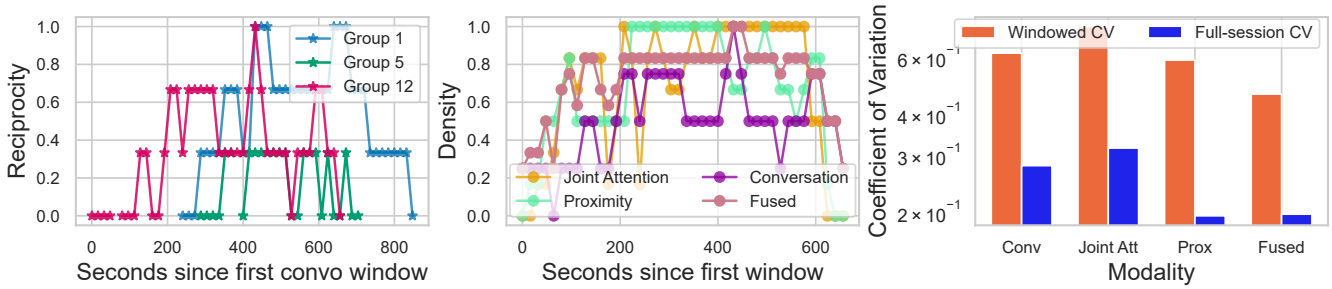

Figure 3: Windowed-session analysis to assess behavioral patterns obscured by session-level aggregation. Conversation reciprocity for representative Groups 1,5,12 (left), Group 12's multimodal density trajectory (center), and density variation (right).

Table 3: Per-group conversation reciprocity and mean fused density computed from active-session windows only.

Group	Mean Reciprocity	σ Reciprocity	Fused Density
1	0.538	0.292	0.509
2	0.305	0.234	0.424
3	0.261	0.253	0.609
4	0.400	0.216	0.669
5	0.179	0.169	0.389
6	0.313	0.281	0.621
7	0.480	0.364	0.748
8	0.308	0.233	0.716
9	0.491	0.242	0.743
10	0.451	0.273	0.723
11	0.515	0.390	0.684
12	0.350	0.268	0.720

Cross-Modal Dependencies. Analysis of 3,822 dyad-windows showed high activity across modalities: conversation in 68.0%, joint attention in 55.9%, and proximity in 54.1%. Proximity and attention exhibited the strongest coupling ($P(\text{prox} \mid \text{att}) = 0.70$; $P(\text{att} \mid \text{prox}) = 0.72$), while conversation conditioned on attention ($P(\text{att}\,|\,\text{conv}) = 0.67$) and proximity ($P(\text{prox}\,|\,\text{conv}) = 0.62$) was somewhat weaker. Lift analysis ($1.15 - -1.29$) confirmed all pairs co-occur more often than chance, validating that these signals are physically and socially linked.

Kernel PCA comparison. To test for nonlinear structure, we compared linear PCA against Radial Basis Function (RBF) and polynomial kernels. Polynomial kernels yielded near-identical results (mean edge rank $\rho = 1.000$), and while RBF was less stable ($\rho = 0.905$) it preserved the top-ranked dyad in 11 of 12 groups. These results confirm the fused sociogram is robust to the fusion method, with linear PCA providing stable, interpretable weights. Structural differences under nonlinear fusion are discussed further in §6.

Leave-one-out ablation. We evaluated modality redundancy by excluding one channel and measuring dyadic rank preservation. Removing any modality caused substantial disruption; dropping conversation resulted in the lowest preservation (mean $\rho = 0.605$; $\rho \geq 0.8$ in 6 of 12 groups), followed by proximity ($\rho = 0.567$; $\rho \geq 0.8$ in 4 of 12) and conversation ($\rho = 0.476$; $\rho \geq 0.8$ in 6 of 12). All three modalities showed mean ρ below 0.7, indicating that no single channel is redundant. We discuss the implications for modality selection in §6.

From this analysis, we validate that the three sensing modalities provide complementary, non-redundant signals essential for capturing the full complexity of group structure.

5.1.3 Network Metrics Across Groups

Edge weight distributions. Each modality produces a distinct profile. Proximity yields the highest total weight per group (8592.6 ± 4325.7) and the widest range (max 2534.8 ± 1773.7). Joint attention has the lowest totals (757.6 ± 781.8) but the highest edge inequality (Gini $= 0.272 \pm 0.122$). Conversation is the most uniform (Gini $= 0.235 \pm 0.076$), reflecting the broadcast edge attribution model described in §3.3.3. Notably, the fused sociogram has the lowest Gini coefficient of all (0.208 ± 0.093), indicating that integrating modalities produces a more balanced representation of group engagement than any single channel. All 12 fused graphs are complete (density $= 1.00$), meaning every participant pair has nonzero fused interaction strength.

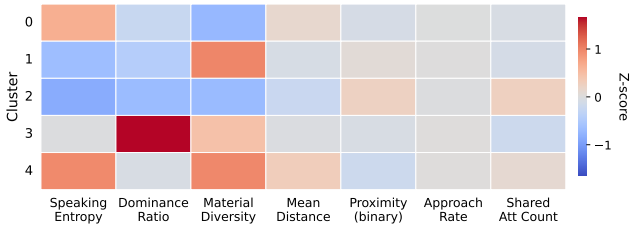

Figure 4: Heatmap of clusters (rows) vs. features (columns), where color intensity shows each feature's deviation from its mean.

Structural role variation. No single participant dominated across groups. Strength ratios between the most and least connected participants ranged from 1.10 (Group 11) to 3.97 (Group 8). Modality-level disagreement was significant: all three sensors agreed on the strongest dyad in only 2 of 12 groups, while in 5 groups, each modality identified a unique strongest pair. This divergence reinforces the necessity of the fused representation and cross-modal findings in §5.1.2.

Illustrative cases. Contrasting patterns in Groups 8 and 11 highlight these dynamics. In Group 8, conversation was negatively correlated with both proximity ($\rho = -0.771$) and attention ($\rho = -0.714$), resulting in the lowest cross-modal agreement and the highest strength ratio. Conversely, Group 11 participation was nearly equal (ratio 1.10), but the modalities still identified different strongest pairs, proving that modality disagreement persists even in egalitarian groups.

Hence, we conclude that the fused representation provides a stabilized view of group structure that diverges from any single modality.

5.2 RQ2: Emergent Behavioral Phases

To answer **RQ2**, we evaluate the behavioral patterns identified by the temporal clustering module. We discuss cluster selection (§5.2.1), the interpretability of features via decision trees (§5.2.2), and validation against egocentric video (§5.2.4).

5.2.1 Cluster Selection & Characterization

The temporal clustering pipeline (§3.4) identified $k = 5$ clusters with high stability (silhouette $= 0.81$, cross-seed $ARI = 1.0$) from 5,334 dyadic windows (§3.4.3)[2]. To link clusters to behavioral patterns, we inspected their z-scored feature profiles (Figure 4). Two of the seven selected features require interpretive context: entropy_speaking and dominance_ratio. High speaking entropy ($\sim$1.58 bits) signifies balanced conversation with natural pauses, while low entropy ($\sim$0) indicates state dominance, such as floor monopolization or silence. The dominance_ratio measures speaking time symmetry: 0.5 indicates balanced turns, with deviations categorized as near-balanced ($|z| < 0.3$), moderately asymmetric ($0.3 \leq |z| < 0.6$), or asymmetric ($|z| \geq 0.6$). The five clusters are characterized as follows:

[2]The differing observation counts between sections reflect distinct windowing configurations optimized for the structural and temporal modules.

Table 4: Entropy of cluster membership across groups, pairs, and actors (lower values = more context-specific). Actor entropy is the mean of per-actor entropies for both dyad members.

Cluster	N	Group Ent.	Pair Ent.	Actor Ent.
C0 (Balanced Narrow Focus)	1498	0.77	2.90	2.04
C1 (Diverse Exploration)	847	0.99	2.44	1.79
C2 (Concentrated Shared Focus)	1250	0.97	2.83	2.17
C3 (Guided Exploration)	1003	0.74	2.46	1.77
C4 (Active Distributed Dialogue)	736	**0.63**	**2.13**	**1.47**

C0 (Balanced Narrow Focus, 28.1%). Characterized by balanced conversation (high speaking entropy, $z = +0.63$), but limited material engagement (low material diversity, $z = -0.73$) and near-balanced dominance ($z = -0.25$).
C1 (Diverse Exploration, 15.9%). Pairs exploring a wide range of materials under a dominant conversational state with skewed speech. High material diversity ($z = +0.97$), low speaking entropy ($z = -0.66$), and moderately asymmetric dominance ($z = -0.45$).
C2 (Concentrated Shared Focus, 23.4%). High joint attention and close physical proximity on a narrow set of materials, with asymmetric speech. Characterized by very low speaking entropy ($z = -0.85$), asymmetric dominance ($z = -0.69$), low material diversity ($z = -0.71$), and close proximity ($z = -0.24$).
C3 (Guided Exploration, 18.8%). Strongly elevated dominance ratio ($z = +1.65$), one participant dominates the floor while both explore varied materials with moderately high material diversity ($z = +0.44$).
C4 (Active Distributed Dialogue, 13.8%). Balanced, active conversation (high speaking entropy, $z = +0.95$) across diverse materials ($z = +0.96$)), and balanced dominance with greater spatial separation ($z = +0.29$ for distance).

5.2.2 Cluster Interpretability

We constructed a surrogate decision tree to distill each behavioral pattern into a rule hierarchy. Shapley additive explanations (SHAP) values [28] confirmed that three features provide nearly all splitting power: material diversity (importance = 0.425), speaking entropy(0.320), and dominance ratio (0.251), with other features contributing less than 0.2%. The root split utilizes dominance ratio to isolate Guided Exploration (C3). An edge case with extreme dist_mean is reclassified as Active Distributed Dialogue (C4). For the remaining windows, speaking entropy distinguishes balanced from asymmetric speech, while material diversity separates narrow-focus from diverse-exploration patterns. This depth-4 tree reproduces cluster assignments with 98.9% accuracy (macro $F1 = 0.987$). This high fidelity affirms that the unsupervised clusters map onto a transparent, logically consistent hierarchy of observable interaction behaviors.

5.2.3 Distribution Across Dyads

To evaluate how interaction styles vary across social contexts, we computed the categorical entropy of cluster membership at the group, pair, and actor levels (Table 4). Lower entropy signifies behaviors concentrated in specific contexts, while higher scores indicate widely shared styles.
Context-Specific Patterns: Active Distributed Dialogue (C4) exhibited the lowest entropy across group (0.63), pair (2.13), and actor (1.47) levels, suggesting this balanced interaction style is highly dependent on specific team dynamics. Balanced Narrow Focus (C0) and Guided Exploration (C3) also showed low group entropy, indicating they are team-specific traits.
Generalized Patterns: Conversely, Diverse Exploration (C1) and Concentrated Shared Focus (C2) exhibited the highest group entropy (0.99 and 0.97), demonstrating that these behaviors occur broadly across the participant pool regardless of group assignment.

5.2.4 Behavioral Correspondence Check

With the procedure described in §4.5, a single coder reviewed 100 stratified windows against synchronized egocentric video, achieving 83% agreement with automated cluster assignments. Clusters 0 through 3 demonstrated high reliability, with F1 scores ranging from 0.83 to 0.90. Balanced Narrow Focus (C0) was the most accurately identified pattern with an F1 of 0.90. Diverse Exploration (C1) and Concentrated Shared Focus (C2) both reached 0.95 recall. Active Distributed Dialogue (C4) achieved perfect precision but lower recall at 0.55, primarily due to feature overlap with Guided Exploration. Overall macro-averaged metrics ($F1 = 0.83$) confirm that the temporal module reliably recovers dominant interaction patterns, with Active Distributed Dialogue presenting the most ambiguous boundary due to its shared features with neighboring clusters.

Table 5: Structural metrics whose low/medium/high tertile distributions show significant association with behavioral clusters. Most metrics use the full aligned sample (N=466); reciprocity metrics have fewer observations because reciprocity is undefined in windows lacking directed edges. Of 14 candidate metrics, 5 collapsed to fewer than three tertile bins and were excluded; the remaining 9 all reached significance.

Metric (binned)	N	χ^2	p	Cramér's V
Joint Attention Eigenvector	466	99.40	<0.001	0.33
Fused Reciprocity	341	68.12	<0.001	0.32
Fused Density	466	76.82	<0.001	0.29
Proximity Density	466	70.80	<0.001	0.28
Conversation Density	466	65.08	<0.001	0.26
Proximity Eigenvector	466	61.97	<0.001	0.26
Fused Eigenvector	466	46.32	<0.001	0.22
Conversation Eigenvector	466	45.70	<0.001	0.22
Conversation Reciprocity†	316	30.24	<0.001	0.22

†Minimum expected cell count $= 2.8 < 5$; chi-squared approximation may be unreliable.

5.3 RQ3: Complementarity of Structural and Temporal Analyses

To answer **RQ3**, we examined whether structural and temporal modules capture overlapping or distinct aspects of collaboration.

5.3.1 Cluster-Metric Cross-Tabulation

To examine how temporal cluster assignments relate to structural network properties we cross-tabulated the five behavioral clusters against tertile-binned network metrics across 466 group-windows. Nine metrics showed significant associations with cluster membership at $p < 0.001$ (Table 5), with joint attention eigenvector ($V = 0.33$) and fused reciprocity ($V = 0.32$) exhibiting the strongest effects. C2 was overrepresented in high joint-attention centrality (58.4%), while C3 showed the highest fused reciprocity, C1 was almost absent from the high bin (1.2%). For fused density, Balanced Narrow Focus (C0) and Guided Exploration leaned toward the low tertile (51–53%), while C4 leaned high (44.9%).

Effect sizes remained small to moderate (Cramér's $V = 0.22$–0.33), indicating partial overlap rather than redundancy.

5.3.2 Structural Stability During Temporal Transitions

To illustrate how the two modules complement each other, we present case analysis of Group 12 in Figure 5 with fused eigenvector centrality alongside temporal cluster labels for all six dyads in Group 12 over a 528 s session. Analysis shows that structural metrics remained stable while dyads concurrently occupied four distinct behavioral clusters with 100% consistency: A–B in Diverse Exploration, A–D in Concentrated Shared Focus, B–C in Active Distributed Dialogue, and B–D in Balanced Narrow Focus. While structural analysis would characterize this state as uniformly cohesive, the temporal module identifies four distinct interaction patterns coexisting within it. Furthermore, synchronized transitions to Active Distributed Dialogue occurred without triggering significant shifts in any of the 14 structural metrics. Following an initial transient, fused eigenvector centrality stabilizes (mean $EV = 0.48$), masking these micro-level behavioral perturbations that only the temporal module detects.

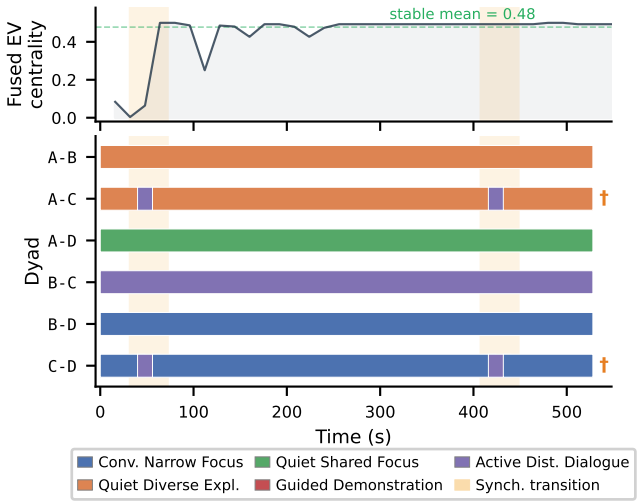

Figure 5: Fused eigenvector centrality (top) and temporal cluster assignments for all six dyads (bottom) in Group 12 over a 528 s session. Orange bands mark periods where two dyads (†) undergo synchronized transitions to *Active Distributed Dialogue*; no structural metric registers a corresponding change ($0/14$, Mann–Whitney U, all $p > 0.07$ uncorrected).

This pattern generalizes to 8 of 12 groups, where three or more dyad-level clusters emerge under stable structural conditions. In Group 8, three dyads underwent synchronized transitions that structural metrics failed to register ($0/14$, $p > 0.20$ uncorrected).

While cross-tabulation indicates partial overlap (Cramér's $V = 0.22$–0.33), these case studies confirm that temporal mode shifts often occur without detectable structural changes, proving the modules capture complementary aspects of group dynamics.

6 Discussion and Limitations

6.1 Summary of Findings

Three research questions guided this work. For RQ1, windowed structural analysis revealed within-session variability that session-level aggregation compresses entirely, and all three sensing modalities contribute meaningfully to the fused sociogram: conversation and attention co-dominate the PCA-derived fusion weights, while proximity, though the weakest contributor, is non-redundant, as removing any single channel disrupts dyadic rankings substantially. For RQ2, the temporal clustering module identified five behaviorally distinct interaction phases with perfect cross-seed stability, separable by just three features (material diversity, speaking entropy, and dominance ratio); a correspondence check against egocentric video confirmed 83% agreement, indicating that the detected clusters map onto observable behavioral patterns. For RQ3, the two modules capture complementary rather than redundant dynamics: behavioral cluster transitions occur within structurally stable states, and while cross-tabulation reveals significant associations between temporal clusters and structural metrics, the effect sizes are moderate, indicating partial overlap rather than reducibility. Together, these findings suggest that structural and temporal views of collaboration each surface patterns the other misses, and that passive sensing from commodity headsets provides sufficient signal to support both levels of analysis.

6.2 Multimodal Fusion and Modality Roles

The PCA-based fusion assigns empirically derived weights that reflect each modality's contribution to shared variance, rather than assuming equal importance across channels. In our task, conversation and attention co-dominated the fusion weights while proximity contributed least, but this weighting is task-specific: physical assembly or navigation tasks, where spatial coordination is central, would likely shift the balance toward proximity. The kernel PCA comparison provides evidence that this linear fusion captures the dominant variance structure: the polynomial kernel recovered near-identical sociograms ($\rho > 0.999$), and even the less stable RBF kernel preserved the top-ranked dyad in 11 of 12 groups. However, fusion inherently compresses modality-specific information into a single summary view. Group 8 illustrates this cost: conversation was negatively correlated with both proximity and joint attention, a pattern invisible in the fused sociogram. Preserving per-modality views alongside the fused representation is therefore not optional but necessary for analysts investigating how different behavioral channels relate to one another. The pipeline supports this by computing and storing all four network types independently, with fusion as an additional layer rather than a replacement.

6.3 Validation Without Ground Truth

Moment-to-moment collaboration states are not objectively observable constructs: they are shaped by internal states, social context, and subjective interpretation [40], and even trained raters typically achieve only moderate agreement when coding behavioral observations from video [18]. Post-hoc self-reports, while valuable for capturing participants' subjective experience, are unreliable indicators of the micro-level behavioral dynamics that automated sensing detects [8]. This is not a limitation specific to **MURMR** but a domain-inherent constraint shared across social signal processing and collaboration analytics. Prior work in these fields validates automated behavioral analysis through internal consistency, temporal stability, and expert correspondence rather than per-frame ground truth: Pentland's sociometric badge research evaluated sensing outputs against organizational outcomes and observer ratings rather than moment-level labels [42], and Echeverria et al.'s collaboration translucence framework relied on expert interpretation of system outputs to establish face validity [12]. We position **MURMR** in this tradition as an exploratory pipeline that assists analysts rather than a classifier requiring labeled ground truth. The relevant validation criteria are reliability (cross-seed ARI $= 1.0$, surrogate tree fidelity $= 98.9\%$) and behavioral correspondence (83% agreement with observed video), not correctness against subjective labels. We are candid about the limits of this correspondence check: a single coder reviewed 100 of 5,334 windows under IRB access restrictions, with no inter-rater reliability data. This establishes behavioral plausibility, not statistical generalization. Expanding coverage, adding a second coder, and computing inter-rater agreement are necessary steps before the cluster labels can be treated as validated categories.

6.4 Limitations

The results reported here are specific to a co-located, object-sorting task performed by four-person groups. The five behavioral clusters, relative modality weightings, and cross-modal dependency patterns should not be assumed to transfer to tasks with different interaction demands, such as debates, remote collaboration, or physical assembly. The seven retained temporal features were empirically selected from this dataset; tasks with scripted turns, asymmetric roles, or different physical constraints may require different feature sets. With 12 groups (48 participants), the sample is sufficient for pipeline demonstration and internal validation but limits statistical generalization, as reflected in the small-to-moderate cross-tabulation effect sizes. The three sensing modalities (gaze, audio, position) were chosen deliberately for lightweight deployment on commodity headsets, but they omit hand gestures, facial expressions, and virtual object manipulation context that shape MR collaboration. The modular architecture supports extension with additional channels, but current results reflect only these three. Conversation edge attribution broadcasts each utterance to all group members, which may overestimate conversational engagement in subgroup-specific dialogue. On the validation side, the correspondence check relied on a single coder reviewing 100 of 5,334 windows with no inter-rater reliability data. Active Distributed Dialogue (C4) showed the lowest recall (0.55, $F1 = 0.71$), with most misclassified windows reassigned to neighboring clusters, indicating that the boundary between balanced high-entropy conversation and adjacent patterns needs refinement. These constraints define the scope within which the pipeline's outputs should be interpreted and point toward the extensions needed before deployment in applied settings.

7 Conclusion

We presented **MURMR**, a passive sensing pipeline that derives both structural and temporal views of group collaboration from commodity MR headset data without external instrumentation or manual annotation. By integrating automated multimodal sociograms with unsupervised deep clustering, the system reveals within-session variability and distinct behavioral phases that traditional session-level aggregation obscures. Our findings establish that headset embedded sensors can provide sufficient signal to identify complex group organization and social roles.

The identified five-phase behavioral taxonomy and modality weightings are context-specific to co-located, four-person sorting tasks. While cross-tabulation demonstrates the modules provide complementary, non-reducible insights, further validation across diverse collaborative settings and richer sensing modalities is required. Ultimately, **MURMR** provides a foundation for scalable, automated tools that augment analyst interpretation, advancing the state-of-the-art in immersive collaboration research.

Acknowledgments

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

References

[1] H. Bai, P. Sasikumar, J. Yang, and M. Billinghurst. 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*, pp. 1–13, 2020. 2

[2] N. Berlin, M. Gueye, and S. Monjon. Feedback and cooperation: An experiment in sorting behavior, 2025. 6

[3] M. F. Bjerre. 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):74–87, 2015. 6

[4] D. Bohus, S. Andrist, A. Feniello, N. Saw, M. Jalobeanu, P. Sweeney, A. L. Thompson, and E. Horvitz. Platform for situated intelligence. *arXiv preprint arXiv:2103.15975*, 2021. 2, 3

[5] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7124–7128. IEEE, 2020. 3

[6] W. Büschel, A. Lehmann, and R. Dachselt. Miria: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–15, 2021. 2, 3

[7] G. Cascini, J. O'Hare, E. Dekoninck, N. Becattini, J.-F. Boujut, F. B. Guefrache, I. Carli, G. Caruso, L. Giunta, and F. Morosi. Exploring the use of ar technology for co-creative product and packaging design. *Computers in Industry*, 123:103308, 2020. 1

[8] Y. Chandio, V. Interrante, and F. M. Anwar. Reaction time as a proxy for presence in mixed reality with distraction. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 6, 9

[9] Y. Chandio, D. Romero, S. Elmalaki, and F. Anwar. What sensors see, what people feel: An exploratory study of subjective collaboration perception in mixed reality. *arXiv preprint arXiv:2504.16373*, 2025. 6

[10] A. Drahota and A. Dewey. The sociogram: A useful tool in the analysis of focus groups. *Nursing research*, 57(4):293–297, 2008. 2

[11] J. Du, Y. Shi, C. Mei, J. Quarles, and W. Yan. Communication by interaction: A multiplayer vr environment for building walkthroughs. In *Construction Research Congress 2016*, pp. 2281–2290, 2016. 1

[12] V. Echeverria, R. Martinez-Maldonado, and S. Buckingham Shum. Towards collaboration translucence: Giving meaning to multimodal group data. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–16, 2019. 2, 3, 9

[13] L. C. Freeman, D. Roeder, and R. R. Mulholland. Centrality in social networks: Ii. experimental results. *Social networks*, 2(2):119–141, 1979. 4, 5

[14] M. Gendron, D. Roberson, J. M. van der Vyver, and L. F. Barrett. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251, 2014. 6

[15] J. Gerup, C. B. Soerensen, and P. Dieckmann. Augmented reality and mixed reality for healthcare education beyond surgery: an integrative review. 11:1–18, jan 2020. 1

[16] M. Gonzalez-Franco, R. Pizarro, J. Cermeron, K. Li, J. Thorn, W. Hutabarat, A. Tiwari, and P. Bermell-Garcia. Immersive Mixed Reality for Manufacturing Training. *Frontiers in Robotics and AI*, 4, feb 2017. Publisher: Frontiers. 1

[17] E. T. Hall, R. L. Birdwhistell, B. Bock, P. Bohannan, A. R. Diebold Jr, M. Durbin, M. S. Edmonson, J. Fischer, D. Hymes, S. T. Kimball, et al. Proxemics [and comments and replies]. *Current anthropology*, 9(2/3):83–108, 1968. 3

[18] K. A. Hallgren. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23, 2012. 9

[19] R. A. Hanneman and M. Riddle. Introduction to social network methods, 2005. 4, 5

[20] K. Huang, T. Bryant, and B. Schneider. Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. *International Educational Data Mining Society*, 2019. 2, 3

[21] S. Hubenschmid, J. Wieland, D. I. Fink, A. Batch, J. Zagermann, N. Elmqvist, and H. Reiterer. Relive: Bridging in-situ and ex-situ visual analytics for analyzing mixed reality user studies. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2022. 2, 3

[22] P. Jansen, J. Britten, A. Häusele, T. Segschneider, M. Colley, and E. Rukzio. Autovis: Enabling mixed-immersive analysis of automotive user interface interaction studies. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–23, 2023. 2, 3

[23] C. Javerliat, S. Villenave, P. Raimbaud, and G. Lavoué. Plume: Record, replay, analyze and share user behavior in 6dof xr experiences. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2087–2097, 2024. 2, 3

[24] T. Kim, E. McFee, D. O. Olguin, B. Waber, and A. Pentland. Sociometric badges: Using sensor technology to capture new forms of collaboration. *Journal of Organizational Behavior*, 33(3):412–427, 2012. 2

[25] B. Kurdi, S. Lozano, and M. R. Banaji. Introducing the open affective standardized image set (oasis). *Behavior research methods*, 49(2):457–470, 2017. 6

[26] A. Lammert, G. Rendle, F. Immohr, A. Neidhardt, K. Brandenburg, A. Raake, and B. Froehlich. Immersive study analyzer: Collaborative immersive analysis of recorded social vr studies. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2, 3

[27] A. Léchappé, C. Fleury, M. Chollet, and C. Dumas. How to categorize collaboration during a collaborative puzzle-solving task? validation of collaboration profiles using multimodal data in virtual reality context. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–46, 2025. 2, 3

[28] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 8

[29] H. Ma, Z. Zhang, W. Li, and S. Lu. Unsupervised human activity representation learning with multi-task deep clustering. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–25, 2021. 2, 3, 5

[30] K. Mahadevan, Q. Zhou, G. Fitzmaurice, T. Grossman, and F. Anderson. Tesseract: Querying spatial design recordings by manipulating worlds in miniature. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2023. 2, 3

[31] J. Malmberg, S. Järvelä, and H. Järvenoja. Capturing temporal and sequential patterns of self-, co-, and socially shared regulation in the context of collaborative learning. *Contemporary Educational Psychology*, 49:160–174, 2017. 2

[32] Meta. Import Meta XR Packages|Meta Horizon OS Developers, 2024. 6

[33] Meta. Meta Quest Pro: Premium Mixed Reality|Meta Store, 2024. 6

[34] D. L. Mills. Internet time synchronization: the network time protocol. *IEEE Transactions on communications*, 39(10):1482–1493, 2002. 3

[35] J. L. Moreno. Foundations of sociometry: An introduction. *Sociometry*, pp. 15–35, 1941. 2, 4

[36] M. Nebeling, M. Speicher, X. Wang, S. Rajaram, B. D. Hall, Z. Xie, A. R. Raistrick, M. Aebersold, E. G. Happ, J. Wang, et al. Mrat: The mixed reality analytics toolkit. In *Proceedings of the 2020 CHI Conference on human factors in computing systems*, pp. 1–12, 2020. 2, 3

[37] T. Nguyen, D. Phung, S. Gupta, and S. Venkatesh. Extraction of latent patterns and contexts from social honest signals using hierarchical dirichlet processes. In *2013 IEEE international conference on pervasive computing and communications (PerCom)*, pp. 47–55. IEEE, 2013. 2, 3

[38] D. O. Olguın, P. A. Gloor, and A. S. Pentland. Capturing individual and group behavior with wearable sensors. In *Proceedings of the 2009 aaai spring symposium on human behavior modeling, SSS*, vol. 9. 2009. 2

[39] T. Overath, J. H. McDermott, J. M. Zarate, and D. Poeppel. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature neuroscience*, 18(6):903–911, 2015. 3

[40] C. O'Connor and H. Joffe. Intercoder reliability in qualitative research: Debates and practical guidelines. *International journal of qualitative methods*, 19:1609406919899220, 2020. 9

[41] P. Paulus. Groups, teams, and creativity: The creative potential of idea-generating groups. *Applied psychology*, 49(2):237–262, 2000. 1

[42] A. S. Pentland. The new science of building great teams. *Harvard business review*, 90(4):60–69, 2012. 2, 3, 9

[43] S. Puntambekar and R. Luckin. Documenting collaborative interactions: issues and approaches. pp. 737–738, 2023. 2

[44] A. Rorissa and S. K. Hastings. Free sorting of images: Attributes used for categorization. *Proceedings of the American Society for Information Science and Technology*, 41(1):360–366, 2004. 6

[45] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, dec 1980. 6

[46] J. Scott. Social network analysis: developments, advances, and prospects. *Social network analysis and mining*, 1:21–26, 2011. 4, 5

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

[48] I. S. University. 8.2 Defining Small Groups and Teams. Book Title: Introduction to Public Communication Publisher: Originally published by Indiana State University. 6

[49] W. Van Der Aalst. Process mining: Overview and opportunities. *ACM Transactions on Management Information Systems (TMIS)*, 3(2):1–17, 2012. 2

[50] S. Wasserman and K. Faust. Social network analysis: Methods and applications. 1994. 2

[51] S. Wasserman and K. Faust. Social network analysis: Methods and applications. 1994. 4

[52] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998. 4, 5

[53] W. Wolf, J. Launay, and R. I. M. Dunbar. Joint attention, shared goals, and social bonding. *British Journal of Psychology*, 107(2):322–337, 2016. 3

[54] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pp. 478–487. PMLR, 2016. 5

[55] D.-S. Yang, E. FitzGibbon, and F. Miles. Short-latency disparity-vergence eye movements in humans: sensitivity to simulated orthogonal tropias. *Vision Research*, 43(4):431–443, 2003. 3

[56] Y. Yang, T. Dwyer, M. Wybrow, B. Lee, M. Cordeil, M. Billinghurst, and B. H. Thomas. Towards Immersive Collaborative Sensemaking. *Proceedings of the ACM on Human-Computer Interaction*, 6:722–746, nov 2022. 1, 2, 3

[57] C. Yin, S. Zhang, J. Wang, and N. N. Xiong. Anomaly detection based on convolutional recurrent autoencoder for iot time series. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(1):112–122, 2020. 5

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

[59] Y. Zhang, J. Olenick, C.-H. Chang, S. W. Kozlowski, and H. Hung. 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):1–22, 2018. 2

[60] L. Zhao, V. Echeverria, Z. Swiecki, L. Yan, R. Alfredo, X. Li, D. Gasevic, and R. Martinez-Maldonado. Epistemic network analysis for end-users: Closing the loop in the context of multimodal analytics for collaborative team learning. In *Proceedings of the 14th learning analytics and knowledge conference*, pp. 90–100, 2024. 2