Multimodal Group Interaction Dataset Descriptor

Overview

This dataset contains processed multimodal interaction data from collaborative VR sessions, organized into four primary output directories. Each directory contains different analytical perspectives on the same underlying group interaction data.

Dataset Structure

Core Directories

`pairwise_outputs/`

Content: Pairwise interaction features between participant dyads
Key Files: `group-{N}_pairwise_win{X}s.parquet` (per-group) and combined files
Features: Per-second dyadic metrics including speaking entropy, dominance ratio, proximity measures, shared attention, approach rates
Format: Parquet files with ZSTD compression
Temporal Resolution: Configurable windows (default: 1-second)
NaN Strategy: Either 'fill' with safe defaults or 'drop' missing data rows

`session_output/`

Content: Full-session network analysis and sociometric data
Key Files: `session_metrics.csv`, `session_edges.csv`, `session_nodes.csv`, visualizations
Scope: Complete session-level (window=0) network metrics across four modalities
Networks: Conversation, proximity, shared attention, and fused multimodal networks
Outputs: Network visualizations (PNG), structural metrics, edge lists, node centralities

`task_metrics_output/`

Content: Task performance and completion analytics
Key Files: `task_metrics_detailed.json`, summary files (CSV, HTML, TXT, MD)
Data Source: Object placement/grouping logs from GroupingLogs directories
Metrics: Completion times, interaction counts, object categorization patterns, participant activity timelines
Format: Multiple output formats for different use cases

`windowed_output/`

Content: Temporal network analysis with sliding windows
Key Files: `windowed_metrics.csv`, `windowed_edges.csv`, `windowed_nodes.csv`, visualizations
Temporal Resolution: Sliding window analysis (default: 32s windows, 16s stride)
Networks: Same four modalities as session_output but computed over time windows
Visualizations: Sampled network graphs showing temporal evolution

Data Modalities and Network Types

Network Analysis Modalities

All network-based outputs (`session_output/`, `windowed_output/`) contain four modality types:

`conversation`

- **Source**: Speech activity logs with start time and duration data

- **Network Type**: Directed graph (speaker → listeners)
- **Edge Weights**: Speaking duration directed toward other participants
- **Metrics**: Density, reciprocity, eigenvector centrality, clustering

`proximity`

- **Source**: VR headset position coordinates (X, Y positions)
- **Network Type**: Undirected graph
- **Edge Weights**: Count of time steps within proximity threshold (1.5 feet)
- **Metrics**: Density, clustering, eigenvector centrality

`shared_attention`

- **Source**: Eye tracking logs from virtual object hover events
- **Network Type**: Undirected graph
- **Edge Weights**: Total seconds of simultaneous object attention
- **Metrics**: Joint attention duration, attention synchronization

`fused`

- **Source**: PCA-weighted combination of all three modalities
- **Network Type**: Directed or undirected (depends on component networks)
- **Edge Weights**: Weighted sum using data-driven modal weights
- **Purpose**: Integrated multimodal interaction strength

Pairwise Feature Categories

The `pairwise_outputs/` directory contains dyadic interaction features:

Speaking Dynamics

- **speaking_entropy**: Turn-taking unpredictability (Shannon entropy)
- **dominance_ratio**: Speaking time imbalance between participants
- **speaker_switch**: Turn-taking transition events
- **floor_streak**: Consecutive speaking duration patterns

Spatial Interaction

- **dist_mean**: Average physical distance between participants
- **prox_binary**: Time spent within proximity threshold
- **approach_rate**: Movement velocity toward/away from partner
- **dist_accel**: Acceleration patterns in interpersonal distance

Shared Attention

- **material_diversity**: Entropy of jointly attended virtual objects
- **joint_att_count**: Frequency of simultaneous object fixations
- **shared_att_ratio**: Proportion of overlapping attention time

Temporal Patterns

- **burst_switch_rate**: Short-term turn-taking frequency
- **burst_overlap_rate**: Speaking overlap in recent windows
- **resp_latency**: Response delay after partner's speech

Task Performance Categories

The `task_metrics_output/` directory contains collaboration task metrics:

Completion Metrics

- **completion_time_seconds**: Total task duration
- **total_interactions**: Count of object placement events
- **session_timeline**: Relative timestamps from session start

Participation Patterns

- **participant_interactions**: Per-participant activity counts
- **first/last_interaction_time**: Temporal engagement boundaries
- **unique_objects/categories**: Diversity of task engagement

Feature Categories

File Formats and Data Structure

Pairwise Data (`pairwise_outputs/`)

- **Format**: Parquet files with ZSTD compression
- **Structure**: One row per dyad-window combination
- **Columns**: `group`, `pair`, `window_idx`, `window_start`, plus ~20 interaction features
- **Temporal**: Configurable window size (typically 0.5-1.0 seconds)
- **Pairs**: All possible dyad combinations within each group (A-B, A-C, A-D, B-C, B-D, C-D)

Network Data (`session_output/`, `windowed_output/`)

Metrics Files:

- Columns: `group`, `window`, `modality`, `t_start`, `t_end`, network metrics
- Window=0 for session-level, window>0 for temporal windows
- Metrics: `density`, `avg_clustering`, `eigenvector`, `reciprocity`

Edges Files:

- Columns: `group`, `window`, `modality`, `source`, `target`, `weight`
- Represents network connections with interaction strength weights

Nodes Files:

- Columns: `group`, `window`, `modality`, `participant`, centrality measures
- Centralities: `betweenness`, `degree`, `eigenvector`, `closeness`

Visualizations:

- PNG network graphs with edge thickness proportional to weights
- Nodes positioned in standardized 2×2 grid layout
- Participants labeled A, B, C, D

Task Data (`task_metrics_output/`)

- **JSON**: Complete structured data with metadata and interaction logs
- **CSV**: Tabular summaries for quantitative analysis
- **HTML/TXT/MD**: Human-readable reports with completion statistics
- **Timestamp Format**: Relative seconds from session start (anonymized)