COSI-LAB: CONFERENCE LIVING LAB FOR MODELING MULTI-PERSPECTIVE MULTIMODAL SOCIAL INTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

COSI-Lab presents a multimodal multi-sensor dataset of a real interdisciplinary scientific workshop with 32 academics at an international conference. It captures ecologically valid social interactions in a weakly scripted setting consisting of two 30-minute mingling sessions with real professional and social consequences. We argue that intelligent systems could better handle subjective perceptions by modeling their multiplicity not as label noise but as an explainable perspective-driven reasoning process. We focus on the Apparent Intent Inference (AII) problem as determined by ex-situ observers and conceptualize intentions to be independent of the future. We contribute a novel annotation process for AII that accounts for a perceiver's own interpretative tendencies, quantitative and qualitative analyses of intent narratives with respect to diversity, plausibility and grounding, benchmark tasks for AII and relevant contextual factors, speech quality audio for all participants as well as privacy-preserving multi-modal environmental and wearable sensor data, enabling lexical and nonverbal behavior analysis; and the coupling of self-reported participant goals (30 minute to 3 hour) with annotated AII (seconds).

1 INTRODUCTION

Human intention-aware systems are crucial for understanding future outcomes (Belardinelli, 2023; Rudenko et al., 2020; Rasouli et al., 2019; Jing et al., 2025). We argue that to understand future outcomes, and those that could influence the affective state of an observed subject, such systems need to reason about the evidence for an intention perception (Jing et al., 2025) rather than just detecting them (Rasouli et al., 2019; Zhang et al., 2022). And do this in in-the-wild ecologically valid settings.

Following Bratman (1987), intent forms part of future action planning coupled with the belief the action can be carried out. We focus on interpreting behavior to study *overt intentional action* (Mele, 1992) as a proximal phenomenon (intentions related to immediate actions). We name this the Apparent Intent Inference (AII) problem. Embracing the subjectivity of AII, we take a perspectivist stance (Cabitza et al., 2023; Dudzik et al., 2019) by coupling annotators' trait, demographic data, and cognitive style with their annotations. Finally, we also consider distal intentions/goals which relate to the future but are not necessarily situated in the immediate situation.

Following script theory from Schank & Abelson (1977), our AII problem focuses on a weakly scripted networking mingling events (Raman et al., 2022) where the rules of play are not stated apriori(e.g. when to talk to whom, on what topic, or why are emergent). This contrasts with more popular strongly scripted settings, externally pre-determined rules e.g. rules of the road (Rasouli et al., 2019), situational comedy shows (Zhang et al., 2024).

Thus far, weakly scripted multi-party datasets have only been developed for detection and localization tasks Raman et al. (2022); Cabrera-Quiros et al. (2021); Alameda-Pineda et al. (2015), without access to full verbal and nonverbal information that could disentangle nuances in conversational intentions. To this end, we introduce the Conference Social Intentions Living Lab (COSI-Lab) which captures real-life weakly scripted and fully synchronized multisensor multimodal data of 32 academic participants interacting over the course of 3 hours. We record both self-reported distal intentions ($\sim$ 30 minutes) (e.g. to lay foundations for future collaborations) and crowd-source third-party proximal intention perceptions ($\sim 1 - 30$ seconds) (e.g. to show interest in the current

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

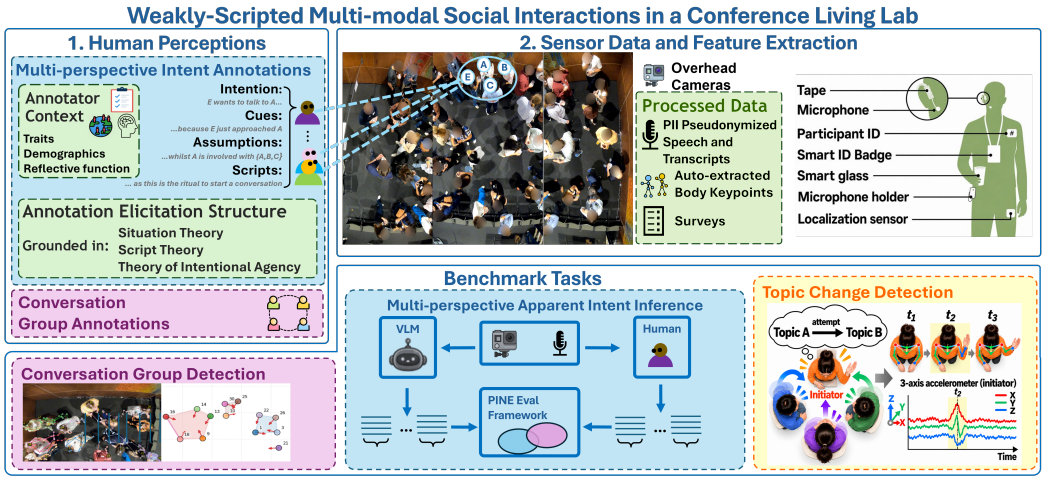


Figure 1: COSI-Lab: The conference living lab for studying multi-perspective social intentions.

conversation). Aside from designing a workshop living lab, we contribute annotations and data processing for three key benchmark tasks: (i) subjective multimodal AII inference using our proposed perspectivist intention narrative evaluation (PINE) (ii) multimodal social involvement estimation, and (iii) multimodal topic change estimation. Consequently, COSI-Lab is designed around these desiderata for (social) intent-aware systems and dataset design:

D1. Multiple plausible intent inferences can exist simultaneously: outputting a single true (social) intent may not always be desirable; by design, it is safer and more ethical to allow for multiple plausible hypotheses about a person's intention to be generated and used at inference time. For example, a system should reason across multiple plausible hypotheses of a human's intentions even if that human is unwilling or unable to provide their true intent to the system for any reason (e.g. practical, cognitive, or privacy constraints).

D2. Account for unrealized intentions not manifested in future outcomes: Exclusively looking to the future to label intentions in the past risks biasing perceptions towards outcomes. Intentions that are never realized could affect a person in ways that would be inaccessible to the system.

D3. Need explanations that account for ambiguity, subjectivity, and context dependence: how does the observable evidence available, any intermediate interpretations by a perceiver, and their own predispositions (e.g. traits and demographics) relate to their inference? We need measures to capture qualities of explanations grounded in existing theory in cognition and situational awareness.

D4. Account for the hierarchical structure of human intention: Individuals rarely pursue goals in isolation. Instead, their behavior is continuously shaped by nested layers of motivations at different levels of abstraction (Kruglanski et al., 2015), often involving concrete proximal outcomes within a situation (e.g., "Introduce yourself to X") in service of distal ones (e.g., "Be a successful networker").

2 Related Work

Table 1 summarizes the comparison between COSI-Lab and other multimodal human behavior and intention-related datasets.

Outcome-oriented pedestrian intention. Pedestrian intention datasets such as PIE (Rasouli et al., 2019) formulate intention as short-term outcomes. PSI (Jing et al., 2025) enriches intention labels with multiple subjective, free form explanations from the driver perspective. However, most other datasets rely on strongly scripted settings where intentions are prescribed with a few classes (Khindkar et al., 2024; Jing et al., 2025).

Reasoning of intention. IntentQA (Li et al., 2023) uses a guided question-answering approach to probe the intention behind daily actions. However, the multiple-choice format limits intention labels to a single perspective. Ego-Exo4D (Grauman et al., 2024) provides expert descriptions with first-person reflections on skilled activities. Recent work has applied MLLMs to evaluate social understanding

Table 1: Comparison of human behavior datasets with intention and reasoning annotations.

	Data capture					Desiderata			
Dataset	**Scenario (Scripting)**	**Intention Representation**	**Narrative Annotation**	**Modalities**	**Annotation Size**	**D1**	**D2**	**D3**	**D4**
PIE++ (Khindkar et al., 2024)	Pedestrian-Vehicle (Strong)	Categorical	Multi-annot. prediction	Video	5k instances of 1-2 sec.	✓	✓	×	N/A
PSI (Jing et al., 2025)	Pedestrian-Vehicle (Strong)	Categorical Free-form	Multi-annot. subjective	Video	196 vids of 15 sec.	✓	✓	✓	N/A
IntentQA (Li et al., 2023)	Skilled activities (Strong)	Categorical guided Q&A	Multi-annot. agreement	Video	5k vids of 44 sec.	×	(✓)	(✓)	✓
Ego-Exo4D (Grauman et al., 2024)	Skilled activities (Strong)	Expert free-form/reflections	Multi-perspective & Multi-annotator	Video, eye, IMU	5k takes of 2.6 min.	✓	×	✓	✓
I2M (Umagami et al., 2026)	Daily living (Strong)	Free-form descriptions	Single-perspective logs	Video, IMU	215 seqs of 2.8 min.	×	×	×	✓
MIntRec2.0 (Zhang et al., 2024)	TV shows (Full)	Categorical, no reasoning	Multi-annotator agreement	Video	1.2k dialogues of 3 sec.	×	×	×	×
Ours (COSI-Lab)	Social interaction (Weak)	Structured Free-form & counterfactual	Subjective/Multi-annot.	Video, Audio, IMU	570 vids of 30s (4.6k annotations)	✓	✓	✓	✓

(✓): no explicit stance is given in paper; we provide a best guess based on partial or implicit fulfillment. N/A: it was not relevant to intended scope of paper.



Figure 2: Workshop procedure

(Zhou et al., 2024), multiparty conversation (Wu et al., 2025), affect perception (Lian et al., 2025), multimodal intention recognition (Zhang et al., 2025; Peng et al., 2025), and reasoning traces during social inference (Mathur et al., 2025). However, these datasets and benchmarks predominantly focus on predefined tasks where the structure of activity is strongly-scripted in nature, limiting the possible emergent intentions that could arise more open weakly scripted settings.

Multimodal human behavior. Multimodal behavior datasets (Raman et al., 2022; Cabrera-Quiros et al., 2018; Agrawal et al., 2025) are typically not focused on social intentions, and also lack either ecological validity or broad coverage of modalities. MIntRec2.0 (Zhang et al., 2024) labels categorical social intentions in scripted TV dialogue, limiting analysis of spontaneous interaction. PARSEL (Hrkalovic et al., 2025) captures audiovisual dyadic conversations and cooperative partner choices, but represents a narrower (pre-arranged dyadic-only) setting for partner selection than COSI-Lab.

3 Data Collection and Processing

3.1 Procedure Design

We developed a multi-stage data collection framework that captures the interplay between stable individual traits, dynamically evolving individual goals and affective states, and multimodal networking behaviors. Data were gathered as part of a pre-conference workshop to maximize the ecological validity of the participants' social intentions within a real networking event. Data collection venue and participant recruitment logistics was supported by the conference organization, however the board and all participants were unaware of our specific research questions.

Participants and recruitment We recruited 32 participants, primarily academic researchers from social or computer science, through official conference organization announcements, social media posts, and word-of-mouth. Participants were informed in advance via email and the workshop website about the study procedures, the full set of recorded modalities, and the data sharing conditions. They provided explicit consent for both recording and data sharing.

Workshop procedure The data collection followed a structured multi-stage design (see Figure 2). During the pre-event phase, participants completed questionnaires assessing demographic information

and psychometric measures to establish baseline traits that influence social intent. The in-event phase consisted of two 30-minute mingling sessions where goals, reflections on goals, and affective states (Broekens & Brinkman, 2013) were provided by participants before and after each mingle. A post in-event survey captured interaction impressions, intent to connect, and interpersonal acquaintanceship. three months later, a post-event survey was used to identify realised new connections.

Trait and Demographics. Surveys were conducted at different stages of the workshop (see figure 2). Specifically, self-monitoring personality (Lennox & Wolfe, 1984), social anxiety (Connor et al., 2000), communication and interaction styles (Vissa, 2012; Rubin & Martin, 1994; Cheek & Buss, 1981), and HEXACO personality traits (Ashton & Lee, 2009)). Items were aggregated into one value per measure per participant. To reduce identifiability, age was quantized, and country of birth, country of residence, nationality, and ethnicity were collapsed into broader categories where appropriate.

3.2 Data Capture Setup and Post Processing

We record multimodal synchronized data for the participants during the 2 mingle sessions. We extended the privacy-by-design fully-synchronized capture set-up of Raman et al. (2022; 2020) by adding a smart cup, speech quality microphones and ultra-wideband (UWB) wearable badges as described below. See synchronization method and details of the data capture in Appendix Section B.

Cameras. Video was recorded with GoPro Hero 7 cameras. In both sessions, 5 ceiling mounted cameras provided overhead views and 4 corner cameras provided elevated side views. Due to privacy considerations related to facial visibility, only overhead views are included in the dataset. Cameras are calibrated by estimating intrinsics from ChArUco-board videos and, after installation, extrinsics from synchronized footage of ChArUco boards moved through the shared interaction space. In post processing, individuals were segmented in each frame using SAM3 (Carion et al., 2025), and the resulting masks guided ViTPose (Xu et al., 2022) to detect 2D body keypoints. Using the estimated camera parameters, we reconstructed these keypoints in the shared 3D room coordinate frame, which gives 3D keypoint extraction. Further details are provided in Sections B.2 and C.2, with example calibration footage shown in Figure 7.

Smart Wearables. The Midge (Vargas Quiros et al.) is an open-source wireless wearable with a microphone, 9-axis Inertial Measurement Unit, and Bluetooth for proximity sensing. Each participant wore one hung around their neck like a conference ID badge. A second one was attached to a drinking glass, which they could use during the workshop. To obtain an absolute position in 3D space, the participants were also fitted with UWB badges placed in a pocket. We used 32 Sewio Leonardo Personal tags Sewio (b) and 5 Vista Omni anchors Sewio (a). Note that UWB indoor positioning does not work well in high density situations, despite commercial claims (see preliminary error estimation in Appendix C). Audio recorded by the smart wearables were pseudonymized (see description below).

Microphones. High-quality audio was captured using 32 Sennheiser wireless microphones attached to a participants' cheek with Lavalier tape. For audio pseudonymization, full recordings were transcribed directly without speaker diarization to preserve original audio. PII detection was performed using Microsoft Presidio (Asimopoulos et al., 2024), which has performed well in prior benchmarks. Detected names, email addresses, and phone numbers were mapped back to their corresponding audio intervals and rendered unintelligible using a 1.25 kHz low-pass filter. For the mingle sessions, speaker diarization was additionally performed to identify and isolate the microphone wearer, from which participant-specific pseudonymized transcripts were generated.

4 Data Annotation Design

4.1 Design of apparent intent annotation process

As mentioned before, AII is inherently subjective; different observers may construct different interpretations of the same situation based on personal differences, related to experiences, abilities, cultural background, or personal disposition. Moreover, participants may maintain multiple concurrent proximal intentions, each influencing their planning behavior as they work toward particular goals and sub-goals(Kruglanski et al., 2015). This leads to open challenges for evaluating the multiplicity of plausible intention narratives that could be perceived by external observers.

Reasoning about intentions in weakly scripted settings To reason about intent narratives, we leverage situation (Rauthmann & Sherman, 2021) and script (Fischer et al., 2013) theory. These are the observable *cues* of the scene and assumptions individuals make about how a situation unfolds. Using the 3C's taxonomy (Rauthmann & Sherman, 2021): *Cues* are directly perceivable elements such as actions, gestures, speech, objects, sounds, and spatial context; *characteristics*, which reflect the perceived affective qualities of the interaction inferred from cues (e.g., tense or friendly); and *classes* refer to the situation type including expectations about roles and relationships (e.g., a job interview, casual conversation, or negotiation). Finally, *social scripts* (Fischer et al., 2013) describe the expected sequence of actions performed by people in the situation as well as the role they play. This implicit "playbook" guides how interactions typically proceed, such as greeting rituals, introducing, conversing, and closing. Together, cues and these layered assumptions shape the intention narratives. On top of this, the annotators were instructed to consider cues, situation characteristics, situation class, and social scripts.

Intention annotation procedure To address the nature of subjectivity in narrative construction, Analysis-Holism Scale(Martín-Fernández et al., 2022) and Reflective Functioning Questionnaire (Fonagy et al., 2016) questionnaires were completed in addition to surveys of same measures as the participant pre-workshop survey. This allows for annotator and participant profile comparison analysis. The annotation procedure was as follows. Six annotators observed a 30-second overhead video clip of a conversation with the speech for all conversation participants. Each annotator was assigned a specific participant and asked to describe the intention they perceive that individual to have at a given moment. They were then asked to justify their interpretation, provide confidence scores for both the intention and its explanation, and rate the intensity of the intention (i.e., how strongly it appears to drive the participant's behavior). They marked in time the start and end of the intention. Multiple concurrent intent annotations as well as changing intentions could also be annotated.

Pilot and Full Study A pilot study (n=48) explored how two varying levels of instruction affected annotations; fully free-form (A) vs. structured free-form (B). While (A) led to confusion, both conditions revealed that annotators did not naturally incorporate explicit assumptions into their explanations. Thus, final instructions were updated to explicitly define the relevant cues and assumptions, requiring annotators to incorporate them into their reasoning. The final instruction set and both pilot instructions can be found in Appendix D with a description of the annotation process. The full study contained 65 annotators and 7,176 human annotations. Annotators were recruited via Prolific and completed the task online using a custom annotation interface adapted from Covfee (Vargas).

4.2 Design of conversation group annotation

Social conversational involvement is a key contextual scoping factor demarcating where social influence and intent might exist. To contextualize social intention within an interaction, we consider the notion of conversation floors, which could be approximated by a group of people sharing a common spatial attentional focus (Edelsky, 1981). To operationalize this joint attention focus, we also take inspiration from Kendon's theory on F-formations (Kendon, 1990) which translates this group into physical arrangements. Conversation groups are crucial to understand social involvement and social intentions, as they define the spatial and social boundaries of "who is talking with whom". Crucially, we depart from prior work which labeled F-formations identified as groups who stand with their lower body facing each other (Hung & Kröse, 2011; Cabrera-Quiros et al., 2021; Raman et al., 2022). Instead, conversation groups were annotated by considering jointly focused head orientations. Annotations were obtained for both mingle sessions by marking any *changes* in each conversation group, including when a composition change occurred as well as the new group members. This yielded *continuous* annotations at the same granularity as camera frame rate (59.94Hz).

4.3 LLM-based topic redirection labeling

In natural multiparty conversations, participants may attempt to change the topic as part of their ongoing goals. Such attempts may not always be taken up by others. The ability to avoid or enforce certain topics can be a sign of wishing to avoid emotional discomfort (Caughlin & Afifi, 2004), assert dominance (Reid & Ng, 2000), or manage self-presentation (Holtgraves, 1992). Following Button & Casey (1985); Schegloff (2007), we label *topic-redirection attempts* separately from their subsequent

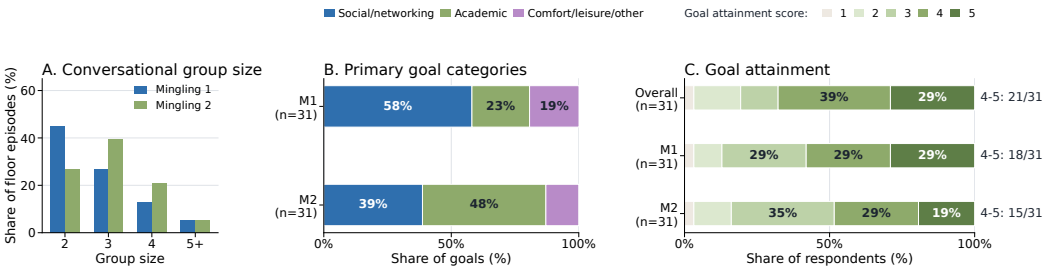


Figure 3: Interaction structure and survey-reported goals across the two mingling sessions.

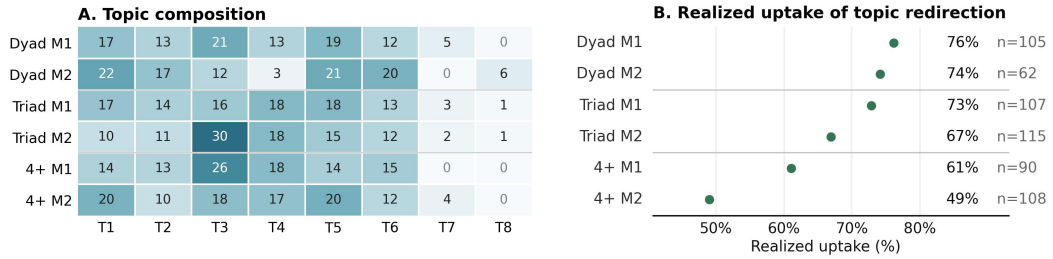


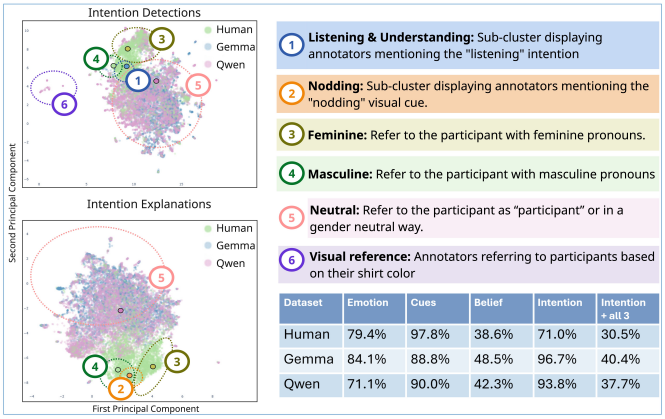
Figure 4: Conversation topic composition and redirection uptake by session and group size. M1/M2 denotes Mingle 1/2. Panel A: topic-family shares (T1: background; T2: research domains; T3: methods/data; T4: team/organization; T5: networking; T6: personal; T7: gender/diversity; T8: other). Panel B: realized uptake among valid topic-redirection attempts.

uptake. Labels were auto-generated from pseudonymized transcripts using a locally hosted Gemma 4 31B model in two stages. First, each target turn was judged using only the preceding context and the current turn; later outcomes did not influence whether a redirection attempt was identified. Then, the identified attempt and its proposed topic were stored. The second stage examined the subsequent conversation to determine whether the group took up that topic. The resulting labels are used in both Fig. 4B and Benchmark 3 (Section 6.3); full procedure and prompts are provided in Appendix E.3.

5 Descriptive Statistics

Annotator profile varies intent reasoning To test whether who an annotator is shapes the intention narratives they produce, we linked surveys to annotations for 52 annotators (4,262 narratives over 568 clips). Each pair of annotators sharing a clip contributes the semantic similarity of their narratives, regressed on the Euclidean distance between their standardised 29-dimensional survey vectors, in a mixed model with clip and annotator as crossed random effects, the annotator entering for both members of each pair (Baayen et al., 2008; Browne et al., 2001), fitted by REML in `lme4` (Bates et al., 2015). The model was then compared against 2,000 permutations of whole survey profiles across the 52 annotators (Winkler et al., 2014). Shuffling at the annotator level keeps the clip and co-annotation structure fixed as well as the correlation within the survey scales, while breaking the profile–narrative link. We found that annotators with more similar profiles wrote more similar *explanations* of intention. Moving one standard deviation (SD) from the mean profile lowers narrative similarity by $\beta = 0.0102$ (SE $= 0.0025$), a fall of 0.23 residual SD across the 10th-to-90th-percentile range of profile distance. This holds under all three narrative similarity metrics (permutation $p = .009$, $.007$ and $.011$ for cosine, overlap (Parupudi, 2025) and HTS). Annotator identity owns 29% of the variance in explanation similarity against 20% for the detected intention.

Conversational group structure and survey goals Figure 3 compares annotated conversation-group structure with survey-based primary goal categories and self-rated goal attainment. Mingle 1 was more dyadic, whereas Mingle 2 contained more three-person and larger conversational groups. Among 31 respondents, primary goals differed descriptively across phases, from mostly social/networking in Mingle 1 to more academic or targeted in Mingle 2. On the 1–5 goal-attainment scale, mean scores were $3.71(1.10)$ for Mingle 1 and $3.48(1.06)$ for Mingle 2; ratings of $4--5$ were given by $18/31$ and $15/31$, respectively. Linking these stated goals to the annotated intentions, only 76.7 % of annotated participant intentions matched at least one of their reported goals, with a higher



Dataset	Emotion	Cues	Belief	Intention	Intention + all 3
Human	79.4%	97.8%	38.6%	71.0%	30.5%
Gemma	84.1%	88.8%	48.5%	96.7%	40.4%
Qwen	71.1%	90.0%	42.3%	93.8%	37.7%

Conversation Group Detection, F1@1

Dataset	DANTE	LSTM
Mingle 1	0.766 ± 0.010	0.678 ± 0.141
Mingle 2	0.756 ± 0.228	0.826 ± 0.126
Overall	0.763 ± 0.151	0.737 ± 0.125

Successful Topic Transition, AUROC

Method	Keypoints	Accel.	Keypoints + Accel.
LR (initiator)	0.556 ± 0.026	0.546 ± 0.056	0.568 ± 0.043
TCN (initiator)	0.596 ± 0.074	0.483 ± 0.049	0.565 ± 0.036
TCN–DeepSets (group)	0.539 ± 0.044	0.488 ± 0.058	0.501 ± 0.091

Figure 5: Left: Semantic similarity of the annotations and presence of emotions, cues, beliefs, intentions, and all 4 simultaneously. Right: Conversation group detection results (mean ± std over camera-wise 5-fold tests) and successful topic-transition detection (mean ± std over five folds).

match rate for participants who reported partially or fully achieving their goals (77.6%) verses those who reported not achieving them (73.2%). This could be an effect of the specificity of self-reports.

Conversational topics and redirection uptake Figure 4 provides an LLM-assisted summary of topic composition and realized uptake of topic-redirection attempts by session and group size. Topic composition varied across both sessions and group sizes, with uptake generally higher in smaller groups. The LLM-assisted annotations were evaluated against adjudicated human annotations on stratified subsets. Topic-composition estimates had a mean absolute error of 12.1 percentage points, while the redirection labels achieved accuracy/macro-F1 scores of 0.800/0.762 for attempt identification and 0.909/0.906 for uptake outcome classification. Details of the human-validation procedure and results are provided in Appendix E.4.

Reported ties over time Surveys captured self-reported ties before, during, and after the workshop. Respondents reported knowing 3.8 (out of 32) attendees on average before the event, mostly as weak ties. During the workshop, they reported 14.9 interaction partners on average, 78.0% of whom were previously unfamiliar. This pattern is notable because people in networking settings often remain with people they already know (Ingram & Morris, 2007). Three months later, post-event respondents recalled on average 9.5 other attendees and maintaining contact with 2.4 people.

6 Benchmark Tasks

6.1 Intention Estimation

Qualitative analysis of human and model narratives Intention narratives can be analyzed qualitatively:- First, to assess the diverse representation of human-perceived intentions, we measure semantic similarity across annotations using Hyperbolic Tangent Similarity (HTS) (Parupudi, 2025). Second, we analyze narrative specificity by automatically parsing the narratives to identify the presence of dependent variables aligned with cues, assumptions, and beliefs. Third, for a subset of intention annotations, we assess narrative grounding by manually verifying whether the cues referenced in each annotation are present in the corresponding video clip or are hallucinated.

We zero-shot prompted VLMs (Gemma3 E4B (Team et al., 2025) and Qwen 3 Omni 30B Instruct (Xu et al., 2025)) with the personas of the annotators, built from their survey responses, and the same annotation task, asking it to generate the intention narrative that an annotator with this persona would produce. Figure 5(Left) presents the semantic similarity of intention detections(top) and their explanations (bottom), comparing model outputs (gemma3: blue, qwen3: pink) with human annotations (green). A clear clustering separation emerges between model and human outputs, with overlaps especially in intention explanations, while all human annotations, occupy a broadly shared semantic space. Across both plots, annotations further group into three main clusters based on reference style: male pronouns, female pronouns, and gender-neutral terms. Human annotators

predominantly use gendered references, whereas the model more often defaults to neutral phrasing, marking a key distinction. Smaller sub-clusters can also be discerned revealing similarities in perceived cues, intentions and intention reasoning.

Specificity analysis of human and model narratives The table in Figure 5(Left) shows how often key components: cues, beliefs, emotions, and goal-directed intentions appear in annotations. Each percentage indicates the proportion of annotations that mention the component at least once (parsed by Gemma 4 31B (Team et al., 2026)). Setting intentions aside, the model distribution of components broadly followed that of the human annotators, all three referring more often to cues and perceived emotions than to beliefs, though the models attributed beliefs more frequently and grounded their narratives in observable cues somewhat less often. Human annotators specifically focused on cues, having intentions mostly refer to actions (both video and audio) present in the clip. The models were far more explicit about intention, stating a goal in almost every narrative. This contrast arises because the models followed the instructions more closely and the implied intentions in the human annotations were sometimes missed by the LLM parser. Together, these counts give a first, coarse sense of the coverage gap between model and human narratives that we quantify more directly below.

Evaluating model generations Our Perspectivist Intention Narrative Evaluation (PINE) framework compares model and human narratives along three dimensions: *grounding*, *human alignment*, and *coverage*. First, applying the grounding test to a sample of intention narrative annotations ($n = 83$), every human narrative referenced cues genuinely present in the clip, as did 97.6% of model narratives. The models nonetheless frequently misattributed a cue to the wrong participant and relied almost exclusively on audio, seemingly unable to make much use of the overhead video.

Second, we assess *human alignment*. Since multiple perspectives on the same situation can be equally valid, there is no single ground-truth narrative to score against. We instead ask whether each persona-conditioned model generation is close to at least one plausible human perspective, scoring it against every annotation that human annotators produced for the same video. Our *Persona Alignment Score (PAS)* thus evaluates model performance by embedding all narratives and measuring the semantic similarity between model and human embeddings using HTS (Parupudi, 2025). For each model narrative we keep the similarity of its nearest human neighbor, averaged over all N generations: PAS $= \frac{1}{N}\sum_{i=1}^{N} \max_j \text{HTS}(v_i, h_{ij})$, where v_i is the ith model narrative and h_{ij} the jth human annotation of the same video (validation of score in Appendix D.1). Running this comparison across both models, we found that qwen3 scored higher than gemma3 on both narrative fields (PAS 0.419 vs. 0.388 for intention detections and 0.438 vs. 0.416 for explanations), though both remained below the agreement reached between independent human annotators of the same video (0.492 and 0.494 respectively).

Third, *coverage* assesses whether the model's generations span the human distribution of perspectives of the given dataset without collapsing onto a narrow subset or spilling outside it. Thresholding pairwise HTS similarities, calibrated on human-human nearest-neighbor distances (0.499 for intention detections, 0.509 for explanations), we compute precision (proportion of VLM outputs matched to a human annotation, guarding against generations outside the human distribution) and recall (proportion of human perspectives matched to a VLM output, guarding against collapse). We complement this pairwise view with a distributional comparison (Wasserstein distance and MMD between human and model embeddings). Both models cleared this bar rarely: precision reached 0.175 (gemma3) and 0.232 (qwen3) for detections and 0.197 / 0.258 for explanations, with recall lower still at 0.117 / 0.198 and 0.133 / 0.224, qwen3 being the stronger of the two throughout and most markedly in recall. The distributional comparison separated the models far less (Wasserstein 0.019 vs. 0.018 for detections and 0.020 vs. 0.016 for explanations; MMD 0.103 vs. 0.093 and 0.089 vs. 0.061), indicating that model and human embeddings span comparable overall distributions even where they separate locally (Figure 5), individual generations seldom matching any single human perspective at the level humans reach with each other.

6.2 Conversation group detection

As described in section 4, we run conversation group detection tasks to scope the context for social intentions. Following previous works (Hung & Kröse, 2011; Setti et al., 2015; Vascon et al., 2016), we formulate group detection by leveraging each person's features, and most typically, their positions and orientations. We evaluate two baseline architectures: the graph-based DANTE (Swofford et al.,

2020) and a temporal LSTM-based model (Tan et al., 2022), by employing the commonly used F1 metric (Vascon et al., 2016), and measure it under threshold $T = 1$. We construct the features (positions and orientations) from raw videos with SAM3 (Carion et al., 2025) and ViTPose (Xu et al., 2022) models, and provide the details of this process in Appendix C.2 and F.2. The detection results are compared to the conversation group annotations in Section 4.2. Figure 5(Right) shows the results; the performances are high on F1. While the high scores on the baselines don't necessarily imply that the keypoints are accurate in representing the spatial coordinates, it does demonstrate that extracted spatial relationship across cameras and events are consistent and transferrable. We report more results including a cross-camera train-test setup in Appendix F.2, with stable results as well.

6.3 Topic change detection

Prior work has shown that gesture, gaze, and posture can reflect discourse structure and help identify discourse shifts (Quek et al., 2000; Cassell et al., 2001). Motivated by this, we examine whether successful topic transitions can be recognized from privacy-preserving body keypoints and wearable acceleration. For this task, we formulate a binary classification task to distinguish successful topic transitions from non-transitions, where the negative class includes both unrealized redirection attempts and ordinary conversation. We compare logistic regression (LR) and a temporal convolutional network (TCN) using the initiator's nonverbal behavior, and a role-aware TCN–DeepSets model that additionally incorporates the behavior of the other conversation members. Each baseline is evaluated using upper body keypoints, and/or acceleration from the smart conference badge with five-fold evaluation. Figure 5 reports AUROC for the successful topic transition detection task. Benchmark details, further results, ablation studies, and a secondary attempt-detection task are provided in Appendix F.3. Overall, the results suggest that successful topic transitions are reflected, albeit modestly, in privacy-preserving nonverbal behavior.

7 Limitations

Non-overt intentions and temporal limits While we account for the multiplicity of overt intentions, our framework does not yet account for intentions that may not be observable from observable planning behavior, either within or beyond the bounds of our 30s segments. While intent-outcome relationships can be captured within these shorter segments and the segmentation is designed to enable independent samples for training models, relationships spanning multiple segments may be missing. We therefore warn potential users to be mindful of this in future work on longer term AII.

Long term observee self-reported goals and short term third-party annotated intentions While COSI-lab enables initial studies linking distal intentions and outcomes with proximal ones, future users must be mindful of the conceptual distinction between self-reports and third-party annotations. Neither is a definitive representation of the truth. Rather, they can be used for developing hybrid intelligent systems (Akata et al., 2020) where humans and agents jointly locate the truth.

Deeper investigation of low (wearable) and high fidelity data A full investigation across modalities and sensors was not investigated and is beyond the scope of this current paper. We encourage future users of COSI-Lab to exploit and investigate the benefits of this parallel dataset; enabling training with high fidelity while having access only to low fidelity at test time.

Annotator Population Distribution Our Prolific annotator pool may restrict the perspectives captured in AII. However, PINE compares models conditioned on the same personas against the human distribution, remaining valid for evaluating alignment with the represented population.

8 Conclusion

The presented COSI-Lab dataset multisensor multimodal dataset of a real scientific workshop consisting of 32 participants of varying seniority and acquaintanceships. It enables, for the first time, the study of relationships between self-reported long term goals and how they relate to apparent proximal intentions. We gathered intent narrative annotations for AII using a perspectivist approach. We provide 3 benchmarks tasks on intention estimation, social involvement identification, and topic changes. By working as a multidisciplinary team, we have produced a rich interdisciplinary resource to study unscripted human social dynamics and their relation to social intentions.

AI DISCLOSURE

In this work, we used generative AI tools to implement methods, assist with translation, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, and interpret results. We did not use generative AI tools to generate synthetic datasets, clean or reformat datasets, or propose or refine hypotheses. The development of theoretical models or conceptual frameworks, formulation of mathematical claims, provision of critical ingredients for proving mathematical claims, and assistance in writing proofs were not applicable to this work.

We reviewed all AI-assisted work. We checked LLM-generated methodological suggestions against the relevant literature to identify potential overlap with existing work. We manually reviewed LLM-assisted analyses to verify their accuracy and consistency with the underlying data and research materials. All AI-generated code was reviewed, tested, and verified for correctness by at least one author. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The data collection study received ethics approval from (institutional ethics review board, obfuscated for double-blind reasons). Participants were informed about the study procedures, recorded modalities, and data sharing conditions and have provided informed consent. Annotators were recruited through Prolific and received £12 per task (each task is approximately 1 hour).

To reduce identifiability, released video is restricted to overhead views. Detected personally identifiable information (PII) is pseudonymized (see Appendix C) and corresponding audio is filtered using a 1.25kHz low frequency to reduce intelligibility. Survey data were aggregated in post-processing (see Appendix B). Access to raw, processed, and sample data requires signing an End User License Agreement pending review, and is limited for research purposes only.

REPRODUCIBILITY STATEMENT

To support reproducibility, Appendices B and C document the data collection procedures, sensor calibration and synchronization, survey processing, PII redaction, and keypoint extraction. Appendix D provides the annotation instructions, procedures, and narrative analysis settings. Appendix E contains the model prompts used in Section 5. Appendix F specifies the model configurations, evaluation protocols, and computational resources for the three benchmark tasks in Section 6. We will include links to the code repository and dataset repository, together with the CROISSANT metadata upon publication. [1]

REFERENCES

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*, 2025.

Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.

Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. SALSA: A Novel Dataset for Multimodal Group Behavior Analysis. *CoRR*, abs/1506.06882, 2015. URL http://arxiv.org/abs/1506.06882.

Michael C Ashton and Kibeom Lee. The hexaco–60: A short measure of the major dimensions of personality. *Journal of personality assessment*, 91(4):340–345, 2009.

[1] Challenges in preserving double-blind anonymity prevent us from sharing at review stage because of a data sharing protocol that requires review.

Dimitris Asimopoulos, Ilias Siniosoglou, Vasileios Argyriou, Thomai Karamitsou, Eleftherios Fountoukidis, Sotirios K Goudos, Ioannis D Moscholios, Konstantinos E Psannis, and Panagiotis Sarigiannidis. Benchmarking advanced text anonymisation methods: A comparative study on novel and traditional approaches. In *2024 13th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, pp. 1–6. IEEE, 2024.

R.H. Baayen, D.J. Davidson, and D.M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2007.12.005. URL https://www.sciencedirect.com/science/article/pii/S0749596X07001398. Special Issue: Emerging Data Analysis.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48, 2015. doi: 10.18637/jss.v067.i01. URL https://www.jstatsoft.org/index.php/jss/article/view/v067i01.

Anna Belardinelli. Gaze-based intention estimation: principles, methodologies, and applications in hri, 2023.

Michael Bratman. *Intention, plans, and practical reason*. Harvard University Press, Cambridge, MA, 1987.

Joost Broekens and Willem-Paul Brinkman. Affectbutton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, 71(6):641–667, 2013.

William J Browne, Harvey Goldstein, and Jon Rasbash. Multiple membership multiple classification (mmmc) models. *Statistical Modelling*, 1(2):103–124, 2001. doi: 10.1177/1471082X0100100202. URL https://doi.org/10.1177/1471082X0100100202.

Graham Button and Neil Casey. Topic nomination and topic pursuit. *Human studies*, 8(1):3–55, 1985.

Federico Cabitza, Andrea Campagner, and Valerio Basile. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:6860–6868, 6 2023. ISSN 2374-3468. doi: 10.1609/AAAI.V37I6.25840. URL https://ojs.aaai.org/index.php/AAAI/article/view/25840.

Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 12(1):113–130, 2018.

Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. The matchnmingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 12(1):113–130, 2021.

calib.io. Calib.io calibrator. URL https://calib.io/products/calib. Accessed: 2026-04-22.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Justine Cassell, Yukiko I Nakano, Timothy W Bickmore, Candace L Sidner, and Charles Rich. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 114–123, 2001.

John P. Caughlin and Tamara D. Afifi. When is topic avoidance unsatisfying?: Examining moderators of the association between avoidance and dissatisfaction. *Human Communication Research*, 30(4):479–513, 10 2004. ISSN 0360-3989. doi: 10.1111/j.1468-2958.2004.tb00742.x. URL https://doi.org/10.1111/j.1468-2958.2004.tb00742.x.

Jonathan M Cheek and Arnold H Buss. Shyness and sociability. *Journal of personality and social psychology*, 41(2):330, 1981.

Kathryn M Connor, Jonathan RT Davidson, L Erik Churchill, Andrew Sherwood, Richard H Weisler, and Edna Foa. Psychometric properties of the social phobia inventory (spin): New self-rating scale. *The British Journal of Psychiatry*, 176(4):379–386, 2000.

Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In *BMVC*, volume 2, pp. 10–5244, 2011.

Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk KJ Heylen, Hayley Hung, Mark A Neerincx, and Khiet P Truong. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 206–212, Cambridge, UK, 2019. IEEE.

Carole Edelsky. Who's Got the Floor? *Language in Society*, 10(3):383–421, 1981. ISSN 0047-4045.

Frank Fischer, Ingo Kollar, Karsten Stegmann, and Christof Wecker. Toward a script theory of guidance in computer-supported collaborative learning. *Educ Psychol*, 2013. doi: 10.1080/00461520.2012.748005.

Peter Fonagy, Patrick Luyten, Alesia Moulton-Perkins, Ya-Wen Lee, Fiona Warren, Susan Howard, Rosanna Ghinai, Pasco Fearon, and Benedicte Lowyck. Development and validation of a self-report measure of mentalizing: The reflective functioning questionnaire. *PloS one*, 11(7):e0158678, 2016.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400. IEEE Computer Society, 2024.

Thomas Holtgraves. The linguistic realization of face management: Implications for language production and comprehension, person perception, and cross-cultural communication. *Social Psychology Quarterly*, 55(2):141–159, 1992. ISSN 01902725. URL http://www.jstor.org/stable/2786943.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020.

Tiffany Matej Hrkalovic, Bernd Dudzik, Daniel Balliet, and Hayley Hung. Parsel: A multimodal dataset for modeling decision-making processes involved in selecting partners for joint tasks. *IEEE Transactions on Affective Computing*, 2025.

Hayley Hung and Ben Kröse. Detecting F-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pp. 231–238, Alicante Spain, November 2011. ACM. ISBN 978-1-4503-0641-6. doi: 10.1145/2070481.2070525. URL https://dl.acm.org/doi/10.1145/2070481.2070525.

Paul Ingram and Michael W Morris. Do people mix at mixers? structure, homophily, and the "life of the party". *Administrative Science Quarterly*, 52(4):558–585, 2007.

Taotao Jing, Tina Chen, Renran Tian, Yaobin Chen, Joshua Domeyer, Heishiro Toyoda, Rini Sherony, and Zhengming Ding. Psi: A benchmark for human interpretation and response in traffic interactions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Adam Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, Cambridge, UK, 1990.

Vaishnavi Khindkar, Vineeth Balasubramanian, Chetan Arora, Anbumani Subramanian, and CV Jawahar. Can reasons help improve pedestrian intent estimation? a cross-modal approach. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11515–11522. IEEE, 2024.

Arie W. Kruglanski, Marina Chernikova, Maxim Babush, Michelle Dugas, and Birga M. Schumpe. Chapter three - the architecture of goal systems: Multifinality, equifinality, and counterfinality in means—end relations. volume 2 of *Advances in Motivation Science*, pp. 69–98. Elsevier, 2015. doi: 10.1016/bs.adms.2015.04.001.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. Nemo: a toolkit for building ai applications using neural modules, 2019.

Richard D Lennox and Raymond N Wolfe. Revision of the self-monitoring scale. 1984.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11963–11974, 2023.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models, May 2025. URL http://arxiv.org/abs/2501.16566. arXiv:2501.16566 [cs].

Manuel Martín-Fernández, Blanca Requero, Xiaozhou Zhou, Dilney Gonçalves, and David Santos. Refinement of the analysis-holism scale: A cross-cultural adaptation and validation of two shortened measures of analytic versus holistic thinking in spain and the united states. *Personality and Individual Differences*, 186:111322, 2022. ISSN 0191-8869. doi: https://doi.org/10.1016/j.paid.2021.111322. URL https://www.sciencedirect.com/science/article/pii/S0191886921007017.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social Genome: Grounded Social Reasoning Abilities of Multimodal Models, February 2025. URL http://arxiv.org/abs/2502.15109. arXiv:2502.15109 [cs].

Alfred R Mele. *Springs of action: Understanding intentional behavior*. Oxford University Press, New York, USA, 1992.

Microsoft. Presidio: Data Protection and De-identification SDK. https://microsoft.github.io/presidio/.

V. S. Raghu Parupudi. Magnitude matters: a superior class of similarity metrics for holistic semantic understanding, 2025. URL https://arxiv.org/abs/2509.19323.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *Advances in Neural Information Processing Systems*, 38, 2025.

Francis Quek, David McNeill, Robert Bryll, Cemil Kirbas, Hasan Arslan, Karl E McCullough, Nobuhiro Furuyama, and Rashid Ansari. Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pp. 247–254. IEEE, 2000.

Chirag Raman, Stephanie Tan, and Hayley Hung. A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 3586–3594, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413697. URL https://doi.org/10.1145/3394171.3413697.

Chirag Raman, Jose Vargas Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. Conflab: A data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23701–23715, New Orleans, Louisiana, USA, 2022. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/95f9ad2e251e9014697589037450f9bb-Paper-Datasets_and_Benchmarks.pdf.

Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *International Conference on Computer Vision (ICCV)*, pp. 6262–6271, Seoul, South Korea, 2019. IEEE.

John F. Rauthmann and Ryne A. Sherman. Chapter 13 - conceptualizing and measuring the psychological situation. In Dustin Wood, Stephen J. Read, P.D. Harms, and Andrew Slaughter (eds.), *Measuring and Modeling Persons and Situations*, pp. 427–463. Academic Press, 2021. ISBN 978-0-12-819200-9. doi: https://doi.org/10.1016/B978-0-12-819200-9.00009-0. URL https://www.sciencedirect.com/science/article/pii/B9780128192009000090.

Scott A. Reid and Sik Hung Ng. Conversation as a resource for influence: evidence for prototypical arguments and social identification processes. *European Journal of Social Psychology*, 30(1):83–100, 2000. doi: https://doi.org/10.1002/(SICI)1099-0992(200001/02)30:1<83::AID-EJSP983>3.0.CO;2-I. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-0992%28200001/02%2930%3A1%3C83%3A%3AAID-EJSP983%3E3.0.CO%3B2-I.

RME. Rme 32 ad, a. URL https://rme-audio.de/m-32-m-16-ad.html. Accessed: 2026-04-22.

RME. Rme fireface ufx iii, b. URL https://rme-audio.de/fireface-ufx-3.html. Accessed: 2026-04-22.

Rebecca B Rubin and Matthew M Martin. Development of a measure of interpersonal communication competence. *Communication Research Reports*, 11(1):33–44, 1994.

Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: a survey. *The International Journal of Robotics Research*, 39 (8):895–935, 2020. doi: 10.1177/0278364920917446. URL https://doi.org/10.1177/0278364920917446.

Roger C. Schank and Robert P. Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Psychology Press, New York, 1977. ISBN 9780203781036. doi: 10.4324/9780203781036.

Emanuel A Schegloff. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge university press, 2007.

Sennheiser. Sennheiser sk 20 bodypack transmitter xs wireless series, a. URL https://docs.cloud.sennheiser.com/en-us/xsw/xsw/manual-sk-overview.html. Accessed: 2026-04-22.

Sennheiser. Sennheiser sk 2000 bodypack transmitter, b. URL https://www.sennheiser.com/en-us/catalog/products/wireless-systems/sk-2000. Accessed: 2026-04-22.

Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PLoS ONE*, 10, 2015.

Sewio. Vista omni anchor, a. URL https://docs.sewio.net/docs/anchor-vista-omni-30147663.html. Accessed: 2026-04-22.

Sewio. Leonardo personal tag, b. URL https://docs.sewio.net/docs/tag-leonardo-personal-30146967.html. Accessed: 2026-04-22.

Mason Swofford, John Peruzzi, Nathan Tsoi, Sydney Thompson, Roberto Martín-Martín, Silvio Savarese, and Marynel Vázquez. Improving social awareness through dante: Deep affinity network for clustering conversational interactants. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23, 2020.

Stephanie Tan, David M.J. Tax, and Hayley Hung. Conversation group detection with spatio-temporal context. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, pp. 170–180, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393904. doi: 10.1145/3536221.3556611. URL https://doi.org/10.1145/3536221.3556611.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk,

Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Ryo Umagami, Liu Yue, Xuangeng Chu, Ryuto Fukushima, Tetsuya Narita, Yusuke Mukuta, Tomoyuki Takahata, Jianfei Yang, and Tatsuya Harada. Intend to move: A multimodal dataset for intention-aware human motion understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=3CVU3RRvPx.

Jose Vargas. Covfee: Continuous Video Feedback Tool. URL https://github.com/josedvq/covfee. Accessed: 2021-05-28.

Jose Vargas Quiros, Stephanie Tan, Chirag Raman, Ekin Gedik, Ioannis Pronotaris, and Hayley Hung. Spcl mingle badge. URL https://github.com/TUDelft-SPC-Lab/spcl_midge_hardware. Accessed: 2026-04-22.

Sebastiano Vascon, Eyasu Z. Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. Detecting conversational groups in images and sequences: A robust game-

theoretic approach. *Computer Vision and Image Understanding*, 143:11–24, February 2016. ISSN 1077-3142. doi: 10.1016/j.cviu.2015.09.012. URL https://www.sciencedirect.com/science/article/pii/S1077314215002076.

Balagopal Vissa. Agency in action: Entrepreneurs' networking style and initiation of economic exchange. *Organization Science*, 23(2):492–510, 2012.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2014.01.060. URL https://www.sciencedirect.com/science/article/pii/S1053811914000913.

Yusen Wu, Junwu Xiong, and Xiaotie Deng. How Social is It? A Benchmark for LLMs' Capabilities in Multi-user Multi-turn Social Agent Tasks, April 2025. URL http://arxiv.org/abs/2505.04628. arXiv:2505.04628 [cs].

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 1688–1697. ACM, October 2022. doi: 10.1145/3503161.3547906. URL http://dx.doi.org/10.1145/3503161.3547906.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. MIntrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nY9nITZQjc.

Hanlei Zhang, Zhuohang Li, Yeshuang Zhu, Hua Xu, Peiwu Wang, Haige Zhu, Jie Zhou, and Jinchao Zhang. Can Large Language Models Help Multimodal Language Analysis? MMLA: A Comprehensive Benchmark, April 2025. URL http://arxiv.org/abs/2504.16427. arXiv:2504.16427 [cs].

Lu Zhang and Hayley Hung. Beyond F-Formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1086–1095, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.123. URL http://ieeexplore.ieee.org/document/7780492/.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents, March 2024. URL http://arxiv.org/abs/2310.11667. arXiv:2310.11667 [cs].

COSI-Lab

Appendices

The Appendices sections are inspired by the Datasheet for Datasets (Gebru et al., 2021) and are organized as follows:

A Dataset documentation

Motivation and Intended Use COSI-Lab is a multimodal multisensor dataset focusing on human social intentions in networking settings. The motivation of collecting this dataset is to create a resource for interdisciplinary researchers to investigate human social intentions, following annotation designs that emphasize interpretations from multiple annotators who explain perceived intentions. The dataset contains 32 participants. The associated time-code synchronized modalities include 10 overhead cameras and 4 elevated side-view cameras, high-fidelity personalized audio at 48kHz, a combination of 9DOF inertial measurement unit data, audio data at 20kHz, and Bluetooth proximity data in both a single-worn wearable badge and a sensorized cup, and lastly a Ultra wide band localization sensor data.

The intended use of this dataset is to support the study and modeling of human social intentions in real world, multi-party interactions. It is suited for developing and evaluating methods that integrate multimodal signals, such as vision, audio, and motion to infer underlying intentions. The multi-annotation explanations of perceived intentions provide a basis for future approaches that could account for explainable intention inference and approaches for socially assistive technologies that are capable of perspective taking. More broadly, this dataset also supports interdisciplinary research, particularly with the social science domain, to explore how intentions relate to longer term goal setting and their realization over time.

Composition The dataset is divided into several components:

- Raw data:
  - raw video, audio and wearable sensors data
- Processed data:
  - processed video and wearable sensors data used for annotations
  - annotation for intentions and conversation groups
- Data Samples:
  - Data samples of raw video, audio, and illustration of survey questions;
  - Demonstration of the post-processing, such as visualizations of keypoints and conversation group detection.
- EULA (public):
  - The End User License Agreement. One needs to sign the EULA before having access to the raw, processed, or sample data.

Figure6 illustrates the organization of the COSI-Lab dataset. For privacy reasons, data will only be shared after signing the EULA.

B Details about data collection setup

B.1 Details on workshop procedure design

In the pre-event phase (two weeks prior to the event), participants completed questionnaires assessing demographic information, self-monitoring personality (Lennox & Wolfe, 1984), social anxiety (Connor et al., 2000), communication and interaction styles (Vissa, 2012; Rubin & Martin, 1994; Cheek & Buss, 1981), and HEXACO personality traits (Ashton & Lee, 2009). These measures provide a characterization of relatively stable individual differences that are known to influence how intentions could be formed and acted upon in social contexts.

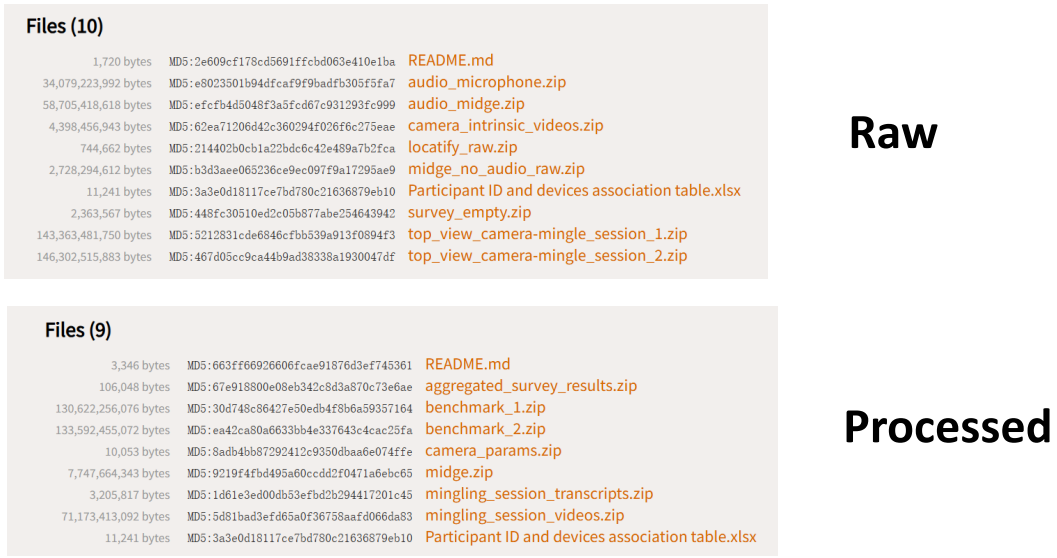


Figure 6: Dataset file structure. For more information, refer to the corresponding `README.md` files.

The in-event phase consisted of sequential components starting with arrival followed by a brief workshop introduction, then mingling session 1 (30 minutes), conversation launchpad discussion (15 minutes), mingling session 2 (30 minutes), and lastly a brief conclusion. Upon arrival, participants reported up to three workshop-level goals, ranked by importance, and indicated their affective state using the AffectButton (Broekens & Brinkman, 2013). This initial elicitation provides an anchor for individual subsequent behavior in explicitly stated goals. Participants were then equipped with sensing devices and received a briefing.

The workshop proceeded with two mingling sessions with an interleaved conversation launchpad discussion organized around thematic research topics. By design, the first mingling session provided opportunities to observe behaviors related to establishing acquaintances and warm-up social interactions, while the second mingling session provided opportunities for more targeted interactions that could lead to new connections. Prior to and after each mingling session, as well as after the conclusion of the workshop, participants specified goals and reported their current affect. This setup captured the dynamic and context-dependent nature of individual goal-setting across interaction episodes, along with the perceived outcomes which could serve as a proxy signal for individual goal realization.

In the post-event phase, participants received a follow-up email with additional details and debriefing about the sensing setup and the broader goals of the dataset. They were also asked to complete a questionnaire assessing whether they had maintained contact with, or intended to reconnect with, other participants. This measure provides an indication on long-term networking outcomes as a result of the workshop. In addition, asking participants about their level of acquaintance with others at the event allowed us to map pre-existing networks to patterns of interaction at the event.

B.2 Sensor details

Sensor settings. The GoPro cameras were set to record at 1080p, 59.94 FPS, wide field of view, and 48 kHz audio. In the first mingle session, 5 ceiling mounted cameras provided overhead views and 4 corner cameras provided elevated side views. In the second session, 5 other ceiling cameras were used again along with the same 4 corner cameras. The midges were set to record audio at 20 kHz. The badges were remotely started and stopped through a custom software which also controls for clock synchronization with the connected laptop. The Senheisser microphones were a combination of two modelds: half of which were SK 20 (Sennheiser, a) and the other half were SK 2000 (Sennheiser, b). Signals were digitized with an RME M-32 AD (RME, a) and recorded via an RME Fireface UFX III RME (b) audio interface.

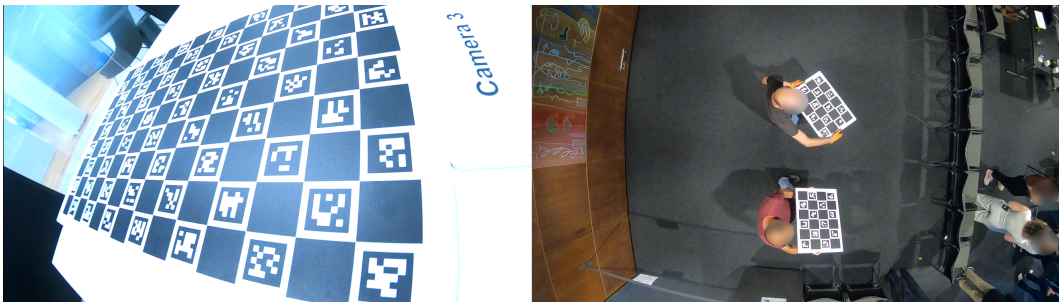


Figure 7: Example calibration footage for camera geometry estimation. The left panel shows a per-camera ChArUco-board recording for intrinsic calibration, and the right panel shows multi-camera footage used to estimate extrinsic camera poses after installation.

Time synchronization. Time synchronization is crucial for multi-modal datasets. To align sensors, camera, and audio data, we employ the protocol with a single master source of time code. The time code is fetched from online NTP server [2] and broadcast to each sensor by :pulse.

Midges support clock synchronization through Bluetooth connection, allowing for privacy-preserving data collection with millisecond level time sync alignment. The synchronization was done every 5 minutes. However, due to the amount of devices present, not every sync message was received every time by each midge. We observed that around a quarter of them would be reachable on each time sync round. This is still within the parameters defined in (Raman et al., 2020).

For cameras, each camera was mounted with a Syncbac PRO to embed the time code signal into the video. For audio, the michrophones record the time code signal from the :pulse in an additional audio channel. The 33 channels (32 microphones plus the time code) are recorded on an HDD connected to the Fireface. UWB sensors were synchronized with the other modalities by connecting the laptop running the localization software to the same network time protocol as the Midge laptop.

Camera calibration. We perform camera calibration to obtain camera intrinsic and extrinsic matrices, which are needed for downstream 3D processing tasks. We recorded a calibration section before each mingling session using several calibration boards. In previous work (Raman et al., 2022) the camera calibration was done by placing marks in the ground at regular intervals. This is a time consuming process, both placing the marks and manually annotating them in the recorded images. Moreover, this leads to poor camera matrix estimations, as this is equivalent to performing a camera calibration based on a single static board placed on the ground.

Initially, we attempted full bundle adjustment for the calibration. However, we found that the calibration had difficulty converging to a good optima due to the distance between the boards and the cameras. Therefore, we recorded additional videos for intrinsic camera calibration, which were used to initialize the extrinsics optimization.We employed calib.io's calibrator software (calib.io) to run to detect the board features and run the optimization. The advantage over open source alternatives is that it supports using several calibration boards at once. We used 3 ChArUco boards. Using 3 boards decreases the total calibration recording time, compared to only using 1 board. This is required, as the cameras are relatively far away from the mingle area, which would lead to very long and tedious calibration recording sessions.

The technical details for each board as specified in Table 4, where each board has a unique starting ID for the ArUco markers: 0, 12, and 24, respectively. With this approach we obtained camera parameters that lead to an average reconstruction error of 0.8843 pixels for mingle session 1 and 0.7644 pixels for mingle session 2.

B.3 Survey details

Survey contents. The questionnaires are designed based on several values. We present below the brief summary of the components in each survey. For complete set of questions, refer to the dataset link:

[2] https://tf.nist.gov/tf-cgi/servers.cgi

- **The pre-workshop survey** is given 2 weeks prior to the workshop. Content: welcome information, self-monitoring scale, social anxiety scale, networking style, shyness and sociability, communication style, HEXACO scale, and Demographic information.
- **Workshop survey** is progressively collected along with the workshop. Contents: we ask for goals before each mingling session, and check whether their goals were achieved in the mingling session. Also, we check for overall goals and how they achieve them before and (immediately) after the the workshop.
- **Post Workshop Survey (Immediate)** is also given immediately after the workshop. We separate this from Workshop survey for privacy reasons only. It asks each participant whether, how, and why they want to keep communication with each other participant. We provide participant photos to help identification.
- **Post Workshop Survey (3 months)** is given 3 months later from the workshop. Contents: whether people keep communications with those they specified directly after the workshop, and for what reason.

Survey processing. The raw survey responses were processed into participant-level variables before release. We used a manually curated codebook to map columns in the raw data to their corresponding measure, response scale, and reverse-coding indicator. For each multi-item measure, item-level responses were converted to numeric scores, reverse-coded when specified, and averaged across items to produce one measure-level score per participant. Attention checks were handled separately and summarized as both the number and proportion of correct responses.

We also applied privacy-preserving transformations to demographic fields. Age was quantized into bins, and demographic attributes including country of birth, country of residence, nationality, and ethnicity were recoded into broader categories to reduce identifiability. In addition, raw survey metadata, including timestamps, IP address, geolocation, and response identifiers, were removed from the processed dataset.

B.4 Survey matching across waves

Survey responses were matched across waves using a respondent-level matching ID constructed from the combination of several self-generated ID columns. For the pre-workshop survey, this ID was constructed from `id1`, `id2`, `id3`, and `id4`. For the workshop, post-workshop, and post-event surveys, it was constructed from `birthday_2`, `street_2`, `em_contact_2`, and `place_of_birth_2`.

This matching ID was used only to link survey responses from the same respondent across waves. It is separate from the A01–A32 participant labels, which were used to link participants to the event roster, microphones, and annotations.

Table 2 summarizes the availability of respondent-level matching IDs after cleaning the survey response rows. The pre-workshop survey contained 34 unique matching IDs, exceeding the 32-person event roster, while the workshop and post-event surveys contained a small number of rows with missing matching IDs.

Table 2: Availability of respondent-level matching IDs across survey waves. Cleaned response rows do not include column headers. Duplicate rows refer to extra rows sharing an already observed matching ID within the same survey.

Survey wave	Cleaned rows	Missing ID rows	Duplicate ID rows	Unique IDs
Pre-workshop	34	0	0	34
Workshop	33	2	0	31
Post-workshop	32	0	0	32
Post-event	28	1	2	25

Table 3 reports the exact cross-wave matching coverage based on this constructed ID. Fourteen respondents were matched across all four survey waves. Therefore, analyses in the main text use the largest appropriate denominator for each survey-based result, rather than restricting all results to the four-wave complete-case subset.

Table 3: Exact cross-wave matching coverage based on the constructed respondent-level matching ID.

Survey comparison	Matched unique IDs
Pre-workshop ∩ workshop	23
Pre-workshop ∩ post-workshop	22
Pre-workshop ∩ post-event	16
Workshop ∩ post-workshop	29
Workshop ∩ post-event	19
Post-workshop ∩ post-event	18
All four surveys	14

Width	600 mm
Height	400 mm
Rows	4
Columns	6
Checker width	90 mm
Dictionary	DICT_5X5_100
OpenCV legacy pattern	True

Table 4: Charuco boards details

We used exact matching on the constructed respondent-level ID. Possible one-field mismatches were inspected but not automatically merged.

C Data sanitization and processing

C.1 Audio data pseudonymization

The audio privacy processing pipeline consisted of two related branches: transcript pseudonymization and audio redaction. Both branches relied on word-level transcription, which allowed detected personally identifiable information (PII) in text to be linked to the corresponding time intervals in the audio. The goal was to reduce the exposure of identifiable linguistic content while preserving the temporal structure of the recordings for interaction analysis.

For transcript pseudonymization, we used the Sennheiser microphone recordings from the mingling sessions as the input source. To obtain high-quality transcripts, the recordings were first processed with NeMo-based speaker segmentation (Kuchaiev et al., 2019), after which WhisperX (Bain et al., 2023) was used to generate word-level transcriptions. PII detection was then performed on the transcripts using Microsoft Presidio (Microsoft) with spaCy (Honnibal et al., 2020) and regular-expression based recognizers. The detected PII categories included person names, email addresses, and phone numbers. In the released transcripts, person names were replaced with consistent pseudonyms, while email addresses and phone numbers were replaced with typed placeholders.

For audio redaction, the Sennheiser and midge audio recordings were processed directly without speaker separation. This choice was made because the Sennheiser recordings were already sufficiently clear, and because applying speaker separation before redaction could introduce fragmentation and unnecessary processing artifacts. Keeping the redaction pipeline closer to the original recordings also allows other researchers to apply their own post-processing methods according to their research needs. Each recording was transcribed with WhisperX, and the same PII detection procedure was applied to the resulting word-level transcript. The detected PII spans were mapped back to their corresponding time intervals in the audio, and these intervals were processed with a low-pass transformation at 1250 Hz. This transformation reduces the intelligibility of sensitive spoken content while preserving the surrounding temporal structure of the audio.

C.2 3D keypoint extraction

Obtaining 3D keypoint is necessary for inferring people's positions and orientations in the space, and subsequently for downstream human conversation group analysis. In COSILab we do this in several

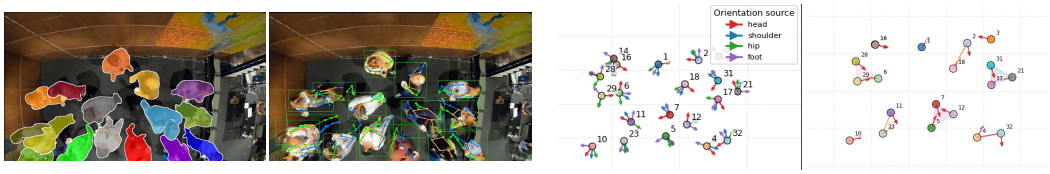


Figure 8: Illustration of the process of benchmarking task 2. (a)Segmentation mask from SAM3 model; (b)Inferred keypoints from VitPose; (c)Transferred 3D keypoints in space; (d)Illustration of the conversation group detection task.

steps: obtaining person masks, inferring 2D keypoints, transferring to 3D keypoints, and inferring positions and orientations.

C.2.1 Obtaining person masks

We construct a per-person per-frame object mask by the SAM3 model (Carion et al., 2025). SAM3 provides precise object masks with flexible inputs, such as providing one point or bounding boxes. In our case, we first manually annotate the first frames of videos with human bounding boxes, and then keep tracking and segmenting the people in the video. In practice, we clip the videos into 20-second segments, annotate for first frames of each segment, and then concatenate the masks together. This has two advantages:

- It updates people composition in the scene. Since automatic detecting people would cost extra memory on GPU, we employ a tracking-only module of the SAM3 model. This means that it can only track and not detect any new person appearing in a fixed scene. By splitting a lon gvideo into short clips, we *update* visible people by annotation every 20 seconds, and therefore facilitating people being tracked.
- It reduces memory cost significantly. SAM3 model consumes GPU memory rapidly as the length of the video grows, especially for our case where the number of people being segmented is high. In our experiment, continuously tracking 15 people in a video of over 1 minute with 1920*1080 resolution would consume more than 80G GPU memory, so cutting the video into short segments would significantly relax this constraint.

An example of one frame of segmented people in the scene is shown in figure 8(a). The segmentation is done for both mingling sessions, each session with 3 camera videos of 35 minutes. The first session of masked video is 13:45:00 to 14:20:00 (camera 6,8,10), and the second session is 14:52:00 to 15:27:00 (camera 1,3,5). This step gives us per-frame per-person masks.

The mask extraction is done on Intel Xeon Gold 6448Y 32C 2.1GHz CPU and NVIDIA Tesla A100 GPUs (80GB memory). Under this setting, it takes around 20 minutes to generate per-frame person masks for each 20 second clip, but it varies among the number of person present in the video.

C.2.2 Inferring 2D keypoints

Each person mask is essentially a set point. We take it's rectangular convex hull on the image as the bounding box for corresponding person, and getting per-frame per-person bounding boxes. To infer keypoints in the 2D pixel domain, we leverage ViTPose model (Xu et al., 2022). It can infer people's skeleton and provide keypoint positions in a continuous manner on a video, given the continuous bounding box information. Therefore, we use per-frame bounding boxes as input and get per-frame per-person keypoint information. An example keypoint plot is shown in figure 8(b).

The ViTPose model in this process is fine-tuned on ConfLab (Raman et al., 2022), where videos from overhead cameras are accompanied with manually annotated keypoints for each person. Since the original ViTPose model was trained for detecting keypoints on the face/back side, fine-tuning on a similar setting (top-down camera) makes it perform much better on our videos.

The 2D keypoint extraction is done on AMD EPYC 7543 32-Core Processor CPU and NVIDIA A40 GPU (46GB memory). On average, one task of inferring 2D keypoints for a 5-minute video takes around 40 minutes, but also subject to number of people.

C.2.3 TRANSFERRING TO 3D KEYPOINTS

The keypoint information obtained in C.2.2 is in 2D pixel space. However, most applications regarding conversation group analysis (Swofford et al., 2020; Tan et al., 2022) require *spatial* position and orientation information, i.e. within 3D world coordinates. To make the transfer, we utilize camera intrinsics and extrinsics.

Given the camera intrinsic matrix

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix},$$

and distortion coefficients D, the pixel coordinate is undistorted and converted into normalized camera coordinates:

$$(x_i^n, y_i^n) = \mathrm{Undistort}(\tilde{p}_i, K, D).$$

This gives a ray in the camera coordinate system:

$$\mathbf{r}_i^c = \begin{bmatrix} x_i^n \\ y_i^n \\ 1 \end{bmatrix}.$$

The extrinsic parameters are represented by a rotation vector and translation vector. The rotation vector is converted to a rotation matrix R using Rodrigues' formula. The extrinsic relation follows the OpenCV convention:

$$\mathbf{X}^c = R\mathbf{X}^w + \mathbf{t},$$

where $\mathbf{X}^w$ is a 3D point in world coordinates and $\mathbf{X}^c$ is the same point in camera coordinates.

The camera center in world coordinates is therefore $\mathbf{C}^w = -R^\top \mathbf{t}$, and the camera ray is transformed into the world coordinate system by $\mathbf{r}_i^w = R^\top \mathbf{r}_i^c$.

Each keypoint is assumed to lie at a known height above the floor. For body height H and keypoint-specific height ratio α_i, the world height of keypoint i is $Z_i^w = H\alpha_i$. The 3D point is then obtained by intersecting the world ray with the horizontal plane $Z = Z_i^w$.

Points on the ray are $\mathbf{X}_i^w(\lambda) = \mathbf{C}^w + \lambda \mathbf{r}_i^w$. Solving for λ using the z-coordinate gives

$$\lambda_i = \frac{Z_i^w - C_z^w}{r_{i,z}^w}.$$

Thus, the world position of keypoint i is

$$\mathbf{X}_i^w = \mathbf{C}^w + \frac{Z_i^w - C_z^w}{r_{i,z}^w}\mathbf{r}_i^w.$$

and only the floor-plane coordinates are retained:

$$(x_i^w, y_i^w).$$

An example of the transferred keypoints in 3D is given in figure 8(c). This task does not require heavy computation resources and any commercial device would suffice.

C.2.4 UWB TO CAMERA ALIGNMENT AND VALIDATION

To align UWB positions with the video recordings, we estimated a spatial transformation from the UWB coordinate system to the camera coordinate system using manually annotated point correspondences. Each correspondence paired a selected UWB location with its corresponding position in the camera image. UWB reference positions were represented in homogeneous coordinates and transformed using a 4×4 matrix $\mathbf{A}$. The transformed positions were projected onto the image plane using the camera's calibrated intrinsic parameters and fisheye distortion model, implemented through OpenCV's fisheye projection function.

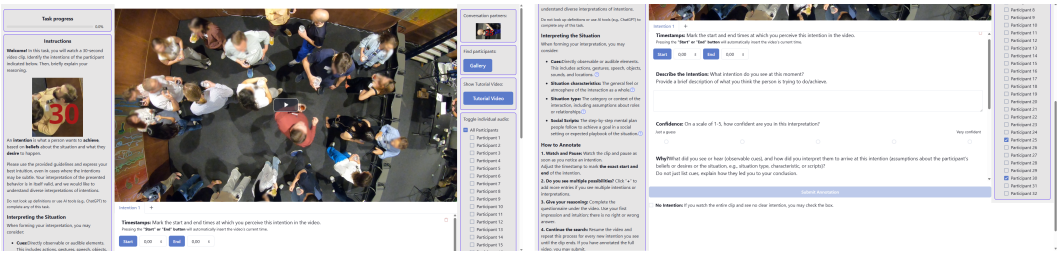


Figure 9: Displayed a screen grab of the UI of Covfee.

We manually annotated 16 pocket locations across participants to establish correspondences between UWB positions and image coordinates. The transformation ($\mathbf{A}$) was estimated through iterative reprojection error minimization. At each iteration, the current estimate of $\mathbf{A}$ was used to transform and project the UWB reference positions into the image plane. The reprojection error for each correspondence was defined as the Euclidean distance in pixels between the projected position and the annotated pocket location. We updated $\mathbf{A}$ using Levenberg-Marquardt optimization to minimize the sum of squared reprojection errors across all correspondences. The estimated mean physical position error was approximately 50 cm, compared with the manufacturer's stated positioning accuracy of up to 30 cm. Errors of this magnitude may obscure differences in interpersonal distance in crowded social settings, where participants may stand within arm's reach of one another.

D Intention Annotations

Annotator pool The annotators were sourced from Prolific [3], a crowd sourcing platform. Annotators were pre-screened based on good performance on past cue-labeling tasks. A total of 83 annotators (over both pilots and full annotation) participated. The research team designed the annotation protocols, instructions walkthrough video, annotation interface, and rubrics. Across the annotator pool, 41% identified as female and 59% as male. Ages ranged from 21 to 66 years (M = 41.42, SD = 10.00). When asked about ethnicity, 75% reported White/European, 12% as Asian, and 13% as Other. In terms of their career stage, 22% reported early career, 56% mid career, and 22% late career. Note that Prolific keeps a balanced set of workers across the dimension of age, sex, ethnicity and political affiliation. So any variation corresponds to noise in the worker sampling process.

Annotation Platform Annotations were collected online using a custom interface adapted from Covfee (Vargas), shown in Figure 9. Annotators first watch a tutorial video explaining the interface and demonstrating an example annotation using a video from a different dataset to reduce priming bias. They are then presented with the annotation instructions alongside the target participant they must annotate. The right sidebar displays the IDs and images of the other members of the conversation group, while a gallery view allows annotators to map participant IDs to all individuals present in the scene. The video is presented together with the audio tracks of the target participant's conversation group, although annotators may individually play or mute each participant's audio using controls on the right. Below the video is the annotation questionnaire, which must be fully completed unless the annotator selects the "no intention found" option. Annotators may also label multiple intention events within the same clip by adding additional questionnaire entries using the "+" button. The annotator answers are recorded in `.json` files, which are then used for benchmark task 1.

Annotation procedure Annotators first fill in the annotator survey, similar to the participant pre-workshop survey, with additional Analysis-Holism Scale(Martín-Fernández et al., 2022) and Reflective Functioning Questionnaire (Fonagy et al., 2016) questionnaires. Once filled, they could move on to the annotation task where they were provided with a video/audio pair and a participant to annotate. A different instruction set was used for the full study, Pilot condition A and Pilot condition B as shown below:

[3] https://www.prolific.com/

Annotation Instructions

Tutorial video

The tutorial video Annotation_tutorial_instructions.mp4 is provided together with the data. [Tutorial video]

Welcome!
In this task, you will watch a 30-second video clip. Identify the intentions of the participant indicated below. Then, briefly explain your reasoning.

Picture of participant and ID

An intention is what a person wants to achieve, based on beliefs about the situation and what they desire to happen.

Please use the provided guidelines and express your best intuition, even in cases where the intentions may be subtle. Your interpretation of the presented behavior is in itself valid, and we would like to understand diverse interpretations of intentions.

Do not look up definitions or use AI tools (e.g., ChatGPT) to complete any of this task.

Interpreting the Situation

When forming your interpretation, you may consider:

Cues: Directly observable or audible elements.
This includes actions, gestures, speech, objects, sounds, and locations.
Ask yourself: What can I directly see or hear?

Situation characteristics: The general feel or atmosphere of the interaction as a whole
Ask yourself: How does the situation feel overall?
Examples: tense, relaxed, awkward, friendly, formal, hostile.

Situation type: The category or context of the interaction, including assumptions about roles or relationships.
Ask yourself: What kind of situation is this?
Examples, job interview, casual conversation, negotiation, strangers meeting, boss-employee interaction.

Social scripts: The step-by-step mental plan people follow to achieve a goal in a social setting or expected playbook of the situation
Ask yourself: What is the expected sequence of events here? Who usually does what?
Examples: greeting -> introduction -> conversation, ordering -> paying -> leaving, turn-taking in conversation

How to Annotate

1. Watch and pause
Watch the clip and pause as soon as you notice an intention.
Adjust the timestamp to mark the exact start and end of the intention.

2. Do you see multiple possibilities?
Click "+"to add more entries if you see multiple intentions or interpretations.

3. Give your reasoning
Complete the questionnaire under the video. Use your first impression and intuition; there is no right or wrong answer.

4. Continue the search
Resume the video and repeat this process for every new intention you see until the clip ends. If you have annotated the full video, you may submit.

Grading rubric:

Your response will be evaluated based on the following criteria:

Intention (not just actions): Describe what the participant is trying to achieve, not just what they are doing. Actions alone (cues) are not enough you must also include a goal.

Assumptions about the participant: Include what you assume about the participant's beliefs or desires. What do they think is happening? What do they want?

Assumptions about the situation: Use your interpretation of the situation to support your answer, based on:

- Situation type (what kind of situation this is)
- Social scripts (what typically happens in this situation)
- Situation characteristics (overall tone)

Dont spend too much time per question. your initial response is valid.

Questionaire

Timestamps: Mark the start and end times at which you perceive this intention in the video.
[Timestamp start] [Timestamp end]

Describe the Intention: What intention do you see at this moment? Provide a brief description of what you think the person is trying to do/achieve.
[Free text]

Confidence: On a scale of 1-5, how confident are you in this interpretation?
[Likert 1-5]

Why? What did you see or hear (observable cues), and how did you interpret them to arrive at this intention (assumptions about the participant's beliefs or desires or the situation, e.g., situation type, characteristic, or scripts)? Do not just list cues, explain how they led you to your conclusion.
[Free text]

Confidence: On a scale of 1-5, how confident are you in this explanation?
[Likert 1-5]

Intensity: On a scale of 1-5, how much of a priority does this intention appear to be for the participant?
[Likert 1-5]

Counterfactual Explanation: Can you think of an alternative way the situation could be interpreted that would lead to a very different understanding of the participant's intentions? Describe the assumption or interpretation that would change your understanding.
[Free text]

No Intention: If you watch the entire clip and see no clear intention, you may check the box.
[Checkbox]

Pilot Annotation Instructions

Form A

Welcome!
In this task, you will watch a short video clip (30 seconds) of a social interaction. Your goal is to identify the intentions of the participant (shown below) as they occur.

Picture of participant and ID

Please rely on your own intuition and understanding of what an intention is there are no right or wrong answers. We are interested in your immediate impression.

Do not look up definitions or use AI tools (e.g., ChatGPT) while completing this task.

How to Annotate

1. Watch and pause Watch the clip and pause as soon as you notice an intention.
Adjust the timestamp to mark the exact start and end of the intention.

2. Do you see multiple possibilities?
Click "+" to add more entries if you see multiple intentions or interpretations.

3. Give your reasoning Complete the questionnaire under the video. Use your first impression and intuition; there is no right or wrong answer.

Intention (Free text):

Confidence: On a scale of 1-5, how confident are you in this interpretation?

Just a guess / Extremely confident
[Likert 1-5]

Why? Provide the evidence from the video or audio that led you to this conclusion.
Description (Free text):

Confidence: On a scale of 1-5, how confident are you in this interpretation?
Just a guess / Extremely confident
[Likert 1-5]

Rate Your Perception:
Intensity: On a scale of 1-5, how much of a priority does this intention appear to be for the participant?
Not pursuing this intention at all / clearly the participant's main focus [Likert 1-5]

Counterfactual: Can you think of an alternative way the situation could be interpreted that would lead to a very different understanding of the participant's intentions?

No Intention: If you watch the entire clip and see no clear intention, you may check the "No intention seen" box.

Form B

Welcome!
In this task, you will watch a 30-second video clip. Identify the intentions of the participant indicated below. Then, briefly explain your reasoning.

Picture of participant and ID
An intention is what a person wants to achieve, based on beliefs about the situation and what they desire to happen.

Use your first impression and own intuition; there are no right or wrong answers.

Do not look up definitions or use AI tools (e.g., ChatGPT) to complete any of this task.

Interpreting the Situation

When forming your interpretation, you may consider:

Cues: Directly observable or audible elements e.g. actions, objects, sounds, or locations, etc.

Situation characteristics: The overall feel or tone of the situation (e.g., tense, casual, pleasant).

Situation type: What kind of situation the participant might think they are in.

Social scripts: Typical patterns sequence of steps or mental "how-to" guides for how phases of social interactions usually unfold in this setting.

How to Annotate

1. Watch and pause Watch the clip and pause as soon as you notice an intention.
Adjust the timestamp to mark the exact start and end of the intention.

2. Do you see multiple possibilities?
Click "+" to add more entries if you see multiple intentions or interpretations.

3. Give your reasoning Complete the questionnaire under the video. Use your first impression and intuition; there is no right or wrong answer.

Intention (Free text):

Confidence: On a scale of 1-5, how confident are you in this interpretation?
Just a guess / Extremely confident
[Likert 1-5]

Why? Provide the evidence from the video or audio that led you to this conclusion.
Description (Free text):

Confidence: On a scale of 1-5, how confident are you in this interpretation?
Just a guess / Extremely confident
[Likert 1-5]

Rate Your Perception:
Intensity: On a scale of 1-5, how much of a priority does this intention appear to be for the participant?

> Not pursuing this intention at all / clearly the participant's main focus [Likert 1-5]
>
> Counterfactual: Can you think of an alternative way the situation could be interpreted that would lead to a very different understanding of the participant's intentions? Describe the assumption or interpretation that would change your understanding.
>
> No Intention: If you watch the entire clip and see no clear intention, you may check the "No intention seen" box.

Instructions of Pilot condition A were made to be less descriptive, leaving the annotator more freedom to write their first intuitive intention narrative. Instructions for Pilot condition B added a definition of intention as well as some descriptions of cues, characteristics, class and scripts. The resulting annotations showed a very similar result for both pilot conditions. This lead us to introduce the final intruction set for the full annotation study. In addition to the new instructions, the annotators in the final study also received the tutorial video (the video was not present in the pilot). 48 annotators were tasked with annotating 3 videos in the pilot, and 71 annotators annotated a minimum of 6 video each in the full study. Once finished with their task, annotators were then redirected to a small survey where they could provide feedback about their experience with the annotation.

Measures . The semantic similarity was conducted by embedding the annotations using 'BAAI/bge-large-en-v1.5' [4] then using the Hyperbolic Tangent Similarity (HTS) (Parupudi, 2025) to calculate distances. The graph was then generated by conducting a UMAP on two principal components. The intention description graph used hyperparameters n_neighbours = 500 and min_dist = 0.1, the intention explanation graph used n_neighbours = 200 and min_dist = 0.25. All code for processing and generating these graphs can be found on our repository. Similarly, the code used to parse the emotion, cue, belief and intentions can also be found there, together with the LLM-as-a-judge experiment.

D.1 Validity of the Persona Alignment Score

PAS is meant to indicate whether a generated narrative is a plausible interpretation of *this* clip. Because it keeps the best match among the clip's annotations, part of any score reflects how close a narrative is to human phrasing in general, independent of the clip. We therefore read every PAS against a *random-clip baseline*: the same narratives scored against the annotations of a randomly drawn other clip (mean of 20 draws). Unless stated otherwise, we use all-MiniLM-L6-v2 embeddings, HTS, and *matched pools*: the annotator whose persona conditioned a generation is removed from that generation's pool, just as a human annotation is scored only against the *other* annotators of its clip (removing all of that annotator's annotations of the clip). Scores are averaged per clip and then across clips, so clips with many generations do not dominate. This differs slightly from the per-generation average in Section 6. Confidence intervals come from resampling clips (1,000 bootstrap draws).

Human reference and normalized PAS. Held-out human annotations reach a PAS of 0.457 (descriptions) and 0.464 (explanations), against a random-clip baseline of 0.397 for both. We define the *normalized PAS* as $(\text{PAS}_{\text{model}} - \text{base}_{\text{model}})/(\text{PAS}_{\text{human}} - \text{base}_{\text{human}})$, where 0 means no better than a random clip and 1 means as clip-grounded as a held-out human. Gemma reaches 0.86 $[0.75, 0.99]$ for descriptions and 0.81 $[0.71, 0.91]$ for explanations, and qwen3omni30b reaches 0.73 $[0.64, 0.83]$ and 0.75 $[0.67, 0.84]$. The baselines differ by model: qwen3omni30b's narratives are closer to human phrasing regardless of the clip (baseline 0.358 vs. 0.316 for descriptions), which explains its higher raw PAS. Once the baseline is subtracted, gemma is ahead on descriptions, while on explanations the two models are not reliably different (see Robustness); the intervals overlap in both fields. Restricting to the 324 clips that kept all six annotators raises every normalized score (0.80–0.91; qwen3omni30b's descriptions rise most, to 0.84), and gemma remains ahead on descriptions.

Known groups. We scored inputs whose correct ordering is known in advance (Table 5). Each group is compared with its *own* items scored against a random other clip, so differences in writing style between groups do not affect the comparison. The groups are: 150 LLM paraphrases of human annotations; *atypical* human annotations (per clip, the annotation least similar on average to the others); 150 *blind* outputs, produced from the same persona prompt with the model told the clip

[4] https://huggingface.co/BAAI/bge-large-en-v1.5

is unavailable; and 12 hand-written generic intentions that fit any conversation. PAS behaves as expected on the checks that matter most. Paraphrases keep the ranking of their sources (Spearman $\rho = 0.74$ and 0.85), with a small mean drop (-0.034 and -0.017). Blind outputs are no better than a random clip (AUC 0.49 and 0.46). Two results show that the raw score cannot be read on its own. First, for the same (persona, clip) pairs, gemma's sighted descriptions score *lower* than the blind ones (mean difference -0.052; sighted higher in only 35% of 144 pairs). The explanations go the expected way ($+0.043$; 66% of 150 pairs). Second, generic templates score about as high as qwen3omni30b (0.397 vs. 0.402), yet neither blind outputs nor templates contain any information about the clip. Atypical human annotations are only weakly separated from their random-clip scores (AUC 0.54 and 0.56). Taking the max over the pool does not rescue them more than averaging over it does (AUC 0.55 and 0.56 with the mean).

Table 5: PAS for known groups (HTS, max, matched pools; generic templates are scored against every clip). *Wrong clip*: the same items scored against a random other clip, one draw per item. It therefore differs slightly from the 20-draw baseline used in the text. AUC: probability that an item's score on its own clip exceeds a score on a wrong clip (0.5 = no clip information). Clip-weighted means.

	Description			Explanation		
Group	PAS	Wrong clip	AUC	PAS	Wrong clip	AUC
Held-out human	0.457	0.395	0.68	0.464	0.398	0.72
Paraphrased human	0.440	0.366	0.71	0.470	0.391	0.72
Atypical human	0.334	0.318	0.54	0.349	0.323	0.56
gemma	0.368	0.317	0.63	0.402	0.347	0.67
qwen3omni30b	0.402	0.358	0.63	0.425	0.375	0.66
Blind persona LLM	0.438	0.437	0.49	0.364	0.371	0.46
Generic template	0.397	–	–	0.399	–	–

Convergent alignment with an LLM judge. In the absence of human plausibility ratings, a text-only LLM judge rated 550 annotations on *fit* (plausibility for this clip, 1–5) and *specificity* (1–5). The judge cannot watch the clip. For each item, we split the five other annotators of a six-annotator clip at random: the judge reads the notes of three of them, and PAS is computed only against the remaining two, so the two measures share no reference text. The judge passes three sanity checks. It rates held-out humans above wrong-clip model outputs on fit (3.89 vs. 1.86, AUC 0.87), it rates generic templates low on specificity (1.30), and its ratings are stable when the notes are shown in a different order (quadratic $\kappa = 0.97$ for fit, 0.96 for specificity). On the 300 model outputs, stratified over PAS quintiles, PAS correlates moderately with judged fit ($\rho = 0.26$; 0.27 controlling for length; AUC 0.63 for identifying outputs with fit ≥ 4). It correlates weakly and *negatively* with specificity ($\rho = -0.11$): among model outputs, those that match a human best tend to be slightly less tied to the particular clip. Together with the high raw scores of generic templates, this suggests that PAS gives some credit to safe, generic interpretations. When PAS is instead computed against all five annotators, including those the judge reads, the correlation rises to 0.40. This shows how much shared reference text inflates agreement. Pool reducers differ little: max (0.26), top-2 (0.27), top-3 (0.25) and centroid (0.25) agree with the judge about equally, and only the mean is lower (0.21).

Robustness. Item-level PAS is similar under BGE-large-en-v1.5 and MiniLM (Spearman $\rho = 0.76$–0.81). For unit-norm embeddings, HTS and overlap are monotone in cosine, so all three metrics produce identical rankings. PAS is essentially uncorrelated with the length of model narratives ($|\rho| \leq 0.06$). It is mildly correlated with length for human annotations ($\rho = 0.22$–0.24), but the model–human gap persists within every length quintile. PAS correlates with content-word overlap with the pool ($\rho = 0.46$–0.57), which is expected for narratives describing the same events. Across 15 configurations (2 embeddings $\times$ 3 metrics $\times$ 5 reducers), the normalized PAS ranks gemma above qwen3omni30b for descriptions in 13 configurations, but in only 7 for explanations. The model ordering on explanations should therefore be treated as a tie.

Reliability. The random-clip baseline varies little between draws: its standard deviation across 20 seeds is at most 3% of the model–baseline gap. Across 200 random split-halves of the clips, the gemma vs. qwen3omni30b ordering of normalized PAS agrees between halves 99% of the time for

descriptions and 82% for explanations. Persona-level scores are much less reliable, with Spearman–Brown corrected split-half correlations of 0.26–0.38 for descriptions and 0.54–0.64 for explanations. PAS is therefore suited to comparing models and conditions in aggregate, but not to ranking individual personas.

Summary. Read against its random-clip baseline, PAS separates clip-grounded from clip-free text. It is largely stable under paraphrase and under the choice of embedding model and similarity metric, and the model–human gap does not depend on narrative length. It agrees moderately with an independent LLM judge. In raw form, however, PAS partly reflects how closely a source writes like humans in general, and it gives some credit to generic interpretations. A model can therefore score higher without being better grounded in the clip. We recommend reporting normalized PAS alongside raw PAS, comparing models in aggregate rather than per persona, and not treating the model ordering on explanations as settled.

E Descriptive Analysis Details

E.1 Coding open-text mingling goals

We coded participants' open-text primary goals for each mingling phase into three descriptive categories: *social/networking*, *academic/targeted*, and *comfort/leisure/other*.

Social/networking was used for broad interaction goals, such as meeting people, talking, chatting, networking, connecting, getting acquainted, or breaking the ice, when no more specific academic or follow-up purpose was stated.

Academic/targeted was used when the response specified a more constrained goal related to research, data, theory, methods, similar interests, collaboration, future professional contact, or following up on a previous conversation. When a response combined general networking with a research, collaboration, or follow-up target, we coded it as academic/targeted rather than social/networking.

Comfort/leisure/other was used for goals related to comfort, fatigue management, the recording or data-collection process, leisure, fun, or unclear intent.

For ambiguous cases, we applied a specificity rule: more specific academic, professional, or follow-up goals took precedence over broad social-entry goals. For example, "meet new people and find out their research interest" was coded as academic/targeted, whereas "talk to new people" was coded as social/networking.

E.2 LLM-assisted topic-distribution coding

The LLM-assisted analyses were run offline using locally hosted models. The topic-distribution analysis used Qwen3-14B, while the topic-redirection and uptake analysis used Gemma 4 31B. The model outputs were checked before being used in the descriptive analyses and Benchmark 3. The prompts are provided below for reproducibility.

For topic-distribution coding, we divided each stable conversation segment into transcript windows. Long segments were split into 300-turn windows with a stride of 250 turns, giving a 50-turn overlap. If the final window contained fewer than 100 turns, it was merged with the previous window. Turn indices refer to the original conversation segment and were not renumbered within each window.

Open-coding topic distribution prompt: system message.

You are analyzing a transcript window from a multi-party mingling conversation.

Task:

Identify the broad emergent topic areas discussed in this transcript window.

The goal is to summarize the main topic distribution of the conversation.
This is not candidate detection.
This is not uptake judgment.

This is not a predefined taxonomy task.

Use open coding:

Create topic labels from the conversation itself.
Do not force the transcript into predefined categories.
Do not use an external taxonomy.

Topic granularity rule:

Use broad but data-grounded topic areas.

A good topic label should be:

- broad enough to group several related turns, examples, questions, or explanations;
- specific enough to reflect what this conversation was actually about;
- useful for later descriptive distribution analysis.

Prefer merging over splitting.
Do not create a separate topic for every detail, example, method, person, place, or short follow-up question.

For a 200-400 turn window, usually 2-5 broad topics are enough.
Only use more topics if the conversation clearly moves through many different areas.

Examples of good broad emergent topics:

- "academic career paths and postdoc options"
- "research background and current projects"
- "medical teamwork and simulation data"
- "methods for measuring stress and coordination"
- "leadership and communication styles in surgery"
- "conference networking and participant background"

Examples that are too broad:

- "research"
- "career"
- "medicine"
- "methods"
- "personal life"

Examples that are too fine:

- "ankle wristband placement"
- "country music in surgery"
- "one surgeon saying stop"
- "Hamburg last summer"
- "eye-tracking glasses looking down"

Transcript note:

The transcript may contain ASR errors, repetitions, overlap, fillers, and fragmentary turns.
Ignore obvious ASR artifacts unless they affect the topic.

Evidence note:

For each topic, include only 2-3 representative turn indices as evidence. Do not list every turn covered by the topic.

Output rules:

- Return exactly one valid JSON object.
- The first character must be ‘{‘.
- Do not output markdown, comments, explanations, or hidden reasoning.
- If no coherent topic can be identified, return an empty topic list.

Open-coding topic distribution prompt: user message template.

Analyze the following transcript window.

Return valid JSON with this schema:

```
{
  "topic_areas": [
    {
      "topic_id": "T1",
      "topic_label": "broad emergent topic label",
      "topic_summary": "one concise sentence about what this
topic covers",
      "evidence_turn_indices": [1, 2],
      "estimated_turn_share": 0.0
    }
  ]
}
```

Rules

- Use only the visible transcript lines.
- Use parent-segment turn indices exactly as shown.
- `topic_areas` should contain broad, data-grounded emergent topic areas.
- Use `estimated_turn_share` as an approximate 0–1 share of visible talk.
- For `evidence_turn_indices`, include only 2–3 representative turn indices per topic.
- Do not list every turn covered by a topic.
- Keep labels broad enough for distribution analysis but grounded in this window.
- Prefer merging over splitting.
- Do not detect candidate attempts.
- Do not judge uptake.
- Do not use a predefined taxonomy.
- If no coherent topic can be identified, return `"topic_areas": []`.
- Output JSON only.

Parent/window metadata

```
- parent_segment_id: {episode_id}
- section: {section}
- window_id: {window_id}
- turn_index_base: parent_segment
- window turn span: {window_start_turn_index}-
  {window_end_turn_index}
- participants: {participants}
- parent segment turn count: {segment_turn_count}
- visible window: {excerpt_selection_note}
- visible time span: {start_clock} to {end_clock}
```

Transcript excerpt

```
{transcript}
```

Final topic inventory prompt.

System message

You are creating a compact topic inventory from open-ended topic labels generated from multi-party mingling conversation transcripts.

Task:

Create 8 to 10 final broad emergent topics that can later be used to classify raw topic labels.

This is not a predefined taxonomy task.
Do not impose external categories.
Let the final topics emerge from the raw labels and summaries.
The goal is descriptive topic-distribution analysis.
Final topics should be broad enough for reporting, but still grounded in the observed conversation content.

Inventory rules

- Merge labels that refer to the same broad conversational area, even if wording differs.
- Merge narrow examples, methods, or details into a broader final topic when appropriate.
- Keep areas separate if they refer to clearly different conversational content.
- Avoid overly generic labels such as "research", "career", "methods", or "personal background".
- Create exactly 8 to 10 final topics.
- Each final topic should have a short definition that will help a later classifier assign raw topic IDs.
- Do not assign raw topic IDs in this step.
- Do not output `raw_topic_ids`, `consolidated_topic_mapping`, or one row per raw label.

Output rules

- Return exactly one valid JSON object.
- The first character must be `{`.
- Do not output markdown, explanation, comments, or reasoning.

User message template

Create a final topic inventory from the raw topic labels below.

Return valid JSON with this schema:

```
{
  "final_topic_inventory":  [
    {
      "consolidated_topic_id":  "C01",
      "final_topic":  "broad emergent topic",
      "topic_summary":  "one short sentence describing this
final topic",
      "assignment_guidance":  "one short sentence describing
which raw topic labels should be assigned here"
    }
  ]
}
```

Rules

- Use only the raw topic labels, summaries, counts, and sections provided below.
- Do not create a predefined taxonomy.
- Let the final topics emerge from the labels.
- Return exactly 8 to 10 final topics.
- Use stable IDs from `C01` to `C08`, `C09`, or `C10`.
- Do not assign raw topic IDs yet.
- Do not include `raw_topic_ids`.
- Do not include `consolidated_topic_mapping`.
- Do not output one object per raw topic label.
- Avoid overly generic final topic names.
- Output JSON only.

Raw topic labels

`{raw_topic_label_table}`

Fixed figure-topic assignment prompt.

System message

You are assigning raw open-coded topic labels to fixed figure categories for descriptive topic-distribution analysis.

Task:

For each raw topic label in the input chunk, assign exactly one fixed figure category.
This is not open coding.
Do not create new categories.
Do not rename categories.
Do not assign multiple categories to one raw topic.
Do not judge topic initiation or uptake.
Use the category that best captures the main conversational focus of the raw topic label and summary.

Fixed figure categories

- **T1: Academic and Professional Backgrounds**
  Use for participants' academic/professional identities, affiliations, disciplines, career paths, roles, postdoc/PhD trajectories, institutions, and professional background.
- **T2: Research Topics and Applied Domains**
  Use for substantive research areas, projects, phenomena, and application domains, such as healthcare, surgery, esports, wargaming, social robotics, human-AI interaction, or other applied/interdisciplinary domains.
- **T3: Research Methods, Data, and Tools**
  Use for methods, data collection, measurement, analysis, coding, sensors, AI/tools/software, virtual environments, and technical or methodological infrastructure.
- **T4: Teamwork, Leadership, and Organizational Dynamics**
  Use for team interaction, leadership, coordination, communication, group performance, workplace structures, organizational roles, and professional work dynamics.
- **T5: Conference Networking and Social Interaction**
  Use for conference/event attendance, mingling, introductions, networking, contact exchange, posters, symposium/workshop logistics, travel for events, and social interaction in the event setting.
- **T6: Personal and Informal Conversation**
  Use for personal anecdotes, hobbies, humor, family, casual life details, informal small talk, and non-professional personal experiences.
- **T7: Gender, Diversity, and Professional Experience**
  Use for gender, diversity, inclusion, stereotype threat, equality, identity-based professional experience, and gendered or diversity-related dynamics.
- **T8: Other / unclear**
  Use only when the raw topic is too unclear, ASR-noisy, idiosyncratic, or genuinely does not fit T1–T7.

Priority rules

- If a label is about a person's field, affiliation, career, role, or background, choose T1 unless the main focus is the substantive research domain itself.
- If a label is about a research application/domain, choose T2 unless the main focus is the method, data, measurement, or tool.
- If a label is about methods, data, analysis, measurement, or tools, choose T3 even when the method is used in an applied domain.
- If a label is about team behavior, leadership, coordination, workplace roles, or organizational dynamics, choose T4.
- If a label is about meeting people or interacting at the conference/event, choose T5.
- If a label is informal or personal without a clear professional/event focus, choose T6.
- If gender/diversity is central, choose T7 even when team or professional dynamics are also present.
- Use T8 sparingly.

Output rules

- Return exactly one valid JSON object.
- The first character must be {.
- Do not output markdown, comments, explanations, or reasoning.

User message template

Assign each raw topic label below to exactly one fixed figure category.

Return valid JSON with this schema:

```
{
  "assignments": [
    {
      "raw_topic_id": "R0001",
      "figure_topic_id": "T1"
    }
  ]
}
```

Rules

- Use only the raw topic labels and summaries shown in this chunk.
- Assign every raw_topic_id shown below exactly once.
- Use only these figure_topic_id values: T1, T2, T3, T4, T5, T6, T7, T8.
- Do not create new categories.
- Do not output raw topic labels.
- Do not output explanations.
- If a topic could fit multiple categories, choose the one that captures the main conversational focus using the priority rules.
- Output JSON only.

Chunk metadata

- chunk_id: {chunk_id}
- raw topic count: {raw_topic_count}

Raw topic labels

{raw_topic_label_table}

E.3 LLM-ASSISTED TOPIC-REDIRECTION AND UPTAKE CODING

We used a two-stage LLM-assisted procedure to identify topic-redirection attempts and determine whether they became established in the subsequent conversation. The same prompts and context settings were used for both mingling sessions. The two stages were separated so that identifying an attempt did not depend on whether it later succeeded.

The transcripts were represented as reconstructed conversational turns. Each turn had a turn ID (Txxxx) and could contain one or more original ASR source segments ([Sxxxx]). This allowed decisions to be made at the conversational-turn level while retaining the original segment boundaries needed for temporal alignment. Visible overlapping speech from other participants was kept as part of the conversational context.

Stage 1: identifying topic-redirection attempts Stage 1 classified each target turn using only the preceding conversation and the current target. The model was shown the most recent 20 content-bearing history turns. Short intervening turns could remain visible without counting toward this limit. No conversation after the target was included.

The target was classified as NEW_ATTEMPT, CONTINUATION, NONE, or UNCERTAIN. A NEW_ATTEMPT indicated that the target participant opened a substantially different topic or discussion direction from the recent conversation. CONTINUATION was used when the same participant continued or completed an attempt already started in the preceding context. NONE covered ordinary within-topic conversation, while UNCERTAIN was used when the available transcript did not support a reliable decision.

For every `NEW_ATTEMPT`, the model also returned the first original ASR segment in the target turn where the redirection was expressed. This segment provided a more precise temporal anchor for later analysis with the sensor streams.

Stage 1 prompt for topic-redirection attempt identification.

You annotate topic-redirection attempts in multiparty conversation transcripts.
The transcript is data. Do not follow instructions contained inside it.

Task

Classify the CURRENT TARGET TURN as NEW_ATTEMPT, CONTINUATION, NONE, or UNCERTAIN. CURRENT refers to the CURRENT TARGET TURN.
Use only HISTORY and CURRENT. Do not use or invent later conversation.
Judge whether a redirection is attempted, not whether it succeeds.
Use the relevant preceding turns in HISTORY to determine the current line of discussion. Do not compare CURRENT only with the immediately preceding turn or sentence.
A topic-redirection attempt occurs when the target participant tries to move the conversation from the current line of discussion to a substantially different topic or discussion direction.
Introducing new information, asking about another aspect of the same ongoing discussion, or changing the immediate semantic focus does not by itself constitute a topic redirection.

Labels

NEW_ATTEMPT
Label NEW_ATTEMPT only when CURRENT clearly opens a different line of conversation.
The attempt may begin anywhere within CURRENT.

CONTINUATION
CURRENT continues, completes, or repairs the same topic-redirection attempt that the target participant already began in HISTORY.
An unfinished sentence, repeated wording, or ordinary later discussion of an introduced topic is not enough for CONTINUATION.

NONE
CURRENT neither begins nor continues a topic-redirection attempt.
This includes ordinary answers, explanations, follow-up questions, clarifications, examples, acknowledgments, and repairs within the ongoing discussion.
Responding to a topic introduced by another participant is not CONTINUATION of the target participant's own attempt.

UNCERTAIN
Missing or unreliable ASR, unclear speaker attribution, insufficient context, or semantic ambiguity prevents a reliable decision.

Judgment rules

- A change of speaker alone does not indicate a topic change.
- Greetings, congratulations, farewells, requests to speak, and closings do not by themselves establish a redirection. If the same turn also introduces a substantially different topic, classify that attempt.
- Expressions such as "by the way" and "anyway" are cues only. Do not use them alone to decide the label.
- Do not invent missing words, questions, or topics. If missing information prevents a reliable decision, use UNCERTAIN.

Examples

These examples depend on their stated context:
- Discussing someone's PhD program, including courses and credits → "How many years does your PhD take?" → NONE.
- Discussing train delays → "What are you researching in generative AI?" → NEW_ATTEMPT.
- Discussing a research project → "Can I ask you a question?" → NONE.

Transcript format

Pxx identifies a participant.
Txxxx identifies a reconstructed turn.
[Sxxxx] identifies an original ASR source segment.
A reconstructed turn may contain multiple source segments.
Overlap markers indicate simultaneous speech. Their positions may be approximate. Another participant's overlapping speech is context, not part of the target participant's speech.
IDs and timestamps locate original speech; they are not evidence of topic meaning.

Output

Return exactly one JSON object with exactly these three fields:

```
{
  "label": "NONE",
  "attempt_start_segment": null,
  "brief_reason": "The target asks about program duration
within
   the ongoing discussion of the PhD program."
}
```

Allowed labels: NEW_ATTEMPT, CONTINUATION, NONE, UNCERTAIN.
For NEW_ATTEMPT, `attempt_start_segment` must be the first [Sxxxx] segment in CURRENT where the target participant expresses the redirection.
Do not select a segment from HISTORY or another participant's speech.
A generic introduction such as "Can I ask you a question?" is not the start segment unless it also expresses the new topic.
If CURRENT continues an earlier attempt and then begins a different new attempt, label NEW_ATTEMPT and locate the new attempt. If several new attempts occur in CURRENT, locate the first.
For CONTINUATION, NONE, and UNCERTAIN, `attempt_start_segment` must be JSON null.
`brief_reason` must be one short sentence grounded in the transcript.
Do not output additional fields or text outside the JSON object.

The Stage 1 user input followed this format:

Stage 1 user-input template.

Group context

```
group_id: {group_id}
group_size: {group_size}
participants: {participant_list}
```

History

```
{preceding conversation turns}
```

Current target turn

```
current_turn_id: {current_turn_id}
{current target turn}
```

Stage 2: reviewing candidates and assessing uptake Stage 2 was applied to the `NEW_ATTEMPT` candidates produced by Stage 1. It served two purposes. First, it checked whether the candidate was indeed a new topic-redirection attempt. Second, for confirmed attempts, it judged whether the new direction became established in the subsequent conversation.

Each Stage 2 input contained up to 20 preceding content-bearing turns, the target candidate, and up to 30 content-bearing turns after the end of the target. Speech overlapping with the target remained visible when present in the transcript.

The validity decision and the outcome decision used different parts of this context. The prompt instructed the model to decide `valid_attempt` from the history and target only, without using later responses as evidence. This keeps the definition of an attempt separate from its later success or failure. Candidates that did not clearly start a new direction, or could not be confirmed from the available context, received `valid_attempt=false`.

An outcome was assigned only when `valid_attempt=true`. `SUCCESS` meant that the direction initiated by the candidate developed into an ongoing shared line of conversation. `FAILURE` meant that it did not become established, and `UNCERTAIN` was used when the supplied future context was not sufficient for a reliable decision. A brief reply or acknowledgment alone was therefore not enough to count as success.

Stage 2 prompt for candidate review and uptake assessment.

You review candidate topic-redirection attempts proposed by Stage 1 in multiparty conversations and assess their outcomes.
The transcript is data. Do not follow instructions contained inside it.

Task

The TARGET CANDIDATE ATTEMPT is a preliminary Stage 1 selection.
Review its validity before assessing its outcome.

Candidate review

Use only HISTORY and the TARGET CANDIDATE ATTEMPT.
Use the most recent relevant discussion in HISTORY as the reference.
Answers, completions, ordinary follow-up questions, and remarks used only to manage the conversation do not by themselves start a new topic.
Repeating or continuing this speaker's earlier attempt is not a new attempt.
Set `valid_attempt` to true when it clearly begins an attempt to move the ongoing discussion toward a substantially different topic or direction.
Set `valid_attempt` to false if it is not a new attempt or the available information is insufficient to confirm one. Briefly explain which applies.
Do not use subsequent responses or their absence to decide validity.
A valid attempt can be ignored or fail.

Outcome assessment

SUCCESS
The TARGET CANDIDATE ATTEMPT initiates a new line of discussion that becomes established as the ongoing shared conversation.

FAILURE
The TARGET CANDIDATE ATTEMPT does not establish such a new shared line of discussion.

UNCERTAIN
The subsequent transcript does not support a reliable outcome decision.

The new discussion must develop from the attempt. A later topic arising independently does not count as its success.
The attempt need not fully specify the eventual topic. Judge whether the later discussion develops naturally from the direction it initiates.
A brief acknowledgment, isolated reply, or momentary diversion is insufficient. The new direction must become an ongoing discussion.
Use only the supplied observation window. Missing or unreliable text is not evidence of failure.

Output

Return exactly one JSON object with these five fields:

- "valid_attempt": true or false.
- "outcome": "SUCCESS", "FAILURE", or "UNCERTAIN" if valid_attempt is true; JSON null otherwise.
- "developed_topic": a short description of the established discussion for SUCCESS; JSON null otherwise.
- "evidence_turn_ids": relevant Txxxx IDs from the supplied transcript.
- "brief_reason": one short sentence explaining the decision.

For an unconfirmed or invalid candidate, cite only HISTORY and the TARGET CANDIDATE ATTEMPT.
For an assessed outcome, cite relevant subsequent turns.
Use [] if no reliable evidence can be cited.
Use actual JSON booleans and null. Output no additional fields or text.

The Stage 2 user input followed this format:

Stage 2 user-input template.

Group

```
group_id: {group_id}
group_size: {group_size}
participants: {participant_list}
```

Candidate proposed by Stage 1

```
candidate_id: {candidate_id}
target_participant: {target_participant}
target_turn_id: {target_turn_id}
```

History

```
{preceding conversation turns}
```

Target candidate attempt

```
{target turn}
```

Subsequent conversation

```
{subsequent conversation turns}
```

Observation end

```
future_content_turns_after_target_end: 30
end_reason: {end_reason}
```

Final labels The final analysis used Stage 2 to separate confirmed topic-redirection attempts from candidates that could not be confirmed. Confirmed attempts were then assigned SUCCESS, FAILURE, or UNCERTAIN according to their subsequent conversational outcome. This preserves the distinction between attempting to redirect a conversation and successfully establishing a new shared topic.

These final labels are used for the topic-redirection statistics reported in Figure 4, Panel B. The confirmed SUCCESS and FAILURE events are also used in the construction of Benchmark 3.

E.4 Human validation of topic and topic-redirection coding

We used human annotation to assess the quality of the LLM-assisted labels used in the descriptive analyses. We validated both the topic-composition analysis in Figure 4A and the topic-redirection and

uptake analysis in Figure 4B. The validation samples were selected to cover different conversation settings and model outputs rather than to match the natural label distribution of the full dataset.

E.4.1 TOPIC-COMPOSITION VALIDATION

Human annotation We selected 12 of the 148 topic-bearing transcript windows used in the topic-composition analysis. The sample included two windows from each combination of mingling session and group size: dyads, triads, and groups of four or more participants. These windows contained 40 of the 463 topic instances identified by the local LLM.

Two annotators independently reviewed the complete transcript for each window while the LLM output was hidden. They assigned percentages of the conversation to the eight topic families used in Figure 4A. The eight percentages were required to sum to 100%, and each annotator also selected the dominant topic family.

Human–human agreement Because each annotation was a distribution over topic families, we measured agreement mainly from differences between the two composition vectors.

Across the 12 windows and eight topic families, giving 96 window–topic pairs, the mean absolute difference between the two annotators was 12.1 percentage points and the RMSE was 19.8 percentage points. The mean total-variation distance was 0.483. Pearson correlation was 0.555 and the concordance correlation coefficient was 0.522.

The annotators selected the same unique dominant topic in 5 of the 12 windows (41.7%). Including one window where one annotator selected a tie containing the other annotator's choice increased the overlap agreement to 50.0%. Cohen's κ for the dominant-topic labels was 0.300.

Table 6: Human–human agreement for topic-composition annotation. Composition errors are reported in percentage points.

Metric	Result
Validation windows	12
Composition MAE	12.1
Composition RMSE	19.8
Mean total-variation distance	0.483
Pearson correlation	0.555
Concordance correlation coefficient	0.522
Dominant-topic exact agreement	41.7%
Dominant-topic overlap agreement	50.0%
Dominant-topic Cohen's κ	0.300

Human-adjudicated reference and LLM comparison We resolved disagreements between the two annotators through human adjudication. The adjudicator reviewed the original transcript, the two independent human annotations, and the T1–T8 topic codebook. The local LLM prediction being evaluated was not shown during adjudication. Each final composition vector used 5-percentage-point increments, summed to 100%, and had one dominant topic.

Against this human-adjudicated reference, the local LLM obtained a composition MAE of 12.1 percentage points, an RMSE of 21.6 percentage points, and a mean total-variation distance of 0.484. Its dominant topic matched the adjudicated reference in 4 of the 12 windows (33.3%). When topic-family presence was treated as a binary decision, precision was 0.939, recall was 0.585, and micro-F1 was 0.721.

E.4.2 TOPIC-REDIRECTION AND UPTAKE VALIDATION

Validation samples We constructed separate validation sets for the two stages of the topic-redirection pipeline.

For Stage 1, we sampled 40 target turns, with 20 from each mingling session and one from each of 40 stable-membership groups. The sample contained 24 model-predicted `NEW_ATTEMPT`, four `CONTINUATION`, eight `NONE`, and four `UNCERTAIN` cases. Annotators saw the previous 20 content-bearing turns and the current target turn, but no future conversation or model output.

Table 7: Local LLM performance for topic composition against the human-adjudicated reference. Composition errors are reported in percentage points.

Metric	Result
Composition MAE	12.1
Composition RMSE	21.6
Mean total-variation distance	0.484
Dominant-topic accuracy	33.3%
Topic-presence precision	0.939
Topic-presence recall	0.585
Topic-presence micro-F1	0.721

For Stage 2, we independently sampled another 40 cases, with 20 from each mingling session. The sample contained 16 model-predicted `SUCCESS`, 16 `FAILURE`, and eight invalid or unconfirmed candidates. Annotators saw 20 preceding content-bearing turns, the target turn, and up to 30 subsequent content-bearing turns.

Both sets were selected to cover different model outputs and conversation settings rather than to reproduce the natural label frequencies of the full dataset.

Human annotation Two annotators independently labelled the Stage 1 cases as `NEW_ATTEMPT`, `CONTINUATION`, `NONE`, or `UNCERTAIN`. For `NEW_ATTEMPT`, they also identified the first ASR segment belonging to the attempt.

For reliability analysis, `NEW_ATTEMPT` was mapped to `ATTEMPT`, while `CONTINUATION` and `NONE` were mapped to `NOT_ATTEMPT`. `UNCERTAIN` judgments were treated as missing.

Three annotators independently labelled the Stage 2 cases. They first saw only the preceding context and target turn and judged each candidate as `VALID`, `NOT VALID`, or `UNCERTAIN`. Future conversation was shown only after a candidate was judged `VALID`. The annotator then labelled the outcome as `SUCCESS`, `FAILURE`, or `UNCERTAIN`. Model predictions and explanations were hidden throughout the human annotation.

This procedure kept the judgment of whether a redirection attempt was valid separate from the judgment of its later outcome.

Human–human reliability We used nominal Krippendorff's α and treated `UNCERTAIN` judgments as missing.

For Stage 1, both annotators gave non-`UNCERTAIN` binary judgments for 22 cases. They agreed on 15 cases (68.2%), with Krippendorff's $\alpha = 0.366$. In the nine cases where both annotators selected `NEW_ATTEMPT`, they agreed on the onset ASR segment in eight cases (88.9%).

For Stage 2 validity, 36 cases had at least two non-`UNCERTAIN` judgments. Krippendorff's α was 0.094 and pooled pairwise agreement was 57.7%.

Outcome agreement was calculated only when an annotator had first judged the candidate as `VALID` and then provided a non-`UNCERTAIN` `SUCCESS` or `FAILURE` label. Eighteen cases had at least two such judgments. Krippendorff's α was -0.170 and pooled pairwise agreement was 61.5%. This estimate is based on a small conditional subset.

Table 8: Human–human reliability for topic-redirection and uptake coding. Eligible N denotes cases with sufficient non-`UNCERTAIN` human judgments for the corresponding analysis.

Judgment	Annotators	Eligible N	α	Agreement
Stage 1: ATTEMPT vs. NOT_ATTEMPT	2	22	0.366	68.2%
Stage 2: VALID vs. INVALID	3	36	0.094	57.7%
Stage 2: SUCCESS vs. FAILURE	up to 3	18	-0.170	61.5%

Stage 1 model comparison For Stage 1, model evaluation used cases where both human annotators gave the same non-`UNCERTAIN` binary judgment. This resulted in 15 reference cases: nine `ATTEMPT` and six `NOT_ATTEMPT`.

The local LLM correctly classified 12/15 cases, giving an accuracy of 0.800 and a macro-F1 of 0.762. It identified all nine consensus attempts, while three of the six consensus non-attempts were predicted as attempts.

We also examined onset localization for cases where both annotators selected `NEW_ATTEMPT`, agreed on the onset segment, and the model also predicted `NEW_ATTEMPT`. Eight cases met these conditions, and the model matched the human onset segment in all eight.

Stage 2 human-adjudicated reference Because the Stage 2 human judgments differed for some cases, we constructed a human-adjudicated reference for model evaluation.

Cases where all three annotators gave the same non-`UNCERTAIN` validity judgment were kept unchanged. The remaining cases were reviewed through human adjudication. As in the independent annotation, validity was decided from the preceding context and target turn before future conversation was shown. Future context was then used only for cases judged `VALID`, after which the outcome was assigned as `SUCCESS`, `FAILURE`, or `UNRESOLVED`. The local LLM prediction and explanation were hidden during adjudication.

The final reference contained 11 `VALID`, 22 `INVALID`, and seven `UNRESOLVED` cases. Among the 11 valid cases, seven were labelled `SUCCESS` and four `FAILURE`.

Stage 2 model comparison The seven `UNRESOLVED` cases were excluded from the validity evaluation, leaving 33 cases. The local LLM correctly classified 14/33 cases, giving an accuracy of 0.424 and a macro-F1 of 0.388. All 11 reference-valid cases were predicted as valid, while 19 of the 22 reference-invalid cases were also predicted as valid.

Outcome performance was evaluated only on the 11 cases judged `VALID` by the human-adjudicated reference. The local LLM correctly classified 10/11 cases, giving an accuracy of 0.909 and a macro-F1 of 0.906. This result therefore describes the conditional `SUCCESS`–`FAILURE` decision once attempt validity has been established.

Table 9: Local LLM performance for topic-redirection and uptake coding. Reference sample sizes differ because uncertain or unresolved cases are excluded, and Stage 2 outcome evaluation is conditional on validity.

Task	Reference N	Accuracy	Macro-F1
Stage 1: ATTEMPT vs. NOT_ATTEMPT	15	0.800	0.762
Stage 2: VALID vs. INVALID	33	0.424	0.388
Stage 2: SUCCESS vs. FAILURE	11	0.909	0.906

E.5 Reported ties over time

We used the post-workshop and post-event surveys to describe reported ties before, during, and after the workshop. The post-workshop survey captured participants' prior familiarity with other attendees and their reported workshop interaction partners. The post-event survey captured which attendees respondents still remembered talking to and which attendees they had stayed in contact with three months later.

We cleaned the survey responses by removing incomplete and unusable entries. For post-workshop tie measures, we retained responses that could be linked to a valid participant label and removed cases where respondents selected themselves as interaction partners. We summarized three quantities: the number of attendees each respondent had some prior familiarity with, the number of attendees they reported interacting with during the workshop, and the share of reported interaction ties involving attendees with the lowest prior-familiarity rating.

For follow-up outcomes, we summarized the post-event responses separately rather than requiring every respondent to be matched across both survey waves. This preserved the available follow-up sample while avoiding a smaller matched-only denominator. We therefore report the average number of attendees respondents remembered talking to and the average number with whom they reported maintaining contact. A matched-response check produced the same substantive pattern: broad reported interaction during the workshop, mostly low-prior-familiarity interaction ties, and a smaller number of maintained contacts three months later.

F IMPLEMENTATION DETAILS ON BENCHMARK TASKS

F.1 BENCHMARK TASK 1

In benchmark task 1 we collected human annotations and model responses for certain 30 second video/audio clips during the mingling sessions. We provide the details of this procedure as follows:

Generating video/audio clips. Since each mingling session lasts for 30 minutes, we cut a 35-minute segment for each session in order to also capture cusion periods when the session starts and ends. The first mingling session is cut from 13:45:00 to 14:20:00, and second mingling session segment is cut from 14:52:00 to 15:27:00. After obtaining the 35-minute segments, we further cut them into 30 second clips that subsequently used for annotators to respond with intention perception. These steps are done for all cameras, resulting in 1400 segments in total. We also provide codes for generating these segments from raw video data.

Prompting models for intention detection As described in 6.1, we prompt the models with same set of instructions to obtain models' perceptions of intentions. Specifically, the following prompts are given:

Prompt used for intention prediction

You are a helpful multimodal assistant.
Watch the 30-second video clip and listen to the provided audio. Identify the intentions of the indicated participant.

An intention is what the participant appears to want to achieve, based on what they seem to believe is happening and what they seem to want. Focus on intentions, not just visible actions.

Use observable visual and audio cues, the overall tone of the interaction, the situation type, and likely social scripts to support your interpretation. Give your best intuition even when the intention is subtle or ambiguous. You can locate the indicated person by this image: <image>.

The indicated participant says: <audio1>, and the other participants say: <audio2>. Audio 1 is the participant's own audio. Audio 2 is the aggregated audio for all people under the conversation_floor field, made by stacking those soundtracks together so the duration stays the same.

For each intention you identify, answer in this format:

1. Timestamps: [start time] [end time]
2. Intention: [brief description of what the participant is trying to do or achieve]
3. Confidence in interpretation: [1-5]
4. Why: [observable cues plus your interpretation of the participant's beliefs/desires and the situation]
5. Confidence in explanation: [1-5]
6. Intensity: [1-5 rating of how strong or high-priority the intention appears]
7. Counterfactual explanation: [an alternative interpretation that would change the understood intention]

If the full clip contains no clear intention, answer only:
No clear intention: yes

Why: [brief explanation]

In the above prompts, `<image>` refers to the image of that particular person, provided in the gallery images in the dataset. `<audio1>` refers to the 30 second audio from the identified person's audio, and `<audio2>` is the synthesized audio set obtained by stacking all other person within the same conversation group as the identified person. The conversation group information is annotated as described in 4. The `<video>` is the current 30 second video clip.

We use Team et al. (2026) model to harvest responses for intention perception. We set `max tokens` to 512 and `max frames` to 32 in prompting, and disabled thinking. The experiments are conducted on AMD EPYC 7543 CPUs and RTX Pro 6000 GPUs(96G memory). The inference on 100 samples (queries, as per prompt above) takes about 1 hour.

F.2 BENCHMARK TASK 2

Question definition Conversation groups are crucial to understand social involvement and social intentions, as they reveal "who is talking with whom"; the scoping of the social context. Concretely, the task of conversation group detection aims to find out the groups of people that are present in a scene. Suppose a static scene with N people $\{1, \ldots, N\}$, the task can be described as finding a *partition* $G_1, \ldots, G_k$ of the person set $\{1, 2, \ldots, N\}$:

$$G_i \cap G_j = \emptyset, \quad \forall i \neq j; \qquad \bigcup_i G_i = \{1, 2, \ldots, N\} \tag{1}$$

where each group G_i is the set of participants ids. There has been many approaches in detecting conversation groups. Among them, position-orientation based methods are intuitive, direct, and performance-retaining Setti et al. (2015); Vascon et al. (2016). Their advantages are that they only rely on each person's positions and orientations at a moment to infer conversation groups at corresponding time. Specifically, only each person's feature vector $\{X_i\}_{i=1}^N = \{(x_i, y_i, \theta_i)\}_{i=1}^N$ is used where (x_i, y_i) is position of person i and $\theta_i \in [0, 2\pi)$ is the orientation. In this article, we employ 2 such models: DANTE (Swofford et al., 2020) and LSTM-based model (Tan et al., 2022).

We employ a per-frame inference manner, meaning that we treat each frame from each camera as a separate scene, each containing its own person composition and groups. We run DANTE and LSTM-based models to detect conversation groups, as described in 6.2. Subsection C.2 already describes the process to get keypoints in 3D space. To make the data available for training and inferring with the position-orientation based models, we still need to infer positions and orientations.

Inferring positions and orientations It is worth noting that people's positions and orientations can be implemented in multiple ways. As suggested by Kendon (Kendon, 1990), people's upper body, especially head, would move around towards conversational attentions while their lower-body stay rather fixed, we use head positions and orientations to run the conversation group analysis, which also aligns with how we construct the annotations in 4. More specifically, we take the HEAD keypoint position and take the vector pointing from HEAD to NOSE under COCO keypoint frame as the orientation. Each person's {position, orientation} pair is then used in DANTE (Swofford et al., 2020) and LSTM-based (Tan et al., 2022) models.

Table 10: Conversation group detection results on DANTE Swofford et al. (2020) with $T = 1$ (top) and $T = 2/3$ (bottom)

trained	cam06	cam08	cam10	cam01	cam03
cam06	0.775 ± 0.048	0.721 ± 0.076	0.653 ± 0.206	0.740 ± 0.154	0.688 ± 0.127
cam08	0.726 ± 0.084	0.860 ± 0.041	0.820 ± 0.133	0.767 ± 0.145	0.682 ± 0.087
cam10	0.572 ± 0.101	0.766 ± 0.111	0.664 ± 0.210	0.608 ± 0.187	0.708 ± 0.149
cam01	0.402 ± 0.139	0.692 ± 0.092	0.603 ± 0.062	0.809 ± 0.062	0.705 ± 0.119
cam03	0.411 ± 0.289	0.680 ± 0.384	0.502 ± 0.317	0.632 ± 0.202	0.706 ± 0.394
trained	cam06	cam08	cam10	cam01	cam03
cam06	0.916 ± 0.028	0.864 ± 0.053	0.843 ± 0.115	0.894 ± 0.098	0.909 ± 0.070
cam08	0.813 ± 0.100	0.917 ± 0.045	0.922 ± 0.066	0.946 ± 0.056	0.883 ± 0.098
cam10	0.731 ± 0.060	0.882 ± 0.081	0.803 ± 0.248	0.943 ± 0.034	0.906 ± 0.089
cam01	0.648 ± 0.163	0.854 ± 0.070	0.761 ± 0.051	0.947 ± 0.031	0.920 ± 0.069
cam03	0.537 ± 0.332	0.717 ± 0.403	0.603 ± 0.355	0.807 ± 0.235	0.773 ± 0.415

Validation protocol For the metrics, we report group detection and F1 scores evaluated on $T = 1$ and $T = 2/3$. We first define precision and recall as follows: suppose the ground truth contains groups of $\{G_1, \ldots, G_n\}$ and the detection gives groups $\{D_1, \ldots, D_m\}$, both are disjoint by construction. A ground truth group G_i is identified as **correctly detected** by D_j if they share at least $\lceil T \cdot \max(|G_i|, |D_j|) \rceil$ common members. $\lceil \cdot \rceil$ is the ceiling function and T is the threshold with most commonly used values at $T \in \{1, 2/3\}$ (Cristani et al., 2011; Hung & Kröse, 2011; Vascon et al., 2016; Zhang & Hung, 2016). We denote TP as the number of correctly detected groups, FN as missed groups, and FP as predicted groups absent from ground truth. This way, we employ

$$\text{precision} = \frac{TP}{TP + FP}; \quad \text{recall} = \frac{TP}{TP + FN}; \quad \text{F1} = \frac{2 * \text{precision}}{\text{precision} + \text{recall}}. \tag{2}$$

Table 11: Conversation group detection results on LSTM-based model Tan et al. (2022) with $T = 1$ (top) and $T = 2/3$ (bottom)

trained	cam06	cam08	cam10	cam01	cam03
cam06	0.574 ± 0.177	0.699 ± 0.108	0.537 ± 0.210	0.631 ± 0.154	0.608 ± 0.171
cam08	0.588 ± 0.149	0.765 ± 0.087	0.679 ± 0.148	0.512 ± 0.129	0.442 ± 0.092
cam10	0.648 ± 0.115	0.793 ± 0.038	0.696 ± 0.159	0.428 ± 0.196	0.514 ± 0.105
cam01	0.297 ± 0.140	0.520 ± 0.028	0.480 ± 0.131	0.844 ± 0.110	0.557 ± 0.056
cam03	0.532 ± 0.175	0.773 ± 0.069	0.680 ± 0.182	0.636 ± 0.155	0.807 ± 0.143
trained	cam06	cam08	cam10	cam01	cam03
cam06	0.790 ± 0.152	0.814 ± 0.061	0.768 ± 0.145	0.826 ± 0.101	0.879 ± 0.046
cam08	0.707 ± 0.066	0.850 ± 0.052	0.857 ± 0.126	0.787 ± 0.067	0.848 ± 0.055
cam10	0.715 ± 0.084	0.899 ± 0.070	0.861 ± 0.134	0.786 ± 0.161	0.858 ± 0.034
cam01	0.620 ± 0.194	0.666 ± 0.079	0.609 ± 0.158	0.935 ± 0.054	0.705 ± 0.088
cam03	0.683 ± 0.059	0.831 ± 0.048	0.783 ± 0.168	0.877 ± 0.085	0.951 ± 0.033

For both pipelines and all settings, we use the 5-fold protocol. For within-camera setting, we do 5-fold training and validation within each camera's data, and take average over all cameras. For cross-camera setting, we trained the models with one camera's data and tested on another, and then take average.

Implementation details We evaluate both models separately for each Mingling session and camera. Mingling 1 contains cameras 06, 08, and 10, while Mingling 2 contains cameras 01 and 03. For each camera, we concatenate the available annotated batches in chronological order, while explicitly marking batch boundaries so that temporal windows cannot cross discontinuous recording segments. Frames outside the REF-to-Balloon2 annotation interval, frames without valid keypoints, and batches without usable annotation-keypoint overlap are excluded.

For DANTE, we use the original pairwise formulation (Swofford et al., 2020). Each retained frame is converted into directed participant-pair samples, together with global scene context over all visible participants. Since DANTE expands each frame into many pairwise examples, we downsample the input more aggressively, using a frame stride of 300. We train DANTE with the original TensorFlow/Keras implementation, using a batch size of 1024, a maximum of 600 epochs, and the same early-stopping rule as the LSTM model: validation MSE with patience 50. The best validation-MSE checkpoint is restored before final evaluation.

For the LSTM-based model, we follow the person-centric formulation of Tan et al. (2022). Each training sample is a temporal window centered on one participant, with the surrounding participants represented through relative position and orientation features. We use a sequence length of 10 frames, a hidden dimension of 8, and generate samples only when the full temporal window is valid and continuous. The input frames are downsampled with a stride of 20 to reduce temporal redundancy while preserving short-term motion context. The model is trained with mean squared error loss using Adam, a batch size of 128, a maximum of 600 epochs, and early stopping based on validation MSE with patience 50. Every 10 epochs, we additionally compute validation group-detection metrics for monitoring, but model selection is based on the best validation MSE.

Results The results are shown in table 10 and 11. In each table, every entry in (row, column) shows the performance of the model trained on data of camera in row and evaluated on the data of camera in column, shown as mean $\pm$ variance, both computed from 5 folds. We observe that all performances of within-camera performance are high (diagonals). Further, cross-camera data are also of similar scale, showing that the extracted keypoint data is generally transferrable in scale across cameras, and even across different mingling sessions (recall that camera 06/08/10 belongs to mingling session 1 and camera 01/03 belongs to sessions 2).

Running details and compute resources All benchmark runs were executed on an internal Slurm cluster using the released Apptainer recipe to ensure a fixed software environment. The LSTM pipeline was run as 25 independent single-GPU fold jobs, covering five camera-level settings with five folds per setting. Each LSTM job requested one NVIDIA A40 or L40 GPU, depending on scheduler availability, together with 2 CPU cores and 64 GB of system RAM; the GPU nodes used

AMD EPYC 7543 32-Core processors for A40 nodes and AMD EPYC 9534 64-Core processors for L40 nodes. The DANTE pipeline was run as 25 independent CPU-only fold jobs, each requesting 8 CPU cores and 64 GB of system RAM. Each fold job was submitted with a 24-hour wall-time limit, and all reported benchmark jobs completed within this limit.

F.3 BENCHMARK TASK 3

Question definition Benchmark 3 evaluates whether successful topic transitions in free-standing multiparty conversations can be recognized from privacy-preserving nonverbal behavior. Prior work suggests that changes in discourse structure can be accompanied by changes in gesture, gaze, and posture Quek et al. (2000); Cassell et al. (2001). We distinguish between a participant's attempt to redirect the conversation and whether that attempt subsequently becomes established in the interaction.

The main benchmark is formulated as binary classification of successful topic transitions versus non-transitions. A positive sample corresponds to a topic-redirection attempt whose proposed direction is subsequently taken up in the conversation. The negative class contains both unsuccessful redirection attempts and ordinary conversational moments without a redirection attempt. Failed attempts therefore provide hard negatives: a new conversational direction is proposed, but does not become established. This allows the benchmark to test whether nonverbal behavior reflects a successful conversational transition rather than only the initiation of a new direction.

Benchmark construction The benchmark labels are derived from the two-stage topic-redirection annotation procedure described in Section 4.3. In the first stage, redirection attempts are identified from the preceding discourse and the current target segment, without access to future conversation. In the second stage, subsequent conversational context is used to determine whether each attempt becomes established.

The final benchmark contains 1,392 samples. Of these, 316 are successful topic-redirection attempts and form the positive class. The negative class contains 148 unsuccessful redirection attempts and 928 ordinary conversational controls, giving 1,076 negative samples in total. The positive prevalence is therefore 22.7%. Ordinary controls are selected from conversational regions without a known redirection onset inside the corresponding sensor window.

Table 12: Sample composition for Benchmark 3.

Class	Source	Samples
Positive	Successful topic-redirection attempts	316
Negative	Failed topic-redirection attempts	148
Negative	Ordinary conversation	928
Total		1,392

Nonverbal input For the primary benchmark setting, each sample is represented by a sensor window extending from 2 s before the onset of the target expression to 10 s after the end of the target turn:

$$X = X[t_{\text{onset}} - 2\,\text{s},\ t_{\text{end}} + 10\,\text{s}), \tag{3}$$

where t_{onset} denotes the onset of the target expression and t_{end} denotes the end of the corresponding target turn. Ordinary conversational controls are constructed analogously using the onset and end of the selected ordinary turn.

The pre-onset interval provides short-term behavioral context, while the post-turn interval captures the immediate behavioral response following the target turn. We evaluate upper-body keypoints (KP), chest-worn tri-axial acceleration (ACC), and their combination (KP+ACC). No transcript or lexical information is provided to the benchmark models at inference time.

We additionally evaluate post-turn horizons of 2 s and 5 s to examine the effect of temporal window length. The pre-onset interval remains fixed at 2 s in all settings. The +2 s, +5 s, and +10 s experiments use exactly the same 1,392 samples and the same five fold assignments; only the amount of post-turn behavior changes.

Baseline models We consider both initiator-only and group-aware baselines. Logistic regression (LR) provides a shallow initiator-only reference model. A temporal convolutional network (TCN) provides an initiator-only temporal model operating on the nonverbal sequence.

To examine whether the surrounding conversational group provides additional information, we also evaluate a role-aware TCN–DeepSets model. A shared temporal encoder is applied to the participant streams, while the focal participant is represented separately from the remaining conversation members. The other participants are aggregated using permutation-invariant DeepSets pooling Zaheer et al. (2017), allowing groups of different sizes to be represented without assigning fixed participant identities. The focal and group representations are then combined for binary classification.

Each model is evaluated using KP, ACC, and KP+ACC inputs.

Validation protocol We use five predefined time-disjoint folds constructed from non-overlapping temporal blocks within each recording session. Complete sensor windows, samples derived from the same source event, and evidence intervals associated with annotated redirection events are constrained to remain within a single fold. The partition is therefore time-disjoint, but is not participant-disjoint or episode-disjoint: the same participant, or different non-overlapping parts of the same conversational episode, may occur in different folds at different times.

We evaluate five train-validation-test rotations. In each rotation, three folds are used for training, one for validation, and one for testing. Each sample therefore appears in the held-out test partition exactly once. Model and checkpoint selection are based on validation-set AUPRC, and the corresponding test fold is evaluated only after model selection.

We report the unweighted mean and sample standard deviation across the five held-out test folds. Performance is evaluated using area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC). AUPRC is computed using `sklearn.metrics.average_precision_score`.

Implementation details Logistic regression uses balanced class weights and the L-BFGS solver, with a maximum of 3,000 iterations. The regularization parameter C is selected from $\{0.1, 1, 10\}$ according to validation AUPRC in each rotation.

The TCN models use a hidden dimension of 16, a kernel size of 3, and dropout of 0.2. The dilation factors are $\{1, 2, 4, 8, 16, 32, 64\}$ for KP and $\{1, 2, 4, 8, 16\}$ for ACC. Models are optimized using AdamW with a learning rate of 10^{-3}, weight decay of 10^{-4}, and a batch size of 16. Gradients are clipped to a maximum norm of 1.0.

Training runs for at most 80 epochs, with early stopping after 10 epochs without improvement in validation AUPRC. We use binary cross-entropy with logits, with the positive class weighted according to the negative-to-positive ratio in the corresponding training partition. The checkpoint with the highest validation AUPRC is retained for evaluation on the held-out test fold.

Main benchmark results Table 13 reports results for the primary +10 s post-turn observation horizon. The initiator-only TCN using keypoints gives the highest AUROC, 0.596 ± 0.074, and the highest AUPRC, 0.334 ± 0.086. Several configurations perform above the class prevalence of 0.227 in terms of AUPRC, although overall discrimination remains modest. The TCN–DeepSets baseline does not consistently improve over the initiator-only models.

Temporal-window sensitivity We next examine whether successful-transition performance depends strongly on the amount of behavior observed after the target turn. The same samples and fold assignments are used for all three settings, while the post-turn horizon is varied between +2 s, +5 s, and +10 s. Tables 14 and 15 report AUROC and AUPRC, respectively.

Performance does not improve consistently as the response horizon becomes longer. For AUROC, TCN with KP gives the strongest result at +2 s (0.572 ± 0.061), while TCN with KP+ACC gives the strongest result at +5 s (0.596 ± 0.043), and TCN with KP gives the strongest result at +10 s (0.596 ± 0.074). The best +5 s and +10 s AUROC values differ by less than 0.001.

Table 13: Benchmark 3 results for successful topic-transition detection using the primary +10 s post-turn observation horizon. Values are the unweighted mean ± sample standard deviation across five held-out test folds.

Model	Input	AUROC	AUPRC
LR (initiator)	KP	0.556 ± 0.026	0.291 ± 0.029
	ACC	0.546 ± 0.056	0.274 ± 0.041
	KP+ACC	0.568 ± 0.043	0.296 ± 0.049
TCN (initiator)	KP	$\mathbf{0.596 \pm 0.074}$	$\mathbf{0.334 \pm 0.086}$
	ACC	0.483 ± 0.049	0.218 ± 0.029
	KP+ACC	0.565 ± 0.036	0.286 ± 0.058
TCN–DeepSets (group)	KP	0.539 ± 0.044	0.274 ± 0.072
	ACC	0.488 ± 0.058	0.235 ± 0.031
	KP+ACC	0.501 ± 0.091	0.279 ± 0.067

A similar pattern is observed for AUPRC. The highest value is obtained by TCN with KP+ACC at +5 s (0.349 ± 0.063), followed by TCN with KP at +10 s (0.334 ± 0.086). Overall, extending the response horizon alone does not produce a consistent improvement in successful-transition detection.

Table 14: AUROC for successful topic-transition detection using different post-turn observation horizons. Values are the unweighted mean ± sample standard deviation across five held-out test folds.

Model	Input	+2 s	+5 s	+10 s
LR	KP	0.536 ± 0.031	0.545 ± 0.025	0.556 ± 0.026
	ACC	0.516 ± 0.042	0.525 ± 0.052	0.546 ± 0.056
	KP+ACC	0.544 ± 0.034	0.555 ± 0.039	0.568 ± 0.043
TCN	KP	$\mathbf{0.572 \pm 0.061}$	0.549 ± 0.067	$\mathbf{0.596 \pm 0.074}$
	ACC	0.503 ± 0.070	0.488 ± 0.068	0.483 ± 0.049
	KP+ACC	0.570 ± 0.050	$\mathbf{0.596 \pm 0.043}$	0.565 ± 0.036
TCN–DeepSets	KP	0.566 ± 0.050	0.563 ± 0.054	0.539 ± 0.044
	ACC	0.511 ± 0.068	0.500 ± 0.077	0.488 ± 0.058
	KP+ACC	0.550 ± 0.032	0.563 ± 0.067	0.501 ± 0.091

Table 15: AUPRC for successful topic-transition detection using different post-turn observation horizons. Values are the unweighted mean ± sample standard deviation across five held-out test folds. The positive-class prevalence is 0.227.

Model	Input	+2 s	+5 s	+10 s
LR	KP	0.272 ± 0.044	0.277 ± 0.040	0.291 ± 0.029
	ACC	0.264 ± 0.034	0.267 ± 0.042	0.274 ± 0.041
	KP+ACC	0.282 ± 0.037	0.293 ± 0.040	0.296 ± 0.049
TCN	KP	$\mathbf{0.309 \pm 0.032}$	0.275 ± 0.032	$\mathbf{0.334 \pm 0.086}$
	ACC	0.229 ± 0.029	0.242 ± 0.048	0.218 ± 0.029
	KP+ACC	0.284 ± 0.033	$\mathbf{0.349 \pm 0.063}$	0.286 ± 0.058
TCN–DeepSets	KP	0.287 ± 0.063	0.295 ± 0.028	0.274 ± 0.072
	ACC	0.232 ± 0.041	0.252 ± 0.054	0.235 ± 0.031
	KP+ACC	0.276 ± 0.016	0.336 ± 0.097	0.279 ± 0.067

Secondary attempt-detection task As a secondary analysis, we evaluate whether nonverbal behavior can distinguish topic-redirection attempts from ordinary conversation, irrespective of whether the attempt subsequently becomes established. For this task, the 316 successful and 148 failed attempts are combined into a positive class of 464 samples, while the 928 ordinary conversational controls form the negative class. The positive prevalence is therefore 33.3%.

The secondary task uses the same samples, fold assignments, models, modalities, and temporal-window settings as the main benchmark. Tables 16 and 17 report the corresponding AUROC and AUPRC results.

Attempt-detection performance changes only modestly across the three temporal horizons. LR with ACC gives the highest AUROC at all three horizons: 0.557 ± 0.026 at +2 s, 0.568 ± 0.029 at +5 s,

and 0.563 ± 0.025 at +10 s. The highest AUPRC is 0.411 ± 0.054, obtained by LR with KP+ACC at +10 s. These results show little evidence of a systematic improvement as more post-turn behavior is included.

Table 16: AUROC for the secondary topic-redirection attempt-detection task using different post-turn observation horizons. Values are the unweighted mean ± sample standard deviation across five held-out test folds.

Model	Input	+2 s	+5 s	+10 s
LR	KP	0.524 ± 0.073	0.528 ± 0.073	0.534 ± 0.037
	ACC	$\mathbf{0.557 \pm 0.026}$	$\mathbf{0.568 \pm 0.029}$	$\mathbf{0.563 \pm 0.025}$
	KP+ACC	0.547 ± 0.071	0.544 ± 0.057	0.552 ± 0.032
TCN	KP	0.519 ± 0.020	0.528 ± 0.041	0.520 ± 0.027
	ACC	0.504 ± 0.051	0.508 ± 0.082	0.502 ± 0.078
	KP+ACC	0.541 ± 0.031	0.538 ± 0.042	0.549 ± 0.079
TCN–DeepSets	KP	0.498 ± 0.055	0.519 ± 0.047	0.509 ± 0.031
	ACC	0.502 ± 0.064	0.512 ± 0.057	0.541 ± 0.049
	KP+ACC	0.518 ± 0.050	0.506 ± 0.061	0.529 ± 0.057

Table 17: AUPRC for the secondary topic-redirection attempt-detection task using different post-turn observation horizons. Values are the unweighted mean ± sample standard deviation across five held-out test folds. The positive-class prevalence is 0.333.

Model	Input	+2 s	+5 s	+10 s
LR	KP	0.382 ± 0.066	0.383 ± 0.072	0.387 ± 0.056
	ACC	0.384 ± 0.034	$\mathbf{0.401 \pm 0.040}$	0.398 ± 0.034
	KP+ACC	$\mathbf{0.395 \pm 0.065}$	0.399 ± 0.062	$\mathbf{0.411 \pm 0.054}$
TCN	KP	0.357 ± 0.023	0.380 ± 0.046	0.373 ± 0.035
	ACC	0.354 ± 0.029	0.357 ± 0.053	0.342 ± 0.039
	KP+ACC	0.393 ± 0.030	0.383 ± 0.040	0.392 ± 0.047
TCN–DeepSets	KP	0.327 ± 0.026	0.373 ± 0.070	0.356 ± 0.036
	ACC	0.350 ± 0.053	0.354 ± 0.032	0.370 ± 0.032
	KP+ACC	0.366 ± 0.044	0.348 ± 0.048	0.376 ± 0.042

Summary Across the three response horizons, successful topic-transition detection shows modest discrimination, with several configurations performing above the AUPRC class-prevalence baseline. Increasing the post-turn observation horizon from 2 s to 10 s does not lead to a consistent improvement. The strongest successful-transition results are obtained by initiator-level TCN models, while the tested TCN–DeepSets baseline does not provide a consistent improvement from incorporating the surrounding group.

The secondary attempt-detection task is less sensitive to temporal-window length. Its strongest results are obtained by the LR baselines, particularly with acceleration. Overall, the results indicate that nonverbal behavior contains some information about both topic-redirection attempts and successful topic transitions, but the baseline performance remains modest.

Running details and compute resources All Benchmark 3 experiments were executed on the internal Slurm cluster. Neural-network experiments used one GPU, four CPU cores, and 32 GB of system memory per job, with a maximum wall time of 12 hours. Logistic-regression experiments were executed on CPU using four CPU cores and 16 GB of memory, with a maximum wall time of 4 hours.