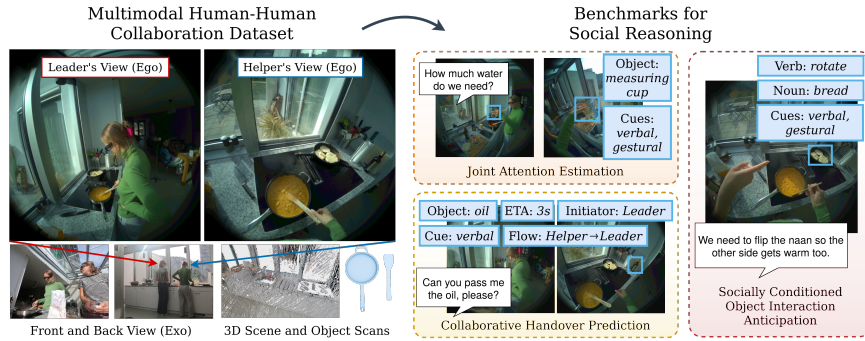


# CoMind: Understanding Collaborative Human Activity from Multiple Minds and Views

Alexey Gavryushin<sup>\*†1</sup> Dingxi Zhang<sup>\*1</sup> Zhao Huang<sup>\*1</sup>  
 Alexandros Delitzas<sup>1,2</sup> Jiaqi Chen<sup>1</sup> Ben Ellis<sup>1</sup> Cedric Zöllner<sup>1</sup>  
 Manthan Patel<sup>1</sup> Manuel Kaufmann<sup>1</sup> Marc Pollefeys<sup>1,3</sup> Xi Wang<sup>1,4,5</sup>

<sup>1</sup>ETH Zurich, Switzerland <sup>2</sup>MPI for Informatics, Germany <sup>3</sup>Microsoft Switzerland

<sup>4</sup>TU Munich, Germany <sup>5</sup>MCML, Germany



**Fig. 1: The CoMind dataset** captures 41 hours of human-human collaboration across 80 sessions, combining egocentric/exocentric videos, audio, gaze, and high-quality 3D scene and object scans, as well as rich annotations for three novel social reasoning tasks.

**Abstract.** Human-human collaboration is a fundamental aspect of everyday life, essential to success in a wide range of goal-directed activities from household tasks to professional teamwork. While much research has focused on modeling coordination and task execution, the cognitive processes that support such collaboration, particularly Theory of Mind (the ability to infer others’ mental states), remain difficult to study in natural settings. To address this gap, we introduce a novel egocentric and exocentric video dataset capturing real-world collaboration in cooking scenarios. The dataset integrates multi-perspective video, high-quality audio, gaze tracking, and 3D scene and object scans, with annotations for shared gaze attention, social cues and interactions between agents, as well as agent-object interactions. We establish benchmarks for Joint Attention Estimation, Action Anticipation, and Collaborative Handover Prediction, enabling research on multimodal perception, proactive assistance, and collaborative planning. By providing temporally aligned, richly annotated multimodal data, CoMind facilitates the development and evaluation of AI systems capable of modeling complex social interactions and reasoning about human behaviors in collaborative environments. Our dataset and benchmarks are made available at <https://comind.ethz.ch/>.

**Keywords:** Egocentric Vision · Theory of Mind · Human-Human Collaboration

<sup>\*</sup> Equal contribution, alphabetical order. <sup>†</sup> Corresponding author.

## 1 Introduction

Human collaboration is central to success in many goal-directed activities, from everyday tasks to professional teamwork. Understanding it is key to building AI systems that can seamlessly work with humans. Effective collaboration requires the ability to infer and adapt to others’ mental states, intentions, and goals, a concept also known as *Theory of Mind* (ToM) [5, 80]. ToM is achieved through the continuous exchange of *social cues* [2], which are multimodal signals (*e.g.*, eye gaze, body pose, and targeted dialogue) that agents use to establish shared mental states, transmit intent, and facilitate social interaction.

However, computational modeling of ToM remains underexplored as previous ToM benchmarks [16, 41, 62, 74, 81, 91] mainly focus on text-based reasoning with multiple-choice prompts. This is due to the lack of datasets capturing the multimodal social cues for human collaboration in natural, real-world settings [18, 71]. While egocentric datasets [9, 19, 32, 33, 37, 38, 56, 73, 83, 90] capture single-agent activity, few focus on *multi-agent collaboration*. Existing datasets with *multiple* humans [34, 39, 40, 43, 85, 90] involve short-horizon tasks, limited interactions, or a minimal recording setup that omits critical social cues such as eye gaze or verbal communication. Other works in simulation [13, 59, 89] lack the authenticity of unscripted collaboration. These limitations hinder learning from and modeling ToM in real-world contexts.

In this work, we present CoMind, a **new egocentric–exocentric multi-modal dataset**, designed to study ToM in collaborative, long-horizon tasks. Our dataset focuses on unscripted cooking tasks, a domain inherently rich in shared intentions and interactions. Our dataset features (1) synchronized egocentric and exocentric video, (2) gaze and hand tracking, (3) audio recordings and transcripts, (4) camera trajectories with spatial mapping, (5) dense 3D scene scans, (6) a set of 3D scans of commonly used objects, and (7) fine-grained social cue annotations.

We additionally design **three tasks** of increasing difficulty to formalize ToM benchmarking. Our first task, *Joint Attention Estimation* tests the ability to establish mutual awareness. The second task, *Socially Conditioned Object Interaction Anticipation* predicts an upcoming action and its social cues, testing the ability to infer intent. Our third task, *Collaborative Handover Prediction* predicts the timing and specifics of physical assistance before any reaching motion begins, testing proactive physical coordination.

We benchmark recent vision-language models across these tasks and reveal a significant performance deficiency, showing that current state-of-the-art methods struggle with social perception. We further validate the utility of CoMind by fine-tuning an open-source baseline model on our training data, which yields a substantial improvement in performance and establishes our dataset as a crucial foundation for socially aware AI systems.

## 2 Related Work

**Ego-Exo Datasets and Benchmarks.** Egocentric vision research has produced several large-scale datasets that capture human activity from first-person



**Table 1: Overview of egocentric datasets.** For Modality, 📹 denotes video, 👁 denotes gaze, 📶 denotes IMU, ☁ denotes semi-dense point cloud, 🌐 denotes dense 3D scene scans, 📦 denotes 3D object scans, 🖐 denotes hand tracking. Hours (Ego) refers to the cumulative duration of distinct egocentric video recordings. Our dataset provides unprecedented depth by offering rich social cues and collaborative annotations alongside multi-view 3D grounding.

| Benchmark                  | Domain             | Modality     | Hours (Ego) | Average Duration | Exo | Collaboration Dynamics | Social Cue Annotation | #Humans per session | Total participants |
|----------------------------|--------------------|--------------|-------------|------------------|-----|------------------------|-----------------------|---------------------|--------------------|
| EPIC-KITCHENS-100 [20]     | Kitchen            | 📹            | 100         | 8.5 min          | ✗   | ✗                      | ✗                     | 1                   | 32                 |
| HD-EPIC [56]               | Kitchen            | 📹 ** 📶 🖐     | 41          | 15.9 min         | ✗   | ✗                      | ✗                     | 1                   | 9                  |
| EPFL-Smart-Kitchen-30 [11] | Kitchen            | 📹 ** 📶 🖐     | 29.7        | 35.9 min         | ✓   | ✗                      | ✗                     | 1                   | 16                 |
| Assembly101 [73]           | Task Execution     | 📹            | 167         | 7.1 min          | ✓   | ✗                      | ✗                     | 1                   | 53                 |
| Ego4D [32]                 | Daily Activities   | 📹 ** 📶 🖐     | 3,670       | 22.8 min         | ✗   | ✗                      | ✓                     | 1                   | 923                |
| EgoExo4D [33]              | Skilled Activities | 📹 ** 📶 🖐     | 221         | 2.6 min          | ✓   | ✓                      | ✗                     | 1-2                 | 740                |
| EgoExoLearn [37]           | Task Execution     | 📹 ** 📶       | 120         | 13.4 min         | ✓   | ✗                      | ✗                     | 1                   | -                  |
| HoloAssist [83]            | Assistive Task     | 📹 ** 📶 🖐     | 166         | 4.8 min          | ✓   | ✓                      | ✗                     | 1-2                 | 222                |
| LEMMA [39]                 | Task Execution     | 📹 📶          | 10          | 2 min            | ✓   | ✓                      | ✗                     | 1-2                 | 8                  |
| EgoLife [86]               | Daily Life         | 📹 ** 📶 🖐     | 266         | 44.3 h           | ✓   | ✓                      | ✗                     | 6                   | 6                  |
| IndEgo [14]                | Industrial Tasks   | 📹 ** 📶 🖐     | 197         | 3.4 min          | ✓   | ✓                      | ✗                     | 1-2                 | 20                 |
| <b>Ours</b>                | Kitchen            | 📹 ** 📶 🖐 🌐 📦 | 81          | 30.5 min         | ✓   | ✓                      | ✓                     | 2                   | 125                |

perspectives, providing valuable multimodal information such as gaze, hand pose, audio, and depth [9, 19, 32, 33, 37, 38, 56, 73, 83, 90]. Large scale datasets like EPIC-KITCHENS-100 [19, 20] and Ego4D [32] have significantly advanced single-agent action understanding through multi-modal annotation. More recent efforts capture ego- and exocentric paired views [33, 37, 39, 61, 73] to bridge first- and third-person perspectives. Ego-Exo4D [33] synchronizes head cameras with external views for skilled physical activities, while HoloAssist [83] captures multi-view instructional collaboration. However, these datasets still predominantly treat the primary actor in isolation, focusing on procedural execution rather than joint coordination. Conversely, socially aware datasets [40, 46, 90] capture multi-person interactions but focus on conversational dynamics and human poses rather than shared physical goals. As detailed in Table 1, existing literature lacks works that capture continuous, goal-driven physical collaboration with detailed annotations for social cues, which are important for modeling intent during collaboration. Our dataset addresses this by integrating long-context multi-agent ego-exo videos with diverse sensor data, rich annotations, and 3D scans of the environment and objects, providing the multimodal data sources necessary to study real-world human-human collaboration from multiple viewpoints.

**Multi-Agent Collaboration.** Beyond single-agent settings, recent works [13, 58, 59, 76, 89] have investigated multi-agent collaboration in simulation, examining how humans and embodied agents coordinate in shared tasks. While these simulators excel at modeling high-level task coordination, they abstract away the sensor noise and lack the rich, multimodal real-world cues essential for natural human teamwork. On the other hand, real-world datasets [14, 39, 40, 43, 85, 86, 90] capture the complexity of physical interaction in shared spaces. LEMMA [39] collects multi-view videos in short-term collaborative activities without verbal communication; HOH [85] focuses on object handover sequences and geometric execution of the process; EgoLife [86] targets daily household activities for training long-term egocentric assistants, and Indego [14] focuses on industrial scenarios and procedural task understanding. While these datasets offer rich sensory information, most center on short-duration or narrowly scoped interactions and

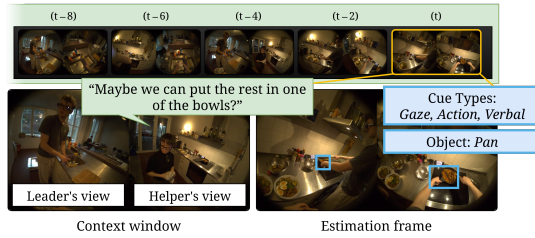
provide limited insight into the dynamic and socially cued interactions in multi-agent collaboration.

**Benchmarking Theory of Mind.** While the importance of Theory of Mind (ToM) in collaboration is well-established [57, 72], its computational evaluation remains a challenge. Recent works benchmark modern AI systems, such as multimodal large language models, for ToM [16, 41, 62, 74, 81, 91]. However, these works typically evaluate for ToM as a passive Video Question Answering task, probing abstract belief systems through text, often using unrealistic multiple-choice prompts. In natural physical collaboration, there are no predefined prompts; agents must continuously infer mental states and respond with real-time physical actions. Therefore, rather than performing post-hoc textual analysis, we target open-ended, complex ToM reasoning. To benchmark for this ToM ability, we formalize Theory of Mind as three interdependent vision tasks.

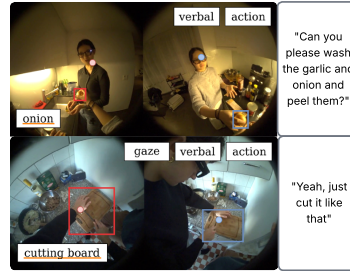
*Joint Attention Estimation.* Gaze behavior is a fundamental cue for social interactions. While gaze estimation has been extensively studied in exocentric [17, 64] and egocentric settings [26, 32, 61], most works predict *single* agent gaze targets. Multi-agent research has addressed shared attention [10, 25] and mutual gaze [44, 47], but typically in third-person videos that lack the synchronized first-person perspective needed for understanding egocentric collaboration. Egocentric works on social gaze [36, 87] explore when another person looks at the camera wearer, and few works explicitly target joint attention in goal-oriented tasks. By providing synchronized multi-agent gaze data, we frame “joint attention” as not just identifying the geometric gaze intersection, but as understanding shared mental attention [65], which is paramount to understanding socially cued actions.

*Object Interaction Anticipation.* Object interaction anticipation has been widely studied in exocentric [1, 15, 30, 66, 82] and egocentric [19, 29, 31, 32, 55, 60, 67] videos. Approaches range from recurrent models [28, 54, 75] and transformer-based methods [31, 32, 45] to multimodal variants that integrate audio or object features [28, 49, 88], and more recently to LLMs [35, 92]. However, most egocentric action anticipation works focus on single-agents. While works like LEMMA [39] incorporate multi-agent scenarios using 3D-CNNs [54] and transformers [84], they generally predict individual actions independently, ignoring inter-agent conditioned behaviors. Our benchmark addresses this critical gap anticipating a *partner’s* future STA variables (verb, noun, and spatial bounding box) strictly conditioned on prior social cues.

*Collaborative Handover Prediction.* The physical transfer of objects is a critical component of collaboration. Previous research studies handovers from a kinematic, grasp synthesis, and safety perspective [27, 78, 79, 88], and predominantly within the context of Human-Robot Interaction [12, 53]. In computer vision, works analyzing handovers [42, 45, 85] focus on segmenting and reconstructing the mechanics of the handover itself (i.e. hand trajectories, grasp phases, contact points). Rather than just detecting the geometric reconstruction of interacting hands, our task serves to understand the cognitive precursors of a “handover” before any reaching motion occurs: who will hand what to whom, when, and prompted by what social cue if at all. We frame “handover prediction” as proactive



**Fig. 2: Joint Attention Estimation Task.** From the audio or transcribed speech in a context window, together with a target frame (green), the model predicts the jointly attended object, its bounding box in the left and right views, and the social cue type (blue).



**Fig. 3: Example annotations.** Frames show joint attention cues, gaze points, and object bounding boxes with labels.

intent inference, and measure a model’s ability to anticipate agents’ needs and provide seamless assistance.

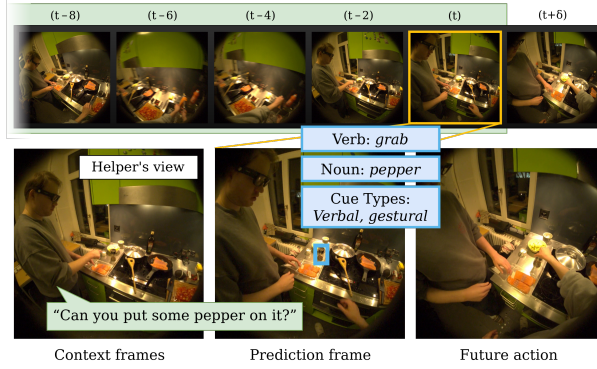
### 3 The CoMind Tasks and Dataset

This section introduces the CoMind dataset and the three tasks we design to formalize ToM benchmarking. We begin by formally defining each task in subsection 3.1, describing the inputs, outputs, and evaluation objectives for Joint Attention Estimation, Socially Conditioned Object Interaction Anticipation, and Collaborative Handover Prediction. We then subsequently describe the data collection protocols, data processing, annotation pipeline, and dataset statistics.

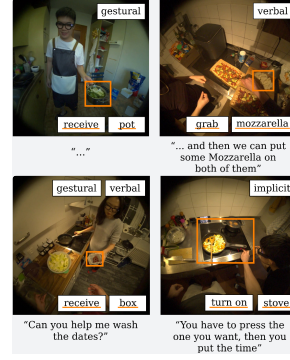
#### 3.1 Proposed Tasks

We propose three novel tasks designed to evaluate a model’s ability to interpret social cues and understand multi-agent coordination. We designate a “leader” and a “helper” for each task, where the leader drives the primary task and the helper must proactively anticipate the leader’s needs by understanding social cues. This directly mirrors proactive multi-agent collaboration. All tasks are grounded in a collaborative cooking scenario, as the kitchen provides a natural yet controlled environment well-suited for studying social coordination.

**Task 1: Joint Attention Estimation.** We define joint attention as a state of temporally and spatially synchronized focus on a shared semantic entity. The Joint Attention Estimation task is shown in Figure 2: For a frame at time  $t$  (with a known “joint attention” event), given the associated context in the form of audio or transcribed speech  $T_{t'-t}$  of the last 10 seconds ( $t' = t - 10s$ ) for both participants  $p \in \{\ell, h\}$  and a final prediction frame  $V_t^{(\ell, h)}$  showing both views, we predict  $(b^{(\ell)}, b^{(h)}, c, q)$ , where  $b^{(p)} = (x^{(p)}, y^{(p)}, w^{(p)}, h^{(p)})$  denotes a view-specific bounding box of the jointly attended object for each participant  $p$ . We further predict an object category  $c$ , as well as the type  $q \subseteq \{\text{action, gaze, verbal, gestural}\}$



**Fig. 4: Socially Conditioned Object Interaction Anticipation Task.** From context frames (or video) and audio or transcribed speech together with a prediction frame (green), the model predicts the noun and verb of the action to be performed, the interacted object’s bounding box, and the social cue type (blue).



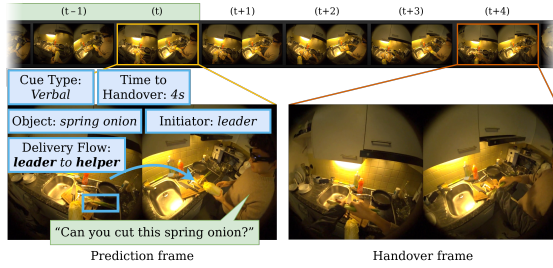
**Fig. 5: Example annotations.** The prediction frame shows the ground-truth action verb, noun, object bounding box, and cue type(s).

of the cue eliciting the joint attention on the object. This task measures the capacity to recognize a collaborative agent’s shared mental state.

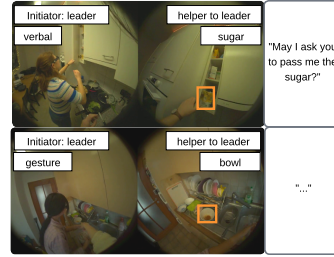
**Task 2: Socially Conditioned Object Interaction Anticipation.** We define a socially conditioned action of participant  $i \in \{\ell, h\}$  as any action performed that is directly triggered by a social cue from the other participant  $j \in \{\ell, h\}, j \neq i$ . In our setting, we fix  $i = h$ , i.e. the action is always cued by the leader and performed by the helper. We fix  $i = h$  to mimic a scenario where a helper agent, i.e., a robot, must predict what is expected of it.

Given the helper’s recent observations as a frame history  $V_{t' \rightarrow t}^{(h)}$  and audio or transcribed speech  $T_{t' \rightarrow t}$  of the past 10 seconds ( $t' = t - 10s$ ) of the collaborative session up to a frame  $t$ , the goal of this task is to predict the socially conditioned action which the helper is expected to perform in the immediate future after  $t$ . We thus aim to predict the next structured action  $a_{t+\delta}^{(h)}$  for  $h$  (the helper) that has been socially conditioned on  $\ell$  (the leader) at a future, unknown offset  $0 < \delta \leq 10$  seconds. The output of the task is  $(v, n, b, q)$ , where we use a verb  $v \in \mathcal{V}$  (e.g. *take*) and noun  $n \in \mathcal{N}$  (e.g. *fork*) describing the action (e.g. *take fork*) that the helper will perform at time  $t + \delta$ . Here,  $\mathcal{V}$  and  $\mathcal{N}$  represent the verb and noun taxonomy of our dataset. Additionally, we predict a bounding box  $b = (x, y, w, h)$  enclosing the object that the helper will interact with, in the 2D coordinates of the frame at  $t + \delta$  of  $h$ ’s video stream. Lastly, we also predict the cue type  $q \in \{\text{gestural}, \text{verbal}, \text{gestural} \wedge \text{verbal}, \text{implicit}\}$  of the social cue which can be used to infer  $v, n, b$ . Figure 5 shows a few examples. This task measures the capacity to anticipate a collaborative agent’s intent, and grasp the social cues preceding the intent.

**Task 3: Collaborative Handover Prediction.** We illustrate the task in Figure 6. Given a prediction frame  $f$  before any reaching or grasping motion towards a handed-over object, together with the two synchronized egocentric



**Fig. 6: Collaborative Handover Prediction Task.** From context frames (or video) and audio or transcribed speech, together with a prediction frame (green), the model predicts the delivery flow (who hands to whom), the object category and bounding box in the view of the handing participant, the initiator, and the cue type (blue).



**Fig. 7: Example annotations.** Frames show handover initiator, cue type, delivery flow, and object bounding boxes with labels.

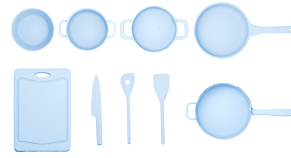
video clips  $V_{t' \rightarrow t}^{(\ell)}$  and  $V_{t' \rightarrow t}^{(h)}$ , captured for 10 seconds before  $t$  ( $t' = t - 10s$ ), as well as the audio or transcribed speech  $T_{t' \rightarrow t}$ , the goal is to anticipate the specifics of the upcoming handover event using  $f$  as a reference. The task is formulated as predicting a single handover event  $(\tau, s \rightarrow r, i, b, c, q)$ , where  $\tau > 0$  denotes how long after  $t$  the handover will occur (also referred to as Time-To-Handover, TTH),  $s \rightarrow r$  indicates the sender-to-receiver object delivery flow with  $s, r \in \{\ell, h\}$  and  $s \neq r$ ,  $i \in \{\ell, h\}$  represents the person initiating the handover through a social cue,  $b$  is a bounding box localizing the object in the sender’s view,  $c$  is the object category, and  $q \subseteq \{\text{verbal, gestural, implicit}\}$  specifies the social cue type(s) used by  $i$  to initiate the handover. This task measures the ability to infer collaborative intent and partner need ahead of physical interaction, reflecting realistic assistive scenarios such as tool use in shared tasks.

### 3.2 Data Recording Protocol and Hardware Setup

Data collection takes place in real-world kitchens to ensure a natural and diverse dataset, and all participants provide written informed consent. Both participants wear the **Meta Aria Glasses** [24] to capture egocentric multi-modal recordings. Prior to each session, we calibrate intrinsics of the two **GoPro cameras** and the Aria glasses are calibrated for each participant for accurate eye-gaze tracking. We place one GoPro camera in front of the cooking area, capturing the participants from the front and providing a clear view of their upper-body and hand movements throughout the recording. A second GoPro is placed behind the cooking area, recording the participants from the back to ensure full-body visibility. The Aria glasses are triggered to start recording synchronously. A trained data collector oversees the recording session and monitors it remotely from another room via a video stream provided by a mobile phone placed in the kitchen space. Moreover, a dense 3D scan of the kitchen is captured by the data collector using a **Leica BLK2GO scanner** prior to the session. Lastly, we reuse a set of common cooking



**Fig. 8: Sensors used to create the CoMind dataset.** Two Aria glasses record the scene from each participant’s view. Two GoPro Hero cameras capture exocentric views. The Leica BLK2Go and Artec 3D Leo are used for 3D scene and object scanning respectively.



**Fig. 9: 3D scans of objects used in the recordings.** Participants work with objects that have been 3D-scanned.

utensils (pans and pots) for each session, which are scanned using an **Artec 3D Leo** [6] scanner to obtain high-fidelity object meshes. We provide a visualization of the sensors and hardware in Figure 8.

### 3.3 Data processing

**Synchronization.** First, the Aria recordings are temporally aligned to sub-millisecond precision using TICSync [48]. Second, the GoPro recordings are time-synced to the Arias with SyncSink [77] using audio cross-correlation. We also manually trim recordings to the same starting timestamp, to eliminate irrelevant frames before the start of the actual cooking.

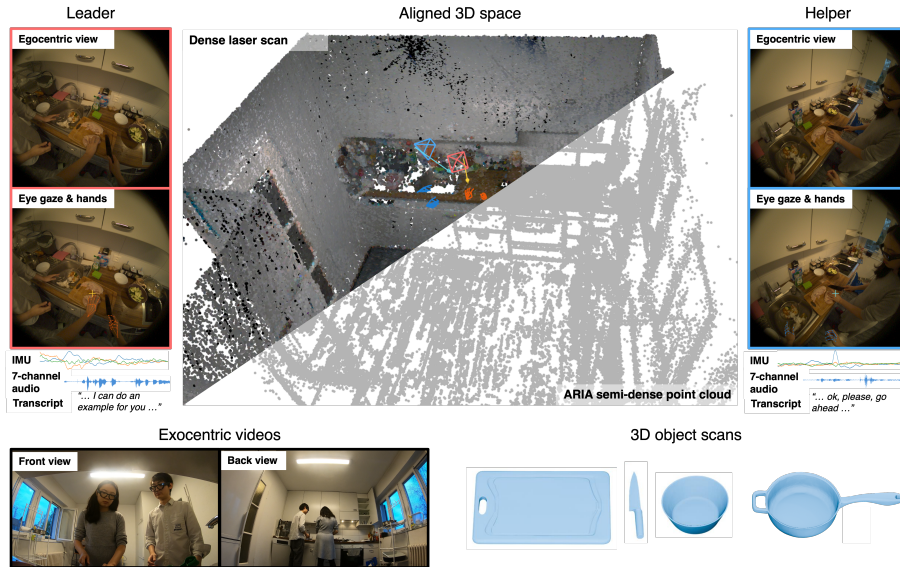
**Post Processing.** We process all Aria recordings using the Machine Perception Services (MPS) [48] to obtain high-fidelity spatial and behavioral data. This pipeline provides 6DoF camera trajectories, semi-dense point clouds, eye-gaze, and hand tracking. Specifically, the Multi-SLAM output from MPS is used to align the data from both participants into a shared coordinate frame, enabling consistent spatial reasoning across each participant pair. Additionally, we extract audio transcripts of the narrations and dialogue using the WhisperX speech recognition model [8, 63]. All modalities available in our dataset are visualized in Figure 10.

**Scan Alignment.** To transform the 3D scan of the environment into Aria’s world frame, we follow a two-stage procedure. In the first stage, we localize the BLK2GO images within Aria’s semi-dense point cloud using Hierarchical Localization [69, 70]. Each successfully localized image yields an independent estimate of the rigid transformation between the two coordinate frames. We reject outlier estimates and compute a robust average over the inlier set to obtain an initial alignment. In the second stage, we refine this estimate using point-to-plane ICP [68] between the BLK2GO scan and the Aria semi-dense point cloud. All alignments are manually verified.

### 3.4 Data Annotation

In addition to the raw captured data, we manually annotate our recordings to provide detailed ground-truth for the three tasks defined above. We use a mix of in-house and external professional human annotators. All annotators follow instructions and tutorial videos designed to ensure consistency and resolve





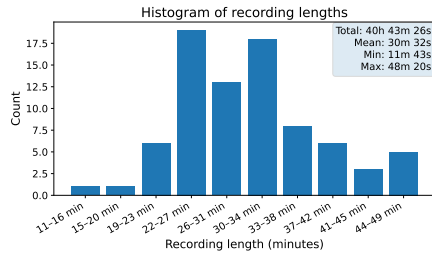
**Fig. 10: Modalities in the CoMind dataset.** The dataset provides synchronized multimodal recordings of collaborative sessions, including egocentric video with eye gaze and hand tracking, audio and IMU signals, exocentric camera views, camera trajectories and spatial mapping, dense 3D scans of the scene, and a set of high-quality 3D scans of commonly used objects.

ambiguity. External annotators pass their work through QA teams, and all annotations are further reviewed by at least one main contributor of our work. Annotations are performed using custom interfaces, built upon the VIA Video Annotator [23], designed for each of the tasks. See Supp. Mat. for more details.

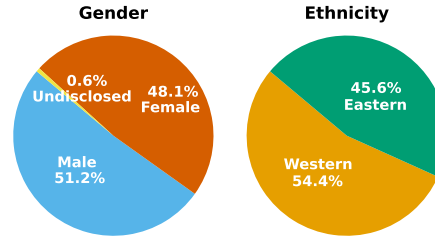
### 3.5 Data Statistics

In total, our dataset contains 80 recordings from 55 kitchen spaces with 125 distinct participants. We accumulate 81h 26m of dual-view, or 40h 43m of single-view data. We provide a statistical analysis of our dataset, showing its diversity and size in terms of participants, recordings as well as the annotation taxonomy created for our considered tasks.

**Participants.** Our dataset features 125 unique participants organized into 80 collaborative pairs. The participants have varying degrees of cooking expertise, and represent diverse nationalities and ethnic backgrounds. A visualization of the demographics is provided in Figure 12. Our dataset exhibits an approximately equal distribution of 48.1% female-identifying participants compared to 51.2% male-identifying and 0.6% participants with undisclosed gender, whereas other egocentric datasets such as Ego4D [32] and EgoExo4D [33] report merely 45% resp. 37% of participants identifying as female. Furthermore, we achieve a good balance of 54.4% participants with Western and 45.6% with Eastern ethnicity.



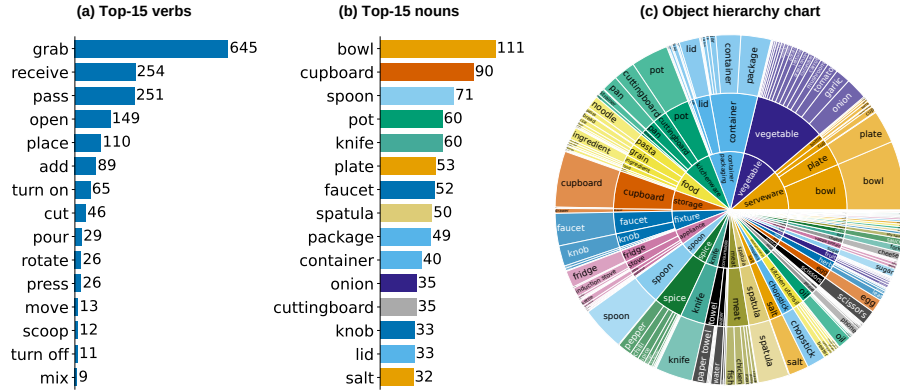
**Fig. 11: Histogram of recording lengths.** Our lengthy recordings allow for the study of human-human collaborative work in long-context settings. In total, we record 41 hours of two-person resp. 81 hours of single-view data.



**Fig. 12: Participant gender and ethnicity distribution.** We balance participants from Eastern and Western backgrounds, and maintain an equal gender distribution of our participants.

The participants’ age at the time of recording ranges from 18 to 38, with most of them in their early to mid-20s.

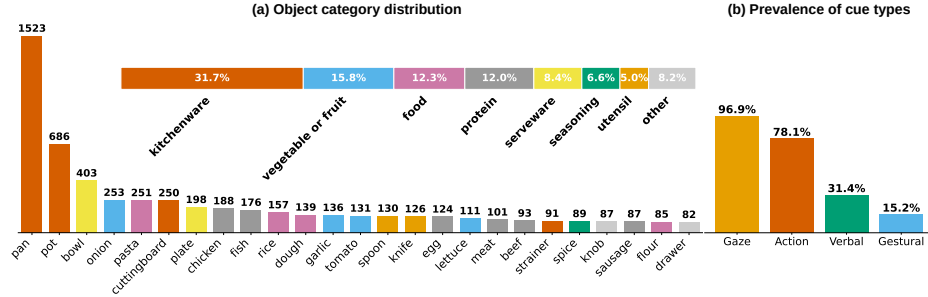
**Recordings.** As shown in Figure 11, our recordings are on average 30 minutes long, allowing the study of long-horizon social interactions and making our dataset suited for future long-horizon benchmarks such as Moment Queries and Long-Term Action Anticipation from Ego4D [32].



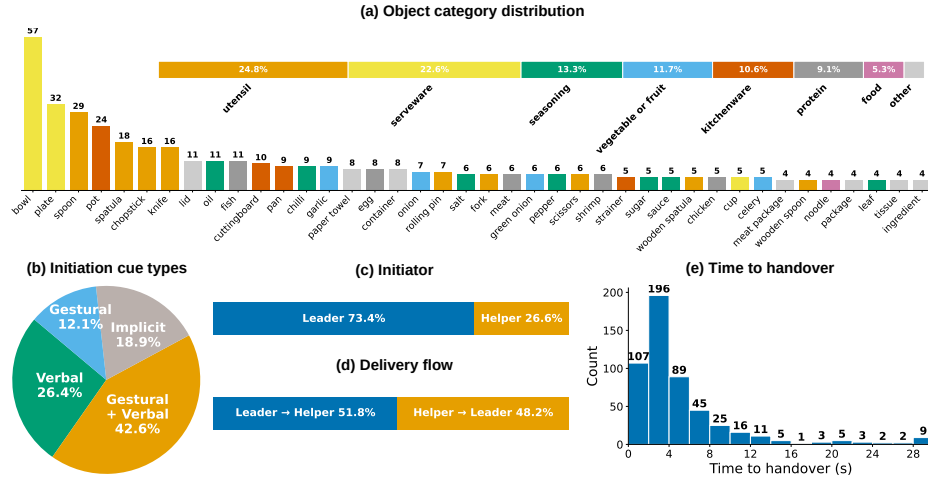
**Fig. 13: Action verb and action noun distribution of the Socially Conditioned Object Interaction Anticipation task.** Our annotations feature a wide range of objects from 13 coarse (L3) and over 180 fine (L1) object categories, as well as 22 unique verbs.

**Annotations.** For all tasks, we create a three-level hierarchy of object categories: L1 (finest) to L3 (coarsest), to account for different possible granularities when describing an object, e.g. noodle (L1) vs. pasta (L2) vs. food (L3). We identify 281, 76 and 14 object categories on levels L1, L2 and L3, respectively. Details are provided in Supp. Mat.





**Fig. 14: Annotation statistics for the Joint Attention Estimation task.** The CoMind dataset allows investigating multiple indicators of joint attention between humans on a broad variety of objects.



**Fig. 15: Annotation statistics for the Collaborative Handover Prediction task.** In addition to the category and bounding box of the handed-over object, we provide annotations about the Time-To-Handover (TTH), delivery flow direction, initiator and initiation cue for each handover event.

For the Socially Conditioned Object Interaction Anticipation task, we visualize this hierarchy, together with the top 15 most frequent verbs and nouns, in Figure 13. We further provide statistics regarding the object categories for the Joint Attention Detection task in Figure 14. Statistics related to the Collaborative Handover Prediction task are illustrated in Figure 15.

## 4 Baselines and Experiments

We benchmark numerous state-of-the-art closed- and open-source vision-language models (VLMs) on our three tasks, establishing a baseline for future research on human-human collaboration. Specifically, we benchmark Anthropic’s Claude Opus (v4.5, v4.6) [3, 4], Google’s Gemini 3 Flash [22], OpenAI’s GPT-4o [50] and GPT-5 families (v5.1, v5.2) [51, 52], and the open-weight Gemma 4 31B

**Table 2: Quantitative results on Joint Attention Estimation.** **Best**, **second-best**, and **third-best** results are highlighted. L-IoU@0.5 and R-IoU@0.5 denote the fraction of predicted bounding boxes with IoU  $> 0.5$  with the ground-truth jointly attended object in the left and right egocentric streams, respectively. Cue Type reports cue classification accuracy, and Cat. L1 measures the fraction of correct L1 object category predictions. Models marked FT are fine-tuned on CoMind.

| Model                    | L-IoU@0.5 $\uparrow$ | R-IoU@0.5 $\uparrow$ | Cue Type $\uparrow$ | Cat. (L1) $\uparrow$ |
|--------------------------|----------------------|----------------------|---------------------|----------------------|
| Random Sampling          | 0.1772               | 0.1337               | 0.3163              | 0.0533               |
| Most Frequent            | 0.0004               | 0.0004               | 0.4647              | 0.2067               |
| Claude Opus 4.5          | 0.0015               | 0.0025               | 0.2419              | 0.3433               |
| Claude Opus 4.6          | 0.0051               | 0.0015               | 0.3439              | 0.3487               |
| Gemini 3 Flash           | <b>0.2164</b>        | <b>0.2071</b>        | 0.3685              | 0.4206               |
| GPT 4o                   | 0.0026               | 0.0003               | 0.0870              | 0.3451               |
| GPT 5.1                  | 0.0042               | 0.0017               | 0.1133              | 0.4281               |
| GPT 5.2                  | 0.0044               | 0.0017               | 0.1474              | <b>0.4285</b>        |
| Gemma 4 31B IT           | 0.1295               | 0.1174               | 0.4628              | 0.3590               |
| Qwen3-VL 8B Instruct     | 0.0000               | 0.0012               | 0.0074              | 0.2701               |
| Qwen3-VL 8B Instruct FT  | 0.1483               | 0.1345               | <b>0.4775</b>       | 0.3007               |
| Qwen3-VL 32B Instruct    | 0.0001               | 0.0012               | 0.0740              | 0.3494               |
| Qwen3-VL 32B Instruct FT | 0.1562               | 0.1496               | 0.4572              | 0.3323               |

IT [21] and Qwen3-VL series (8B, 32B) [7]. All models receive task-specific prompts instructing them to output predictions as structured JSON objects for standardized parsing. We partition the dataset into training and testing splits with disjunct participant sets, selecting 25 videos to provide test set annotations for all tasks. For visual processing, 5 video frames are uniformly sampled from the 10 seconds preceding the prediction frame, preserving fine-grained social cues while respecting VLM context limits. Transcribed speech of the last 10 seconds is also provided. We further finetune our open-weight Qwen3-VL baselines (denoted with “FT” in the Tables 2, 3, 4) and show substantial improvements, validating the effectiveness of our dataset for training socially aware models. Training details, more comparisons as well as visual results are provided in Supp. Mat. Lastly, we add a Random Sampling (RS) baseline which samples values according to the training split’s distribution, and a Most Frequent (MF) baseline which picks the most frequent values, reflecting natural label priors. RS predicts bounding boxes according to the distribution’s mean center point, width and height. MF uses the mean width and height and the image’s center as the center point of the boxes.

#### 4.1 Joint Attention Estimation

**Metrics.** We separately report the fraction of IoUs over 0.5 for the objects in the view of the left (L-IoU@0.5) as well as the right (R-IoU@0.5) person. We also provide the fraction of prediction-GT matches for the cue type. Lastly, we measure the matches of the object category with the ground-truth L1 category.

**Table 3: Quantitative results on Socially Conditioned Object Interaction Anticipation.** **Best**, **second-best**, and **third-best** results are highlighted. IoU@0.5 denotes the fraction of predicted bounding boxes with IoU > 0.5 with the ground-truth object to be interacted with. Cue Type reports cue classification accuracy, and Act. Verb resp. Act. Noun (L1) measures the fraction of correct verb resp. L1 object category predictions. Models marked FT are fine-tuned on CoMind.

| Model                    | IoU@0.5 $\uparrow$ | Cue Type $\uparrow$ | Act. Verb $\uparrow$ | Act. Noun (L1) $\uparrow$ |
|--------------------------|--------------------|---------------------|----------------------|---------------------------|
| Random Sampling          | 0.0152             | 0.3801              | 0.1854               | 0.0411                    |
| Most Frequent            | 0.0005             | 0.4040              | 0.3430               | 0.0662                    |
| Claude Opus 4.5          | 0.0034             | <b>0.5302</b>       | 0.0819               | 0.3490                    |
| Claude Opus 4.6          | 0.0091             | <b>0.5347</b>       | 0.0987               | 0.3507                    |
| Gemini 3 Flash           | <b>0.1859</b>      | 0.5298              | 0.0781               | 0.3775                    |
| GPT 4o                   | 0.0009             | 0.4358              | 0.1046               | 0.3351                    |
| GPT 5.1                  | 0.0202             | 0.4556              | 0.1325               | <b>0.4119</b>             |
| GPT 5.2                  | 0.0038             | 0.5113              | 0.1060               | <b>0.4371</b>             |
| Gemma 4 31B IT           | 0.1468             | 0.4173              | 0.0679               | 0.3225                    |
| Qwen3-VL 8B Instruct     | 0.0909             | 0.1151              | 0.0873               | 0.2288                    |
| Qwen3-VL 8B Instruct FT  | <b>0.2425</b>      | 0.4299              | <b>0.6429</b>        | 0.4021                    |
| Qwen3-VL 32B Instruct    | 0.0035             | 0.4901              | <b>0.1380</b>        | 0.3509                    |
| Qwen3-VL 32B Instruct FT | <b>0.2513</b>      | <b>0.5503</b>       | <b>0.6508</b>        | <b>0.4312</b>             |

Object categories are organized into three levels of granularity: L1 represents the most detailed set of categories, while L2 and L3 progressively merge them into more general groups, with L3 providing the coarsest categorization.

**Results.** As shown in Table 2, zero-shot spatial localization for joint attention remains exceptionally challenging. Most proprietary VLMs (e.g., Claude, GPT families) fail almost entirely at bounding box generation, scoring near zero. Gemini 3 Flash is the notable exception, achieving the highest performance in spatial grounding (0.2164 resp. 0.2071 for L-IoU and R-IoU). Importantly, fine-tuning the Qwen3-VL models on CoMind yields substantial improvements. The fine-tuned 32B variant jumps from near-zero spatial scores to the second-best spatial performance on both L-IoU and R-IoU, while the finetuned 8B variant secures the highest accuracy for identifying the social Cue Type (0.4775). Finally, the clustered Category scores (0.27 to 0.43) across all models suggest that while current VLMs can semantically infer the target object from context, they severely lack the embodied visual grounding required to localize it across synchronous egocentric views.

## 4.2 Socially Conditioned Object Interaction Anticipation

**Metrics.** IoU@0.5 and Cue Type are calculated as in the Joint Attention Detection task. The fraction of matches with the ground-truth verb and L1 noun. As in the previous task, object categories follow the three-level hierarchy (L1–L3), with evaluation here performed at the most fine-grained level (L1).

**Table 4: Quantitative results on Collaborative Handover Prediction.** **Best**, **second-best**, and **third-best** results are highlighted. IoU@0.5 denotes the fraction of predicted bounding boxes with IoU > 0.5 with the ground-truth object. Del. Flow, Initiator, and Init. Type report the fraction of correct predictions for the delivery flow, the initiating participant, and the initiation cue type, respectively. Cat. (L1) measures the fraction of correct L1 object category predictions. TTH reports the fraction of time-to-handover predictions within 0.25 s of the ground-truth. Models marked FT are fine-tuned on CoMind.

| Model                    | IoU@0.5 ↑     | Del. Flow ↑   | Initiator ↑   | Init. Type ↑  | Cat. (L1) ↑   | TTH ↑         |
|--------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Random Sampling          | 0.0000        | 0.4476        | 0.6810        | 0.2905        | 0.0429        | 0.0714        |
| Most Frequent            | 0.0000        | 0.4857        | 0.7857        | 0.4667        | 0.1048        | 0.0524        |
| Claude Opus 4.5          | 0.0000        | 0.5659        | 0.6195        | 0.3073        | 0.3610        | 0.0976        |
| Claude Opus 4.6          | 0.0025        | 0.6146        | 0.6488        | 0.2976        | 0.3610        | <b>0.1317</b> |
| Gemini 3 Flash           | <b>0.0313</b> | <b>0.6524</b> | 0.5810        | 0.3476        | 0.3524        | 0.0667        |
| GPT 4o                   | 0.0008        | 0.5433        | 0.5240        | 0.2644        | 0.3317        | 0.0962        |
| GPT 5.1                  | 0.0000        | 0.5857        | 0.5952        | 0.2333        | <b>0.4000</b> | 0.0810        |
| GPT 5.2                  | 0.0000        | 0.5952        | 0.5667        | 0.2905        | 0.3714        | 0.0810        |
| Gemma 4 31B IT           | 0.0130        | 0.5050        | 0.4100        | 0.2800        | 0.3400        | 0.0750        |
| Qwen3-VL 8B Instruct     | 0.0000        | 0.5381        | 0.4667        | 0.1286        | 0.1905        | 0.0524        |
| Qwen3-VL 8B Instruct FT  | 0.0025        | 0.5714        | 0.6667        | 0.1429        | 0.1952        | 0.1143        |
| Qwen3-VL 32B Instruct    | 0.0000        | 0.5381        | 0.6143        | 0.2952        | 0.3095        | 0.0381        |
| Qwen3-VL 32B Instruct FT | 0.0130        | 0.6143        | <b>0.7286</b> | <b>0.3714</b> | 0.3619        | 0.0905        |

**Results.** The results in Table 3 indicate that predicting future interactions is significantly more complex than identifying current attention. Among the proprietary VLMs, Gemini 3 Flash maintains its lead in spatial grounding (0.1859 IoU), while Claude Opus 4.6 and GPT 5.2 excel in Cue Type and Noun categorization, respectively. Notably, fine-tuning Qwen3-VL on CoMind allows the 32B variant to take the lead in spatial grounding with 0.2513 IoU, followed by the 8B variant scoring second at 0.2425. The finetuned 32B model further scores highest among all models on the Cue Type and Act. Verb accuracy with 0.5503 and 0.6508, respectively. Notably, fine-tuning the Qwen3-VL models on CoMind results in a seven-fold and five-fold increase in their Action Verb accuracy, from 0.0873 and 0.1380 to 0.6429 and 0.6508. This substantial jump suggests that while zero-shot VLMs can recognize static objects and high-level social signals, domain-specific training is essential to accurately model the temporal action dependencies triggered by collaborative social cues.

### 4.3 Collaborative Handover Prediction

**Metrics.** We report IOU@0.5, the fraction of predicted bounding boxes with over 50% IoU with the ground-truth bounding box. Moreover, we calculate the dataset-wide fraction of matches (accuracy) for the delivering flow, initiator, and initiation type predictions. For the object category (Cat.), we measure the fraction of matches of the predicted category with the ground-truth L1 category, counting synonyms for the GT of the same granularity as correct predictions. Object categories follow the three-level hierarchy (L1–L3), with evaluation performed

at the most fine-grained level (L1). Finally, we report the fraction of time-to-handover predictions falling within 0.25s of the ground-truth TTH.

**Results.** Handover prediction emerges as the most difficult task for spatial grounding, with Gemini 3 Flash leading at a mere 0.0313 IoU. While GPT-5.1 shows strength in semantic categorization (0.4000) and Claude 4.6 leads in temporal accuracy (0.1317 TTH), these proprietary models are noticeably less sensitive to social dynamics. In contrast, fine-tuning Qwen3-VL models on CoMind yields the best results for identifying the Initiator (0.7286) and Initiation Type (0.3714). This performance split suggests that while general VLMs can leverage broad heuristics for “what” and, to a lesser degree, for “when,” domain-specific training on CoMind is critical for teaching models the “who” and “how” of proactive collaboration. The across-the-board low scores highlight that seamless, human-like handover remains a challenging frontier for modern AI.

## 5 Conclusion

We introduced CoMind, a dataset of ego-exocentric recordings collected for the study of human-human interaction in collaborative settings. The dataset features a large set of complementary modalities, involving both egocentric and environment data as well as static and dynamic data. We established benchmarks on three new tasks related to human collaboration, and manually collected rich and diverse ground-truth annotations for them in our dataset. We performed an evaluation of numerous vision-language models on our tasks, showing that the current state-of-the-art leaves much room for improvement. We further demonstrated the practical use of our dataset by finetuning an open-source model on its training set, leading to a boost in test-set performance.

## Acknowledgements

This work was supported by the Swiss National Science Foundation Advanced Grant 216260: Beyond Frozen Worlds: Capturing Functional 3D Digital Twins from the Real World and by European Union’s Horizon Europe research and innovation programme under grant agreement number 101214398 (ELLIOT), fully funded by Swiss State Secretariat for Education, Research and Innovation (SERI). Alexandros Delitzas is also supported by the Max Planck ETH Center for Learning Systems (CLS).

## References

1. Abu Farha, Y., Richard, A., Gall, J.: When will you do what?-anticipating temporal occurrences of activities. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5343–5352 (2018) 4
2. Adams Jr, R.B., Albohn, D.N., Kveraga, K.: Social vision: Applying a social-functional approach to face and expression perception. Current directions in psychological science **26**(3), 243–248 (2017) 2

3. Anthropic: Claude opus 4.5 system card. System card (Nov 2025), <https://www.anthropic.com/claude-opus-4-5-system-card>, accessed 2026-03-04 11
4. Anthropic: Introducing claude opus 4.6. Anthropic News (Feb 2026), <https://www.anthropic.com/news/claude-opus-4-6>, accessed 2026-03-04 11
5. Apperly, I.: Mindreaders: the cognitive basis of "theory of mind". Psychology Press (2010) 2
6. Artec 3D: Artec leo: Wireless professional 3d scanner (2024), <https://www.artec3d.com/portable-3d-scanners/artec-leo>, accessed: 2026-03-04 8
7. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 12
8. Bain, M., Huh, J., Han, T., Zisserman, A.: Whisperx: Time-accurate speech transcription of long-form audio (2023), <https://arxiv.org/abs/2303.00747> 8
9. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Han, S., Zhang, F., Zhang, L., Fountain, J., Miller, E., Basol, S., et al.: Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. arXiv preprint arXiv:2411.19167 (2024) 2, 3
10. de Belen, R.A., Mohammadi, G., Sowmya, A.: Temporal understanding of gaze communication with gazetransformer. In: NeurIPS Workshop on Gaze Meets ML (2023) 4
11. Bonnetto, A., Qi, H., Leong, F., Tashkovska, M., Rad, M., Shokur, S., Hummel, F., Micera, S., Pollefeys, M., Mathis, A.: Epfl-smart-kitchen-30: Densely annotated cooking dataset with 3d kinematics to challenge video and language models. arXiv preprint arXiv:2506.01608 (2025) 3
12. Cakmak, M., Srinivasa, S.S., Lee, M.K., Forlizzi, J., Kiesler, S.: Human preferences for robot-human hand-over configurations. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1986–1993. IEEE (2011) 4
13. Chang, M., Chhablani, G., Clegg, A., Cote, M.D., Desai, R., Hlavac, M., Karashchuk, V., Krantz, J., Mottaghi, R., Parashar, P., et al.: Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. arXiv preprint arXiv:2411.00081 (2024) 2, 3
14. Chavan, V., Imgrund, Y., Dao, T., Bai, S., Wang, B., Lu, Z., Heimann, O., Krüger, J.: Indego: A dataset of industrial scenarios and collaborative work for egocentric assistants. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2025) 3
15. Chen, J., Mittal, G., Yu, Y., Kong, Y., Chen, M.: Gatehub: Gated history unit with background suppression for online action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19925–19934 (2022) 4
16. Chen, Z., Wu, J., Zhou, J., Wen, B., Bi, G., Jiang, G., Cao, Y., Hu, M., Lai, Y., Xiong, Z., et al.: Tombench: Benchmarking theory of mind in large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15959–15983 (2024) 2, 4
17. Chong, E., Wang, Y., Ruiz, N., Rehg, J.M.: Detecting attended visual targets in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5396–5406 (2020) 4
18. Csibra, G., Gergely, G.: Natural pedagogy. Trends in Cognitive Sciences **13**(4), 148–153 (2009) 2
19. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 720–736 (2018) 2, 3, 4



20. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)* **130**, 33–55 (2022), <https://doi.org/10.1007/s11263-021-01531-2> 3
21. DeepMind, G.: Gemma 4. <https://deepmind.google/models/gemma/gemma-4/> (2026), accessed: July 2026 12
22. Doshi, T., the Gemini Team: Gemini 3 flash: Frontier intelligence built for speed. <https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/> (Dec 2025), google Blog 11
23. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: *Proceedings of the 27th ACM International Conference on Multimedia*. MM '19, ACM, New York, NY, USA (2019) 9
24. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561* (2023) 7
25. Fan, L., Wang, W., Huang, S., Tang, X., Zhu, S.C.: Understanding human gaze communication by spatio-temporal graph reasoning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 5724–5733 (2019) 4
26. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. pp. 314–327. Springer (2012) 4
27. Fu, R., Zhang, D., Jiang, A., Fu, W., Funk, A., Ritchie, D., Sridhar, S.: Gigahands: A massive annotated dataset of bimanual hand activities. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 17461–17474 (2025) 4
28. Furnari, A., Farinella, G.M.: Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence* **43**(11), 4021–4036 (2020) 4
29. Furnari, A., Farinella, G.M.: Towards streaming egocentric action anticipation. In: *2022 26th International Conference on Pattern Recognition (ICPR)*. pp. 1250–1257. IEEE (2022) 4
30. Gao, J., Yang, Z., Nevatia, R.: Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818* (2017) 4
31. Girdhar, R., Grauman, K.: Anticipative video transformer. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 13505–13515 (2021) 4
32. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18995–19012 (2022) 2, 3, 4, 9, 10
33. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19383–19400 (2024) 2, 3, 9
34. Guo, Z., Hou, Y., Wang, P., Gao, Z., Xu, M., Li, W.: Ft-hid: a large-scale rgb-d dataset for first-and third-person human interaction analysis. *Neural Computing and Applications* **35**(2), 2007–2024 (2023) 2
35. Huang, D., Hilliges, O., Van Gool, L., Wang, X.: Palm: Predicting actions through language models@ ego4d long-term action anticipation challenge 2023. *arXiv preprint arXiv:2306.16545* (2023) 4

36. Huang, Y., Cai, M., Sato, Y.: An ego-vision system for discovering human joint attention. *IEEE Transactions on Human-Machine Systems* **50**(4), 306–316 (2020) 4
37. Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Lu, D., Wang, Y., Wang, L., Qiao, Y.: Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 2, 3
38. Jang, Y., Sullivan, B., Ludwig, C., Gilchrist, I., Damen, D., Mayol-Cuevas, W.: Epic-tent: An egocentric video dataset for camping tent assembly. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. pp. 0–0 (2019) 2, 3
39. Jia, B., Chen, Y., Huang, S., Zhu, Y., Zhu, S.C.: Lemma: A multi-view dataset for learning multi-agent multi-task activities. In: *European Conference on Computer Vision*. pp. 767–786. Springer (2020) 2, 3, 4
40. Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., Kitani, K.: Egohumans: An egocentric 3d multi-human benchmark. *arXiv preprint arXiv:2305.16487* (2023) 2, 3
41. Li, Y., Veerabadran, V., Iuzzolino, M.L., Roads, B.D., Celikyilmaz, A., Ridgeway, K.: Egotom: Benchmarking theory of mind reasoning from egocentric videos. *arXiv preprint arXiv:2503.22152* (2025) 2, 4
42. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3282–3292 (2022) 4
43. Liu, Y., Zhang, C., Xing, R., Tang, B., Yang, B., Yi, L.: Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. *arXiv preprint arXiv:2406.19353* (2024) 2, 3
44. Marin-Jimenez, M.J., Zisserman, A., Eichner, M., Ferrari, V.: Detecting people looking at each other in videos. *International Journal of Computer Vision (IJCV)* **106**, 282–296 (2014) 4
45. Mascaró, E.V., Ahn, H., Lee, D.: Intention-conditioned long-term human egocentric action anticipation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 6048–6057 (2023) 4
46. McLean, C., Meendering, M., Swartz, T., Gabbay, O., Olsen, A., Jacobs, R., Rosen, N., de Bree, P., Garcia, T., Merrill, G., Sandakly, J., Buffalini, J., Jain, N., Krenn, S., Kumar, M., Markovic, D., Ng, E., Prada, F., Saba, A., Zhang, S., Agrawal, V., Godisart, T., Richard, A., Zollhoefer, M.: Embod3d: A large-scale multimodal motion and behavior dataset. Technical report, *arXiv* (2025), <https://arxiv.org/pdf/2510.16258>, *arXiv preprint* 3
47. Medina, M.J.M.K.V.: Suárez p zisserman laeo-net++: revisiting people looking at each other in videos. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)* **10** (2021) 4
48. Meta Reality Labs Research: Project aria machine perception services (mps). [https://facebookresearch.github.io/projectaria\\_tools/docs/ARK/mps](https://facebookresearch.github.io/projectaria_tools/docs/ARK/mps) (2024), accessed: 2026-03-04 8
49. Mittal, H., Morgado, P., Jain, U., Gupta, A.: Learning state-aware visual representations from audible interactions. *Advances in Neural Information Processing Systems* **35**, 23765–23779 (2022) 4
50. OpenAI: Gpt-4o system card. System card (Aug 2024), <https://openai.com/index/gpt-4o-system-card/>, accessed 2026-03-04 11

51. OpenAI: Gpt-5.1 instant and gpt-5.1 thinking system card addendum. System card addendum (Nov 2025), <https://openai.com/index/gpt-5-system-card-addendum-gpt-5-1/>, accessed 2026-03-04 11
52. OpenAI: Introducing gpt-5.2. OpenAI Index (Dec 2025), <https://openai.com/index/introducing-gpt-5-2/>, accessed 2026-03-04 11
53. Ortenzi, V., Cosgun, A., Pardi, T., Chan, W.P., Croft, E., Kulić, D.: Object handovers: a review for robotics. *IEEE Transactions on Robotics* **37**(6), 1855–1873 (2021) 4
54. Osman, N., Camporese, G., Coscia, P., Ballan, L.: Slowfast rolling-unrolling lstms for action anticipation in egocentric videos. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3437–3445 (2021) 4
55. Pasca, R., Gavryushin, A., Hamza, M., Kuo, Y.L., Mo, K., Van Gool, L., Hilliges, O., Wang, X.: Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction anticipation. In: *Conference on Computer Vision and Pattern Recognition 2024* (2024), <https://eth-ait.github.io/transfusion-proj/> 4
56. Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K.K., Liu, K., Gatti, P., Bansal, S., Flanagan, K., et al.: Hd-epic: A highly-detailed egocentric video dataset. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 23901–23913 (2025) 2, 3
57. Premack, D., Woodruff, G.: Does a chimpanzee have a theory of mind. *Behavioral and Brain Sciences* **1**, 515 – 526 (12 1978) 4
58. Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., Torralba, A.: Virtualhome: Simulating household activities via programs (2018) 3
59. Puig, X., Shu, T., Li, S., Wang, Z., Liao, Y.H., Tenenbaum, J.B., Fidler, S., Torralba, A.: Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890* (2020) 2, 3
60. Qi, Z., Wang, S., Su, C., Su, L., Huang, Q., Tian, Q.: Self-regulated learning for egocentric video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) 4
61. Qiu, H., Shi, Z., Wang, L., Xiong, H., Li, X., Li, H.: Egame: A new dataset and challenge for following me via egocentric view in real world. *arXiv preprint arXiv:2501.19061* (2025) 3, 4
62. Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S.A., Botvinick, M.: Machine theory of mind. In: *International conference on machine learning*. pp. 4218–4227. PMLR (2018) 2, 4
63. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: *International conference on machine learning*. pp. 28492–28518. PMLR (2023) 8
64. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: *NeurIPS* (2015) 4
65. Reddy, V.: Understanding attention to self. *Joint attention: Communication and other minds: Issues in philosophy and psychology* p. 85 (2005) 4
66. Rizve, M.N., Mittal, G., Yu, Y., Hall, M., Sajeev, S., Shah, M., Chen, M.: Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22992–23002 (2023) 4
67. Rodin, I., Furnari, A., Mavroeidis, D., Farinella, G.M.: Untrimmed action anticipation. In: *International Conference on Image Analysis and Processing*. pp. 337–348. Springer (2022) 4

68. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings third international conference on 3-D digital imaging and modeling. pp. 145–152. IEEE (2001) 8
69. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: CVPR (2019) 8
70. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020) 8
71. Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., Vogeley, K.: Toward a second-person neuroscience. *Behavioral and Brain Sciences* **36**(4), 393–414 (2013) 2
72. Sebanz, N., Bekkering, H., Knoblich, G.: Joint action: bodies and minds moving together. *Trends in Cognitive Sciences* **10**(2), 70–76 (2006) 4
73. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022) 2, 3
74. Shi, H., Ye, S., Fang, X., Jin, C., Isik, L., Kuo, Y.L., Shu, T.: Muma-tom: Multi-modal multi-agent theory of mind. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025) 2, 4
75. Shi, Y., Fernando, B., Hartley, R.: Action anticipation with rbf kernelized feature mapping rnn. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 301–317 (2018) 4
76. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10740–10749 (2020) 3
77. Six, J., Leman, M.: Synchronizing Multimodal Recordings Using Audio-To-Audio Alignment. *Journal of Multimodal User Interfaces* **9**(3), 223–229 (2015). <https://doi.org/10.1007/s12193-015-0196-1> 8
78. Strabala, K.W., Lee, M.K., Dragan, A.D., Forlizzi, J.L., Srinivasa, S., Cakmak, M., Micelli, V.: Towards seamless human-robot handovers. *Journal of Human-Robot Interaction* **2**(1), 112–132 (2013) 4
79. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: European conference on computer vision. pp. 581–600. Springer (2020) 4
80. Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.: Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* **28**, 675–735 (11 2005) 2
81. Villa-Cueva, E., Ahmed, S., Chevi, R., Cruz, J.C.B., Elzeky, K., Cristobal, F., Aji, A.F., Wang, S., Mihalcea, R., Solorio, T.: Moments: A comprehensive multimodal benchmark for theory of mind. *arXiv preprint arXiv:2507.04415* (2025) 2, 4
82. Vondrick, C., Pirsivash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 98–106 (2016) 4
83. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., et al.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20270–20281 (2023) 2, 3

84. Wang, Z., Zhao, X., Stepputtis, S., Kim, W., Wu, T., Sycara, K., Xie, Y.: Himem-former: Hierarchical memory-aware transformer for multi-agent action anticipation. arXiv preprint arXiv:2411.01455 (2024) 4
85. Wiederhold, N., Megyeri, A., Paris, D., Banerjee, S., Banerjee, N.: Hoh: Marker-less multimodal human-object-human handover dataset with large object count. *Advances in Neural Information Processing Systems* **36**, 68736–68748 (2023) 2, 3, 4
86. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: EgoLife: Towards egocentric life assistant. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 28885–28900 (2025) 3
87. Ye, Z., Li, Y., Liu, Y., Bridges, C., Rozga, A., Rehg, J.M.: Detecting bids for eye contact using a wearable camera. In: *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. vol. 1, pp. 1–8. IEEE (2015) 4
88. Zhang, C., Fu, C., Wang, S., Agarwal, N., Lee, K., Choi, C., Sun, C.: Object-centric video representation for long-term action anticipation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 6751–6761 (January 2024) 4
89. Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J.B., Shu, T., Gan, C.: Building cooperative embodied agents modularly with large language models. arXiv preprint arXiv:2307.02485 (2023) 2, 3
90. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In: *European conference on computer vision*. pp. 180–200. Springer (2022) 2, 3
91. Zhang, X., Chen, Y., Yeh, S., Li, S.: Metamind: Modeling human social thoughts with metacognitive multi-agent systems. arXiv preprint arXiv:2505.18943 (2025) 2, 4
92. Zhao, Q., Wang, S., Zhang, C., Fu, C., Do, M.Q., Agarwal, N., Lee, K., Sun, C.: AntGPT: Can large language models help long-term action anticipation from videos? In: *The Twelfth International Conference on Learning Representations* (2024) 4