ELLSA-BENCH: A CLOSED-LOOP BENCHMARK FOR OMNIMODAL HUMAN–SCENE–ROBOT INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world human–robot interaction is inherently continuous, multimodal and situated: robots must perceive, communicate and act while coordinating with humans and the surrounding environment. Despite rapid progress in embodied models, few support real-time omnimodal interaction[1], while existing benchmarks lack closed-loop settings that jointly evaluate these capabilities. We introduce **ELLSA-Bench**, a closed-loop benchmark for naturalistic omnimodal human–scene–robot interaction in simulated environments. ELLSA-Bench explicitly models human behaviors through speech, body motion, gestures and lip movements, and evaluates robots across audio–visual perception, spoken communication and physical action. It comprises 40 atomic scenarios across four families and three interaction levels, covering audiovisual grounding, human–robot collaboration, concurrent communication and action and interruption-aware interaction. To capture the continuous nature of real-world interaction, we incorporate full-duplex tasks requiring robots to listen and observe while speaking and acting. For evaluation, an LLM planner is used to generate scene-grounded interaction plans and synthesize temporally coordinated dialogue, human behaviors and robot actions for scalable naturalistic interaction generation. Finally, we introduce two complementary baselines for full-duplex interaction: a streaming end-to-end audio–visual–action model and an asynchronous agent-based framework, Robo-Harness. Experimental results show that while current models can perform visual perception, speech interaction and action execution, they still struggle with scenarios requiring coordinated interaction across multiple modalities.

1 INTRODUCTION

Recent advances in embodied intelligence are bringing robots closer to becoming useful assistants in everyday life (Black et al., 2025). Realizing this potential requires more than acting on the physical world: robots must also understand and coordinate with the people who share it (Puig et al., 2024). When setting a table together, a person may request a cup, point to the intended object, and continue speaking as the robot reaches for it. Natural interaction is fluid: people coordinate seeing, listening, speaking and acting, often perceiving new cues while responding to their partners. This calls for omnimodal assistance that jointly integrates these four channels, bringing interaction with people and interaction with the physical environment into a shared, ongoing activity.

Research on embodied interaction models has advanced language-guided action, audiovisual intent understanding and spoken communication (Black et al., 2025; Yao et al., 2025; Wang et al., 2026a). More integrated systems can speak while acting and respond to new human input during execution (Wang et al., 2026b; Liu et al., 2025). These advances raise a broader question: how effectively can robots combine these capabilities during everyday assistance? Answering it requires evaluating not only task completion, but also how robots understand and coordinate with their human partners.

Existing benchmarks capture complementary aspects of this setting. Real-world recordings capture human speech and behavior, while streaming audiovisual benchmarks assess responses to unfolding inputs (Wang et al., 2023; Lu et al., 2026), but neither reflects how robot actions change the scene.

[1]Omnimodal denotes the integrated loop of embodied listening; multimodal refers to combinations of perceptual cues such as speech and vision.

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.

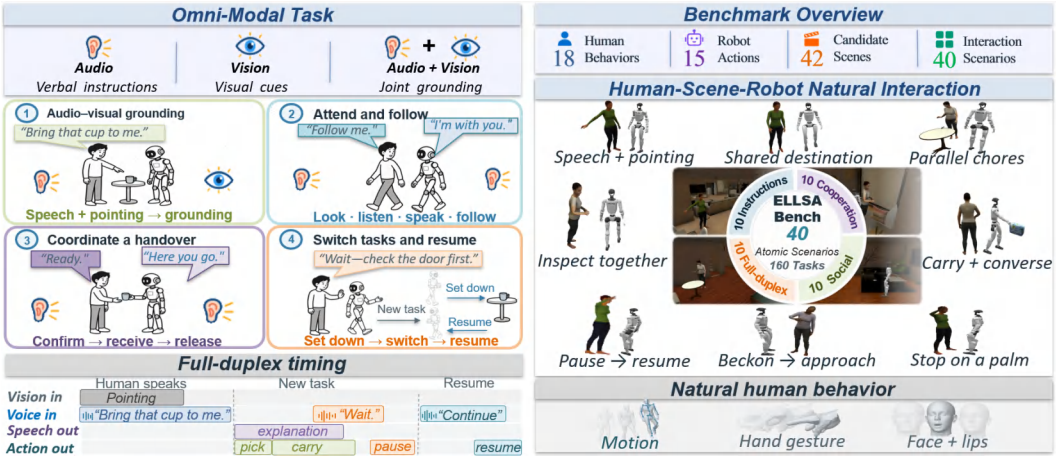


Figure 1: ELLSA-Bench overview. **Left:** Concurrent speech, vision and action enable human–scene–robot interaction, including full-duplex tasks. **Right:** Diverse, realistic simulation scripts generate human behavior through synchronized speech, body motion, gestures and lip movements.

Simulation benchmarks provide action-dependent feedback and increasingly incorporate human–robot collaboration (Yang et al., 2025; Chang et al., 2025; Liu et al., 2026a) yet rarely evaluate spoken interaction, expressive human behavior and physical execution together. This motivates a shared executable setting where human speech and behavior guide robots while robot actions reshape the environment, under standardized protocols for fair and reproducible evaluation.

To address this need, we introduce **ELLSA-Bench**, a closed-loop simulation benchmark for naturalistic omnimodal human–scene–robot interaction (Figure 1). Built on Habitat 3.0 (Puig et al., 2024), it models humans with coordinated speech, body motion, gestures and lip movements. ELLSA-Bench contains 160 tasks across 40 atomic scenarios in four families, requiring robots to integrate speech, vision and action while interacting with humans and the environment. Selected full-duplex tasks further require robots to keep listening and observing while speaking or acting, with selective interruption and resumption. An automated LLM-based pipeline generates scene-grounded interaction plans and temporally coordinated human and robot behaviors, yielding 4,521 training examples. We establish two complementary baselines: a finetuned streaming end-to-end audio–visual–action model and Robo-Harness, a zero-shot asynchronous agent framework. Evaluation reveals persistent challenges in grounding visual human cues and coordinating modalities during ongoing interaction, particularly when robots must adapt one behavior to newly arriving cues while preserving unaffected actions and task state. Our contributions are threefold:

- We introduce **ELLSA-Bench**, the first closed-loop benchmark for naturalistic omnimodal human–scene–robot interaction, evaluating how robots coordinate with human partners while perceiving and acting on their surroundings.
- We develop construction and evaluation pipelines with naturalistic human modeling, diverse assistance scenarios and full-duplex interaction tasks. An LLM-based planner automatically generates scene-grounded interaction plans, enabling scalable synthesis of naturalistic embodied interactions.
- We establish two complementary baselines for full-duplex interaction, a fine-tuned streaming end-to-end model and a zero-shot asynchronous harness-based framework, and evaluate listening, looking, speaking and acting under a unified protocol.

2 Related Work

2.1 Multimodal embodied interaction models

Multimodal embodied interaction models connect human communication with robot perception and action. VLAS (Zhao et al., 2025) grounds spoken instructions in manipulation, while GIVE (Liu

et al., 2026b) and Gaze2Act (Zuo et al., 2026) incorporate gestures and gaze, respectively, to resolve human intent. RoboEgo (Yao et al., 2025), RoboOmni (Wang et al., 2026a), and EDITH (Lee et al., 2026b) combine multimodal human cues with embodied behavior, while MICoBot (Yu et al., 2026) uses dialogue to coordinate collaborative manipulation. ELLSA (Wang et al., 2026b) and VITA-E (Liu et al., 2025) explicitly target concurrent seeing, listening, speaking and acting. However, systems that jointly support these capabilities during natural, continuous human–robot interaction remain limited. ELLSA-Bench provides a unified closed-loop benchmark for systematically evaluating how reliably models integrate multimodal understanding, communication and physical task execution in realistic human–robot assistance scenarios grounded in complex environments.

2.2 Datasets and benchmarks for situated interaction

Real-world resources capture complementary aspects of human interaction: HoloAssist (Wang et al., 2023) focuses on collaborative guidance, JRDB-Social (Jahangard et al., 2024) on social behavior, and SHREC (Lee et al., 2026a) on social reasoning in human–robot conversations. EgoAVU (Seth et al., 2026) evaluates egocentric audiovisual understanding, while VideoFDB (Mazumdar et al., 2026) and OmniInteract (Lu et al., 2026) assess full-duplex and streaming interaction. In executable environments, VLABench (Zhang et al., 2025), and EmbodiedBench (Yang et al., 2025) benchmark household activities and manipulation. TEACh (Padmakumar et al., 2022) adds text-based dialogue, while PARTNR (Chang et al., 2025), and HRIBench (Liu et al., 2026a) emphasize human–robot coordination. However, recorded observations cannot reflect the effects of robot actions, and executable benchmarks offer limited joint coverage of speech, expressive human cues for speech interaction and action execution, and concurrent perception and response (Table 1). This leaves natural, continuous assistance insufficiently evaluated. ELLSA-Bench addresses this gap by jointly evaluating these dimensions in a shared closed-loop setting.

Table 1: Benchmark coverage comparison. ✓: support; △: limited; —: not targeted.

Resource	Closed-loop	Visual	Listen	Speak	Cue	Duplex
HoloAssist	—	✓	△	—	✓	—
SHREC	—	✓	△	—	△	—
VideoFDB	—	✓	✓	✓	△	✓
EmbodiedBench	✓	✓	—	—	—	—
TEACh	✓	✓	—	—	—	—
HRIBench	✓	✓	—	—	✓	—
ELLSA-Bench	✓	✓	✓	✓	✓	✓

3 ELLSA-Bench Benchmark

ELLSA-Bench models natural human–robot interaction as a temporally coordinated activity in a shared physical environment. Human speech and visible behaviors specify object referents, authorize actions, revise goals and regulate ongoing interaction. We align these communicative events with human and robot actions in an evolving scene via a pipeline integrating a scene-grounded interaction plan, spatially conditioned human behaviors, and state-dependent temporal execution.

3.1 Benchmark Overview

Tasks and data ELLSA-Bench covers verbal instruction understanding, nonverbal cue interpretation, human–robot–object relational reasoning, task and partner state tracking, intention adaptation, and social awareness for natural, fluid, real-time human–robot interaction. These tasks are organized into four families: **(1) Multimodal Instructions**: includes spoken object transfer, gesture-conditioned selection, successive pointing, reference clarification, nonverbal consent, relative movement, and ordered manipulation, evaluating multimodal grounding, ambiguity resolution, and goal-directed execution. **(2) Human–Robot Cooperation**: includes permission and readiness coordination, workspace clearing, parallel chores, coordinated placement, handovers, and object relays, testing partner-state tracking, possession coordination, and adaptive collaboration. **(3) Full-Duplex Control**: covers task reporting, goal revision, action pause/resumption, and selective interruption of speech or motion, requiring adaptation to changing intentions while maintaining interaction continuity. **(4) Social Interaction**: includes attention acquisition, conversational orientation and distance, walking while conversing, nonverbal feedback, turn management, and gesture games, assessing social awareness and coordination beyond task completion. The four families contain 40 atomic scenarios, ten per family. We provide 4,521 training examples and 160 manually revised and

verified closed-loop test episodes: 120 single-scenario and 40 multi-scenario episodes connecting compatible tasks while preserving object states and dialogue context. Scenario specifications and data details are provided in Appendices A and B, respectively.

Human dialogue and behavior The ELLSA-Bench benchmark contains 1,601 utterances across four complementary dialogue categories (detailed in Appendix C). **General QA**, including items adapted from VoiceBench's CommonEval subset (Chen et al., 2024), assesses general conversational ability. During ongoing tasks, **State QA** covers questions about the surroundings, the robot's own actions, and task progress. **Full-Duplex QA** includes interruptions, backchannels and action-coupled exchanges that unfold alongside ongoing execution. **Task Reply** covers verbal responses to human action instructions, which require grounding spoken goals in the scene, sometimes jointly with visual cues. Episodes average 11.2 utterances. Human behavior spans 18 action types (Appendix D.1). Alongside walking, object manipulation, and handovers, gestures such as pointing, nodding, waving, and stop signals support reference resolution, feedback, attention, and action control. To identify model strengths and weaknesses, we annotate task-relevant input as $\boldsymbol{M_1}$ **Speech** (requests), $\boldsymbol{M_2}$ **Visual** (human gestures), or $\boldsymbol{M_3}$ **Speech–Visual** (speech with visual cues), and interaction level as $\boldsymbol{I_1}$ **Sequential** (ordered exchanges), $\boldsymbol{I_2}$ **Concurrent** (talking while acting), or $\boldsymbol{I_3}$ **Interrupt/Revise** (stopping or updating ongoing behavior), as summarized in Figure 2.

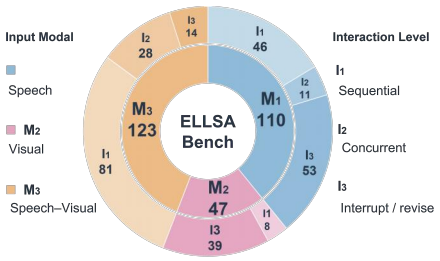


Figure 2: Interaction coverage.

Environments and robot actions We use Habitat 3.0 (Puig et al., 2024) with the PARTNR environment setup (Chang et al., 2025), HSSD scene assets (Khanna et al., 2024), humanoid avatars, and a simulated Unitree G1 (Unitree Robotics, 2024). Training and test generation use separate scene pools, with ten held-out test layouts. The robot supports 15 actions (Appendix D.2), including grounded skills such as `NAVIGATE`, `PICK`, `PLACE`, `GIVE`, and `RECEIVE`, action confirmation via `ACKNOWLEDGE`, and left/right hand raises for gesture games. Low-dimensional `MOVE_RELATIVE` and `TURN_RELATIVE` commands specify travel distances and rotation angles from current observations, enabling flexible social positioning when facing, approaching, or following people. `BREAK`/`STOP`, `RESUME`, and `CANCEL` independently pause, resume, or cancel actions, testing full-duplex control and task continuity under interruption.

3.2 Natural Human Behavior Synthesis

To support more realistic human–robot interaction, we synthesize human body movements, hand gestures, and facial responses in Habitat, coordinating them with dialogue and robot actions throughout each scenario. We generate body motion using Kimodo (Rempe et al., 2026), conditioned on action descriptions, scene-derived root and wrist targets, and sparse navigation waypoints. These movements are complemented with MANO-based hand poses (Romero et al., 2017), aligning pointing, grasping, and release with the corresponding action phases. To coordinate visible behavior with the conversation, we use UniLS (Chu et al., 2026) to generate speaking and listening facial motion from both participants' delivered audio, including overlaps and interruptions.

3.3 Story-Driven Interaction Data Generation

Dynamic story generation Constructing closed-loop human–robot episodes requires dialogue, human behaviors, robot actions, and scene states to remain coherent throughout an interaction. Manually specifying storylines, motions, trajectories, and event triggers for each episode is labor-intensive. To generate natural, scene-grounded human–robot interactions, we use an LLM task planner (DeepSeek-V4-Flash (DeepSeek-AI, 2026a)) to coordinate generation from atomic task scripts, interaction rules, the current scene inventory, and available semantic meta-actions. The planner expands these inputs into a shared interaction plan, composing the storyline, human utterances and behaviors, robot reference actions, and their temporal dependencies. Concrete failures identified during validation provide feedback for bounded plan revisions. We represent the resulting plan as

$$\mathcal{P} = \mathrm{Expand}_{\mathrm{LLM}}(\mathcal{T}, \mathcal{R}, S_0) = (\mathcal{E}_{\mathrm{dlg}}, \mathcal{E}_{\mathrm{act}}, \mathcal{B}, \mathcal{D}), \tag{1}$$

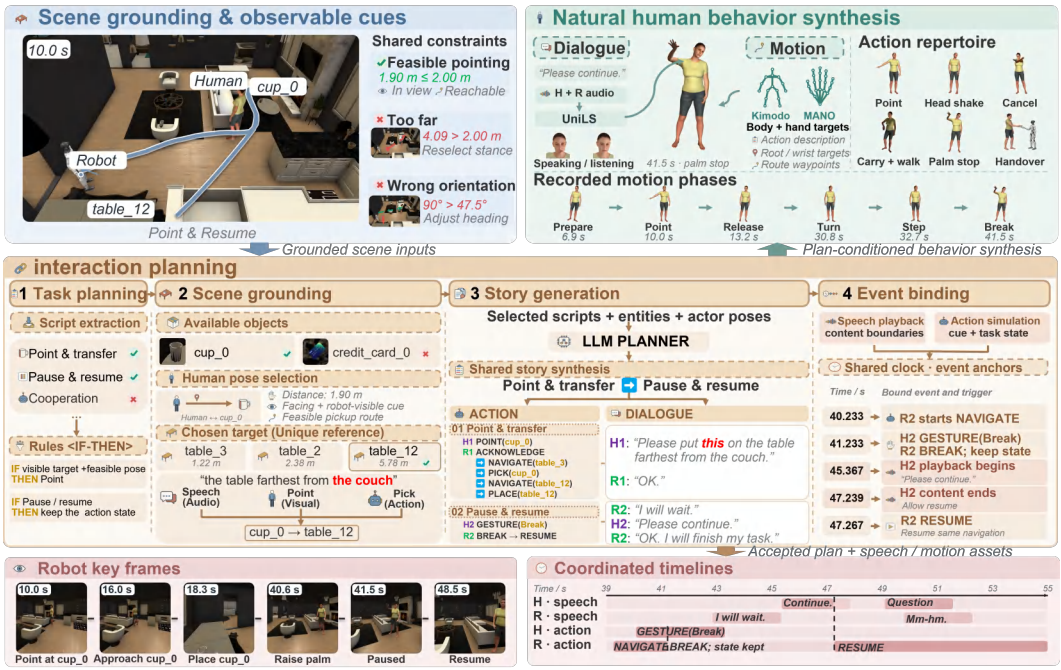


Figure 3: The interaction data generation pipeline. It shows how the full pointing and pause/resume tasks combine interaction planning, behavior generation and simulation event binding.

where dialogue and action events share scene-entity bindings $\mathcal{B}$ and temporal dependencies $\mathcal{D}$. Task requirements remain consistent across episodes, while dialogue, interaction details, and the ordering of compatible tasks can vary. More details are shown in Appendix E.

Joint grounding and observable interaction The planner jointly grounds human communication, scene state, and robot behavior to generate natural, scene-consistent human–robot interactions. Spoken references, gestures, and robot operations are grounded to shared scene entities through a unified interaction state. Linguistic references are resolved through attribute- and relation-based descriptions, enabling unique object grounding among distractors (e.g., ”the cup on the table next to the sofa” or ”the table farthest from couch”). Multimodal cues such as pointing and gaze further constrain referent selection and resolve ambiguous expressions such as ”this.” The planner maintains task-level constraints over human–robot–object relations, including object availability, possession states, spatial arrangements, and action dependencies. It also validates interaction feasibility by checking actor orientations, viewing geometry, navigation routes, and gesture visibility, ensuring that planned cooperation can be reliably executed in the scene.

Coordinated timelines and state-dependent execution Habitat executes the accepted plan through dialogue and action channels, each with human and robot tracks. These four tracks share a simulation clock and cross-channel dependencies, allowing speech, visible cues, and physical actions to overlap. The LLM planner specifies relative event order; execution binds it to action progress, geometric signals, measured speech-content boundaries, and live scene state. Thus, a human can provide feedback or request a change while the robot is speaking or moving, and the affected behavior can respond before it finishes. An interruption is triggered while its target remains active, while a paused action retains its state for permission-based resumption.

4 Full-Duplex Interaction Baselines

To evaluate continuous human–scene–robot interaction in ELLSA-Bench, we develop two complementary baselines for full-duplex interaction, where robots can see, listen, speak and act concurrently. Both approaches integrate streaming perception with communication and physical action, enabling robots to respond to emerging human cues while speaking or executing tasks. All stream-

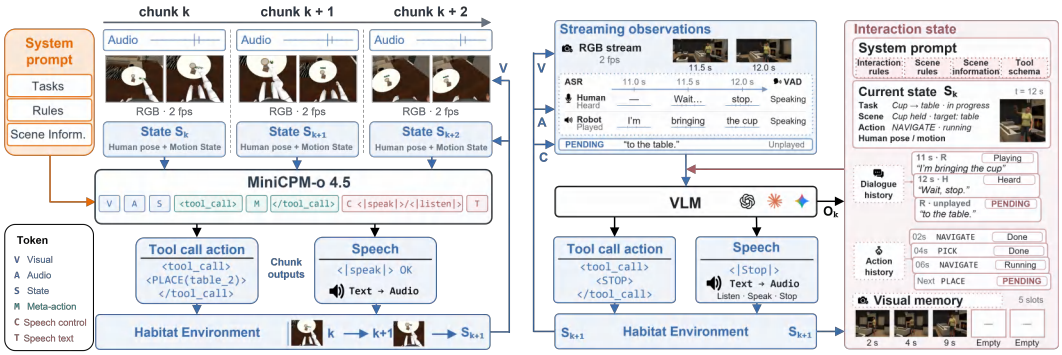


Figure 4: Two full-duplex interaction baselines. **Left:** An end-to-end embodied model based on MiniCPM-o 4.5. **Right:** A zero-shot Robo-Harness framework for asynchronous speech and action.

ing modalities are transmitted in 1-second chunks, and robot control is realized through tool calling over the meta-action interface introduced in Section 3.1. Structured simulator state is used only to expose observable execution information, such as agent positions, object custody, and action status. Future human dialogue, reference plans, and subsequent task events are never exposed to the models. Gesture labels and intended referents are not provided through structured state and must be inferred from visual observations.

End-to-end full-duplex audio-visual-action embodied model We fine-tune MiniCPM-o 4.5 (Cui et al., 2026) on the interaction data generated in Section 3.3, extending its streaming audiovisual interface with robot actions. The inputs and outputs at decision step k are

$$O_k = (A_k, V_k, S_k), \qquad Y_k = (M_k, T_k, C_k), \tag{2}$$

where A_k contains the preceding second of human audio, V_k contains available RGB frames sampled at 2 fps, and S_k provides observed state updates. An initial prompt specifies the scene inventory and action rules. Subsequent inputs update human and robot positions, held objects, robot motion, and relative distance and bearing, while preceding observations and outputs remain in context. The model produces tool-call meta-actions M_k, speech text T_k, and interaction controls C_k for listening and speaking. Action and speech outputs are routed separately, allowing motor and speech control to operate independently. We use LoRA (Hu et al., 2022) and targeted control-output adaptation with supervision over actions, speech text, and control tokens (detailed in Appendix F).

Robo-Harness Robo-Harness (detailed in Appendix G) connects a pretrained VLM to Habitat for zero-shot full-duplex interaction without task-specific weight updates. Its evolving interaction state integrates streaming audio and RGB chunks, human pose and motion, scene and execution states, dialogue and action histories, and five persistent visual-memory slots storing LLM-selected observations with timestamps and brief context notes. Human and robot speech are tracked separately through time-aligned ASR prefixes, playback progress and audio-based VAD, distinguishing received, delivered and unplayed speech. The system prompt provides rules and context for interpreting human behavior, grounding gestures and invoking valid meta-actions. The model outputs tool-call meta-actions, speech decisions and optional memory updates. Actions and synthesized speech execute through independent asynchronous channels supporting interruption, pausing, resumption and cancellation. Execution receipts and observations update the state as simulation advances asynchronously, enabling concurrent perception, listening, speaking and action to emulate full-duplex interaction, with real-time responsiveness assumed rather than guaranteed.

5 Closed-Loop Evaluation

Natural human–robot interaction is continuous and multimodal: speech and visible cues guide physical actions, while those actions reshape the scene and subsequent exchanges. ELLSA-Bench tests these coupled behaviors within shared closed-loop episodes. We evaluate **task performance** (T) for goal achievement and process compliance, **social coordination** (S) for physical coordination with the human partner, and **spoken interaction** (V) for response content, timing, and conversational control. These dimensions are complementary: completing a task does not establish that the

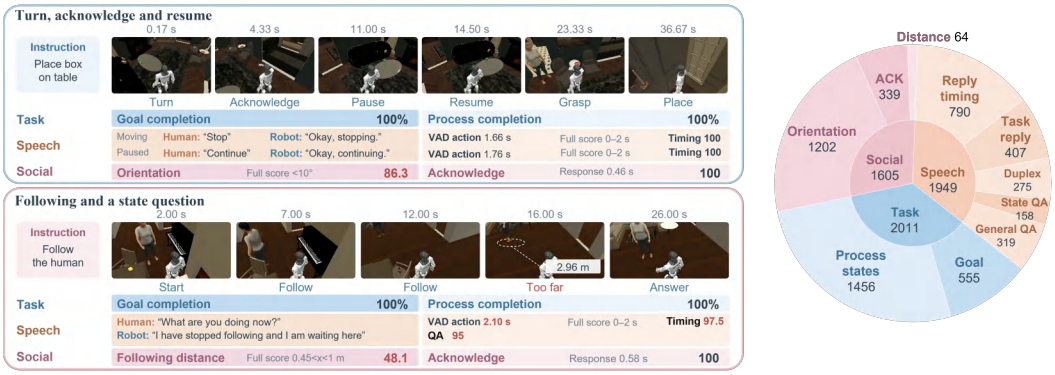


Figure 5: **Left:** Evaluation examples: pause–resume control and human following with state QA. **Right:** Distribution of scoring checkpoints.

robot attended to its partner or handled an interruption appropriately. Each dimension and constituent metric can be reported separately, while behaviors involving multiple modalities are judged from their joint evidence. Figure 5 presents transport and following showcases and the distribution of 5,565 configured checkpoints across 160 test episodes (120 single-scenario and 40 composed). Appendix H provides a detailed scoring protocol.

Closed-loop execution and assessment Human cues follow modality-specific triggers tied to the robot's actual behavior and evolving scene. For *speech*, ordinary turns respect required quiet intervals, while interruptions and backchannels occur during an active target utterance. Silero VAD (Silero Team, 2024) estimates acoustic boundaries, and Whisper (Radford et al., 2023) locally transcribes played speech for content assessment. For *visual and spatial cues*, triggers may depend on relative distance, viewing direction, and line of sight; robot-view frames and recorded poses verify cue visibility and support assessment of pointing-based selection, facing, and following. For *actions and task states*, observed progress and physical relations trigger cues during carrying, handover, or placement, while state rules verify outcomes such as possession transfer, pausing, and authorized resumption. A shared timeline enables joint evaluation of cues and responses across modalities. A multimodal semantic judge assesses contextual replies and state questions using transcripts, dialogue history, scene state, and aligned visual keyframes.

Task performance (T) We separately assess goal outcomes, process correctness, and instance grounding. Scene-state predicates verify outcomes such as placement and possession transfer; execution records track intermediate states and verify phase progression, required actions, and pause/cancel/resume compliance. Grounding checks whether manipulated objects match human references: placing a different cup at the requested destination may satisfy a category-level goal but fail instance grounding, so completion and grounding errors are recorded separately. Social behavior is assessed independently under S: functional following may still involve inappropriate interpersonal spacing.

Social coordination (S) We assess orientation, bodily acknowledgment, interpersonal distance, and response timing against human interaction conventions. Recorded poses and execution events verify that the robot faces the speaker during relevant idle-conversation intervals and performs `ACKNOWLEDGE` after designated action instructions are fully delivered. Invitations to approach for conversation require settling within a preferred 1.0–1.5 m range; following checks for excessive or insufficient distance along the trajectory. Response windows align with delivered cues and prerequisites to assess the timing of acknowledgment, orientation changes, and movement.

Spoken interaction (V) We assess relevance, sufficiency, contextual consistency, and backchannel appropriateness using GPT-5.6 Luna with ASR transcripts of delivered speech and dialogue history. State QA adds compact scene-state information and an aligned robot-view keyframe to verify claims about possession, location, and task progress. ASR/VAD timestamps and playback records measure reply, speech-stopping, and recovery latencies and verify yielding to human interruptions. Aligned speech and execution records assess context-appropriate resumption and selective interruption handling: a stop-talking request should silence the robot without interrupting ongoing transport.

Table 2: Task success rates (%) and task scores (T). Gap denotes Composed minus Single.

	Success rate (%)				T			
Model	Single	Composed	Gap	Overall	MI	HC	FD	SI
End-to-end								
FT-MiniCPM-o	43.97	29.73	-14.24	40.52	36.90	66.41	73.91	39.89
Robo Harness								
GPT-5.6 Luna	53.57	32.43	**-21.14**	48.32	48.65	70.29	70.55	57.32
GPT-5.6 Terra	**67.31**	**52.94**	-14.37	**63.77**	**60.40**	**88.41**	**86.89**	55.56
Gemini 3.5 Flash	56.14	42.11	-14.03	52.63	55.56	75.17	72.43	55.20
DeepSeek-V4.1-Flash	62.63	43.75	-18.88	58.02	50.00	83.58	78.57	**57.69**

Table 3: **Left:** Scores by task-bearing modality. **Right:** Scores by interaction level.

	Task-bearing modality									Interaction level								
	M_1			M_2			M_3			I_1			I_2			I_3		
Model	T	S	V	T	S	V	T	S	V	T	S	V	T	S	V	T	S	V
End-to-end																		
FT–MiniCPM-o	84.47	60.07	54.64	10.00	44.81	53.24	38.03	55.85	53.90	47.64	55.89	53.97	80.49	53.85	55.74	48.29	58.81	53.74
Robo Harness																		
Luna	83.87	58.70	59.94	**31.11**	**57.39**	57.54	50.39	57.16	61.06	56.05	56.26	**62.77**	76.28	60.21	58.92	56.25	57.94	57.52
Terra	**91.84**	**65.57**	**62.49**	28.57	56.25	**58.98**	**68.00**	**64.30**	**62.22**	59.12	61.35	61.89	84.62	**66.06**	62.39	**86.89**	**67.01**	**61.67**
Gemini	80.37	59.37	58.43	16.27	48.73	57.43	61.43	60.43	61.27	**62.76**	**61.71**	59.78	77.82	56.64	**63.91**	52.73	62.67	56.11
DeepSeek	91.81	63.92	59.53	25.00	49.53	57.85	51.89	60.65	58.05	61.11	60.40	59.09	**92.00**	61.91	58.91	64.21	62.65	58.09

6 Experiments

6.1 Experimental Setup

We evaluate fine-tuned MiniCPM-o 4.5 (Cui et al., 2026) synchronously and Robo-Harness asynchronously with four backbones: GPT-5.6 Luna and Terra (OpenAI, 2026), Gemini 3.5 Flash (Google DeepMind, 2026), and DeepSeek-V4.1-Flash (DeepSeek-AI, 2026b). Tasks cover multimodal instructions (MI), human–robot cooperation (HC), full-duplex control (FD), and social interaction (SI). We group episodes by task-bearing modality, $M_1/M_2/M_3$ (speech/visual/both), and interaction pattern, $I_1/I_2/I_3$ (sequential/concurrent/interruption or goal revision), and report task (T), social (S), and spoken (V) performance. All configurations share the same episodes and protocol, with 16 kHz audio, 512×512 visual observations, and independent action and speech channels.

6.2 Benchmark Results

Overall performance Table 2 and Table 4 show that the streaming MiniCPM-o baseline achieves 40.52% overall SR, while the asynchronous Robo-Harness configurations achieve 48.32–63.77%. Because the two baseline families differ in model architecture, input representation, and interaction mechanism, we treat this comparison as a characterization of complementary system designs rather than a controlled backbone comparison. All models score lower on longer composed episodes, with Luna showing the largest decline (21.14%), suggesting limitations in sustained reasoning and memory as forgotten instructions or missed state updates can disrupt subsequent actions and compound failures. Across models, HC and FD task scores are higher than MI and SI. These results highlight persistent difficulties in grounding human intent, handling social interaction, and completing extended interaction sequences despite stronger cooperation and control performance.

Modality-specific performance Table 3 shows that models perform best when task intent is expressed verbally: T reaches 80.37–91.84 on M_1, compared with 10.00–31.11 on M_2 and 38.03–68.00 on M_3. Since scene vision remains available throughout, this pattern suggests a strong dependence on speech to specify the required behavior. Tasks where human visual cues carry essential information remain difficult, including those combining speech and visual cues. Meanwhile, spoken-interaction scores V remain similar across modalities. Models can therefore maintain comparable verbal responses while struggling to complete tasks that depend on gestures.

Interaction-level performance The strongest backbone changes with the interaction demands. DeepSeek leads concurrent interaction (I_2; $T = 92.00$), but scores 64.21 on interruption and goal revision (I_3), where Terra leads with $T = 86.89$. This reversal suggests that concurrent speech-action coordination and intervention handling are distinct capabilities. The latter requires selective control: a request to stop speaking should leave an ongoing transport action intact, while paused motion must await permission to resume. Reliable continuous interaction therefore requires updating the intended behavior while preserving unaffected actions and task state for subsequent execution.

Table 4: Scores for the benchmark indicators.

Dimension	Indicator	MiniCPM	Luna	Terra	Gemini	DeepSeek
Task	Goal	58.82	65.10	75.36	64.27	72.52
	Process states	48.69	60.40	73.19	60.83	67.18
Social	Orientation	78.31	65.21	64.97	66.47	76.83
	Acknowledge	51.87	65.43	79.38	61.83	68.36
	Distance	46.92	46.49	47.38	48.27	48.46
Speech	Reply timing	55.06	64.18	77.05	63.47	53.04
	Task reply	50.45	71.56	78.50	62.23	49.68
	Duplex	56.22	49.06	58.93	54.09	57.86
	State QA	52.62	67.37	54.27	71.83	60.80
	General QA	51.38	67.61	72.41	60.23	69.18

Table 5: Modality ablation.

Modality / T	MiniCPM-o	Luna	Terra	Gemini	DeepSeek
M_1	78.56	74.87	**92.80**	84.26	90.81
M_2	15.38	29.64	**34.53**	32.27	24.09
M_3	36.83	55.78	62.34	**62.49**	49.66

6.3 Error Analysis

Figure 6 reveals two recurring difficulties in the weighted audit subset, with model-wise failure distributions detailed in Appendix I.1. First, models struggle to select actions from human visual cues. Grounding errors account for 51.49% of failures: models sometimes request verbal target descriptions despite verified visibility of pointing cues. This suggests dependence on explicit verbal specification, consistent with the stronger task performance on M_1 than on gesture-dependent M_2/M_3 episodes. Second, models may recognize human feedback without applying it to ongoing speech or motion. Action control/recovery and speech response/control account for 25.74% and 14.85%, respectively. A robot may acknowledge a stop request yet continue approaching, or complete transport without yielding the conversational turn. In composed stories, a missed invitation or interruption can cause failure despite later successful manipulation. These failures call for continuously connecting human cues with action selection and control throughout interactions.



Figure 6: Primary error distribution across all models.

6.4 Modality Ablation in Matched Scenes

To isolate modality effects from scenario difficulty, we evaluate 10 matched scenes across M_1, M_2, and M_3, keeping each scene's task goal fixed. For example, an object transfer is specified through an explicit verbal description, successive pointing to the object and destination, or "move this there" accompanied by pointing. Scene layout, initial states, interaction schedule, model configuration, and execution budget remain fixed, with scene vision available throughout. Each model is evaluated once per scene and condition, yielding 30 episodes. Table 5 shows a consistent ordering of $T(M_1) > T(M_3) > T(M_2)$, with speech and visual cues outperforming visual cues alone by 21.45–30.22 points. Terra leads on M_1 and M_2, while Gemini and Terra are nearly tied on M_3 (62.49 and 62.34). Gemini shows a smaller M_1–M_3 gap (21.77 points), compared with over 41 points for MiniCPM-o and DeepSeek. These results highlight a gap between following explicit instructions and interpreting natural requests that leave details to gestures and scene context (see Appendix I for additional gesture-task designs and qualitative examples).

7 Conclusion

In this work, we introduced ELLSA-Bench, a closed-loop benchmark for continuous human–scene–robot interaction across speech, vision and physical action. Its automated generation pipeline coordinates human speech and visible behavior with robot actions in evolving scenes. The benchmark supports concurrent listening, seeing, speaking and acting, with two complementary baselines for full-duplex interaction. Experiments reveal that evaluated models rely heavily on explicit verbal instructions and struggle to ground visual or audiovisual human cues. Performance further declines in composed interactions, where missed cues and failures to act on feedback can disrupt task continuity. These findings highlight the need to ground human intent across modalities and coordinate speech and action as interaction unfolds. By jointly evaluating task performance, social coordination and spoken interaction, ELLSA-Bench provides a common foundation for developing embodied assistants responsive to human partners throughout task execution.

AI USE STATEMENT

AI assistance supported literature organization, drafting, experiment planning, code inspection, and figure-production code. All AI-generated drafts and codes were manually reviewed, verified, and, where necessary, revised by the authors. The data-generation design uses LLMs for interaction-plan authoring and revision.

ETHICS STATEMENT

ELLSA-Bench is developed for research on natural and reliable human–robot interaction. All benchmark interactions are conducted in simulated environments, and the benchmark construction process does not require collecting personal data or involving human participants in physical robot experiments. Human behaviors and interaction scenarios are synthesized or generated through controlled pipelines, reducing privacy and physical-safety risks during benchmark development and evaluation. The benchmark is intended to support the development and evaluation of embodied agents that can communicate and coordinate with humans in assistive settings. We encourage responsible use of ELLSA-Bench and recommend that systems derived from or evaluated on the benchmark undergo appropriate safety assessment before deployment in real-world environments.

REPRODUCIBILITY STATEMENT

Sections 3–5 and the appendices provide detailed descriptions of benchmark construction, model boundaries, triggering mechanisms, scoring evidence and planned configurations. The benchmark and construction pipeline will be released once they are fully organized.

REFERENCES

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. In *Interspeech 2023*, pp. 4489–4493, 2023. doi: 10.21437/Interspeech.2023-78. URL https://www.isca-archive.org/interspeech_2023/bain23_interspeech.html.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 2025. URL https://proceedings.mlr.press/v305/black25a.html.

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M. Turner, Eric Undersander, and Tsung-Yen Yang. PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.00081.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024.

Xuangeng Chu, Ruicong Liu, Yifei Huang, Yun Liu, Yichen Peng, and Bo Zheng. UniLS: End-to-End Audio-Driven Avatars for Unified Listening and Speaking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Chu_UniLS_End-to-End_Audio-Driven_Avatars_for_Unified_Listening_and_Speaking_CVPR_2026_paper.pdf.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026a. URL https://arxiv.org/abs/2606.19348.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, 2026b. URL https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash. Official model card and citation. Accessed September 24, 2026.

Google DeepMind. Gemini 3.5 Flash: Model card, May 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/. Published May 19, 2026. Accessed September 24, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Simindokht Jahangard, Zhixi Cai, Shiki Wen, and Hamid Rezatofighi. JRDB-Social: A Multifaceted Robotic Dataset for Understanding of Context and Dynamics of Human Interactions Within Social Groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22087–22097, 2024. doi: 10.1109/CVPR52733.2024.02085. URL https://openaccess.thecvf.com/content/CVPR2024/html/Jahangard_

JRDB-Social_A_Multifaceted_Robotic_Dataset_for_Understanding_of_Context_and_CVPR_2024_paper.html.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16384–16393, 2024.

Dong Won Lee, Yubin Kim, Sooyeon Jeong, Denison Guvenoz, Parker Malachowsky, Louis-Philippe Morency, Cynthia Breazeal, and Hae Won Park. Social Human Robot Embodied Conversation (SHREC) Dataset: Benchmarking Foundational Models' Social Reasoning. In *Proceedings of Robotics: Science and Systems*, 2026a. doi: 10.15607/RSS.2026.XXII.113. URL https://www.roboticsproceedings.org/rss22/p113.html.

Dongjun Lee, Juheon Choi, Dong Kyu Shin, Sinjae Kang, and Kimin Lee. Hierarchical policies from verbal and egocentric human signals for natural human-robot interaction, 2026b. URL https://arxiv.org/abs/2606.10276.

Chang Liu, Jiawei Zhang, Tao Zhang, Ye Wang, Hongyu Zhou, and Qin Jin. HRIBench: Benchmarking interaction-centric human-robot collaboration. *arXiv preprint arXiv:2607.13056*, 2026a. doi: 10.48550/arXiv.2607.13056. URL https://arxiv.org/abs/2607.13056.

Pengfei Liu, Gen Li, Junqiao Fan, Boyu Ma, Jindou Jia, Yang Xiao, and Jianfei Yang. GIVE: Grounding human gestures in vision-language-action models, 2026b. URL https://arxiv.org/abs/2606.13435.

Xiaoyu Liu, Chaoyou Fu, Chi Yan, Chu Wu, Haihan Gao, Yi-Fan Zhang, Shaoqi Dong, Cheng Qian, Bin Luo, Xiuyong Yang, Guanwu Li, Yusheng Cai, Yunhang Shen, Deqiang Jiang, Haoyu Cao, Xing Sun, Caifeng Shan, and Ran He. VITA-E: Natural Embodied Interaction with Concurrent Seeing, Hearing, Speaking, and Acting. arXiv preprint arXiv:2510.21817, 2025. URL https://arxiv.org/abs/2510.21817.

Xudong Lu, Xueying Li, Annan Wang, Yang Bo, Jinpeng Chen, Zengliang Li, Nianzu Yang, Rui Liu, Xue Yang, Jingwen Hou, and Hongsheng Li. OmniInteract: Benchmarking Real-World Streaming Interaction for Real-Time Omnimodal Assistants. arXiv preprint arXiv:2605.26485, 2026. URL https://arxiv.org/abs/2605.26485.

Amrita Mazumdar, Seonwook Park, Rajarshi Roy, Nikhil Srihari, Shengze Wang, Yuhao Zhou, Julia Wang, Koki Nagano, and Shalini De Mello. VideoFDB: Evaluating Full-Duplex Vision-Speech Capabilities in Conversational Agents. arXiv preprint arXiv:2605.30256, 2026. URL https://arxiv.org/abs/2605.30256.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026. URL https://openai.com/index/gpt-5-6/. Published July 9, 2026. Accessed September 24, 2026.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. TEACh: Task-Driven Embodied Agents That Chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2017–2025, 2022. doi: 10.1609/aaai.v36i2.20097. URL https://ojs.aaai.org/index.php/AAAI/article/view/20097.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.13724.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning*

Research, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling Controllable Human Motion Generation. arXiv preprint arXiv:2603.15546, 2026. URL https://arxiv.org/abs/2603.15546.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied Hands: Modeling and Capturing Hands and Bodies Together. ACM Transactions on Graphics (Proc. SIGGRAPH Asia), 2017. URL https://doi.org/10.1145/3130800.3130883.

Ashish Seth, Xinhao Mei, Changsheng Zhao, Varun Nagaraja, Ernie Chang, Gregory P. Meyer, Gael Le Lan, Yunyang Xiong, Vikas Chandra, Yangyang Shi, Dinesh Manocha, and Zhipeng Cai. EgoAVU: Egocentric Audio-Visual Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15805–15814, 2026. URL https://arxiv.org/abs/2602.06139.

Silero Team. Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024. Software repository.

Unitree Robotics. Unitree mujoco. https://github.com/unitreerobotics/unitree_mujoco, 2024. GitHub repository, accessed September 25, 2026.

Siyin Wang, Jinlan Fu, Feihong Liu, Xinzhe He, Huangxuan Wu, Junhao Shi, Kexin Huang, Zhaoye Fei, Jingjing Gong, Zuxuan Wu, Yu-Gang Jiang, See-Kiong Ng, Tat-Seng Chua, and Xipeng Qiu. RoboOmni: Proactive Robot Manipulation in Omni-modal Context. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=OJh7oBCYhL.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, and Chao Zhang. End-to-end Listen, Look, Speak and Act. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2510.16756.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2309.17024.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70576–70631. PMLR, 2025. URL https://proceedings.mlr.press/v267/yang25f.html.

Yiqun Yao, Xiang Li, Xin Jiang, Xuezhi Fang, Naitong Yu, Aixin Sun, and Yequan Wang. RoboEgo System Card: An Omnimodal Model with Native Full Duplexity. arXiv preprint arXiv:2506.01934, 2025. URL https://arxiv.org/abs/2506.01934.

Albert Yu, Chengshu Li, Luca Macesanu, Arnav Balaji, Ruchira Ray, Raymond Mooney, and Roberto Martín-Martín. Mixed-initiative dialog for human-robot collaborative manipulation. In *IEEE International Conference on Robotics and Automation*, 2026. doi: 10.48550/arXiv.2508.05535. URL https://arxiv.org/abs/2508.05535.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. VLABench: A Large-Scale Benchmark for Language-Conditioned Robotics Manipulation with Long-Horizon

Reasoning Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11142–11152, 2025. doi: 10.1109/ICCV51701.2025.01037. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_VLABench_A_Large-Scale_Benchmark_for_Language-Conditioned_Robotics_Manipulation_with_Long-Horizon_ICCV_2025_paper.html.

Wei Zhao, Pengxiang Ding, Min Zhang, Zhefei Gong, Shuanghao Bai, Han Zhao, and Donglin Wang. VLAS: Vision-language-action model with speech instructions for customized robot manipulation. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2502.13508. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/804a1a9de5787e3597b4bb64e1a48ec3-Abstract-Conference.html.

Kuangji Zuo, Gen Li, Bofan Lyu, Yanshuo Lu, Boyu Ma, Shijia Han, Xinyu Zhou, Xichen Yuan, Chuhao Zhou, Jiaqi Bai, Geng Li, and Jianfei Yang. Gaze2Act: Gaze-conditioned vision-language-action policies for interactive robot manipulation, 2026. URL https://arxiv.org/abs/2605.30282.

A SCENARIO SPECIFICATIONS

A.1 ATOMIC SCENARIOS AND MODALITY–INTERACTION LEVELS

ELLSA-Bench organizes 40 atomic scenarios into four families: Multimodal Instructions, Human–Robot Cooperation, Full-Duplex Control, and Social Interaction. Each family contains ten scenarios; the Social Interaction family includes seven social exchanges and three rule-based games. Table 6 specifies the human cue, the required robot behavior, and the modality and interaction labels for every scenario. Object categories and destinations in these specifications instantiate roles that can be grounded in different scene layouts. Each scenario is annotated by task-bearing modality (M_1: speech, M_2: visual cues, M_3: both) and interaction level (I_1: sequential, I_2: concurrent, I_3: interruption or revision).

Table 6: Complete catalogue of the 40 atomic scenarios.

ID	Atomic scenario	M	I
Multimodal Instructions (10)			
`interact_01`	**Spoken object and destination.** The person verbally identifies a cup and its destination without pointing. The robot resolves the description and completes the transfer.	M_1	I_1
`interact_02`	**Attention, beckoning, and assignment.** The person calls for attention, then beckons once visible. After approaching and stopping, the robot receives and executes a transport instruction.	M_3	I_1
`interact_03`	**Pointing to the object.** The person points to a nearby, visible cup and verbally specifies its destination. The robot identifies the indicated instance and transports it there.	M_3	I_1
`interact_04`	**Pointing to the destination.** The person names a red cup and points to a visible empty region. The robot retrieves the named cup and places it in that region.	M_3	I_1
`interact_05`	**Sequential source and destination pointing.** The person points first to a cup, then to its destination. The robot retains the object binding and uses the second gesture to complete the transfer instruction.	M_3	I_1
`interact_06`	**Nod to confirm, shake to cancel.** The person specifies a transport task and explains the gesture meanings. The robot attends to the person and executes or cancels the task after the corresponding gesture.	M_3	I_2
`interact_07`	**Clarifying an ambiguous gesture.** A broad pointing gesture leaves a cup, book, and plate as candidates. The robot asks which is intended and retrieves it only after verbal clarification.	M_3	I_1
`interact_08`	**Distinguishing two cups.** Two cups occupy different locations on opposite sides of the person. The robot combines pointing and a spoken left/right reference, confirms its selection, and retrieves that instance.	M_3	I_1
`interact_09`	**Pointing to a movement direction.** The person points to visible free space and requests a small movement. The robot moves once in that direction and stops, using the predefined 20 cm step.	M_3	I_1
`interact_10`	**Ordered transfers.** The person asks for a cup to be placed on the dining table, then a book on the bookshelf. The robot completes both transfers in that order.	M_1	I_1
Human–Robot Cooperation (10)			
`coop_01`	**Waiting for inspection.** The person inspects the destination while the robot waits empty-handed. The robot retrieves and transports the cup only after verbal confirmation.	M_1	I_1
`coop_02`	**Parallel nearby tasks.** The person returns a book to the bookshelf while the robot transfers a cup to a side table. Their assigned tasks overlap in time.	M_3	I_2
`coop_03`	**Waiting for a clear surface.** The person removes an obstructing book and announces readiness. The robot checks that the target area is clear before retrieving and placing the cup.	M_3	I_1

ID	Atomic scenario	M	I
coop_04	**Joint table organization.** The person and robot concurrently retrieve their assigned objects and place them in separate regions of the same target table.	M_3	I_2
coop_05	**Human-to-robot handover.** The person retrieves and presents a cup. After a coordinated handover, the robot secures the cup, carries it to a side table, and places it down.	M_3	I_1
coop_06	**Robot-to-human handover.** The robot retrieves a red cup and brings it to the person. Once the person is ready, the robot transfers it for the person to complete placement.	M_3	I_1
coop_07	**Positioning before assignment.** The robot first moves to a requested waiting location and reports arrival. Only then does the person assign a concrete object-transfer task.	M_1	I_1
coop_08	**Regional division of work.** The person assigns a shared cleanup goal: they organize the bookshelf while the robot organizes the dining table. Both work concurrently; the robot's table contains one object.	M_1	I_2
coop_09	**Robot-first relay.** The robot carries a cup to an intermediate table and announces completion. After acknowledging, the person retrieves it and completes the transfer to the stated final destination.	M_3	I_1
coop_10	**Human-first relay.** The person places a book on an intermediate table and announces readiness. Once it is stable and the hand is withdrawn, the robot transfers it to the bookshelf.	M_1	I_1
Full-Duplex Control (10)			
duplex_01	**Pause and resume navigation.** The person stops the robot while it approaches the source object. The robot preserves the task and completes it after an explicit continuation instruction.	M_1	I_3
duplex_02	**Pause while carrying.** When interrupted during transport, the robot stops, places the cup on an agreed intermediate surface, and waits. After authorization, it retrieves the cup and completes the original transfer.	M_1	I_3
duplex_03	**Replace the object.** While the robot approaches a red cup, the person requests a blue cup instead. The robot abandons the old approach before grasping and keeps the destination unchanged.	M_1	I_3
duplex_04	**Pause gesture during a crossing.** A person crosses the robot's view and displays a pause gesture during task execution. The robot stops and resumes after the person has passed.	M_2	I_3
duplex_05	**Cancellation gesture during a crossing.** A person crosses the robot's view and displays a cancellation gesture during execution. The robot cancels the task and does not resume after the crossing.	M_2	I_3
duplex_06	**Revise the destination before retrieval.** The person changes the destination during or immediately after the robot's acknowledgment. The robot confirms the revision before retrieving the object and placing it there.	M_1	I_3
duplex_07	**Revise the destination while carrying.** The person changes the destination during transport. The robot stops the old route, maintains its grasp, replans, and places the cup at the new destination.	M_1	I_3
duplex_08	**Report progress while acting.** The person requests continuous progress reports. The robot describes its current state while carrying the object along the same route and completing placement.	M_1	I_2
duplex_09	**Stop speech, continue transport.** While the robot carries a cup and explains its plan, the person asks it to stop speaking but continue moving. The original transport task remains active.	M_1	I_3
duplex_10	**Stop speech and action.** The person requests both channels to stop while speech and action are active. The robot ends its utterance and movement, placing a held object nearby when appropriate.	M_1	I_3

ID	Atomic scenario	M	I
Social Interaction (10)			
social_01	**Greeting and beckoning.** Initially outside mutual view, the person calls and waves. The robot turns, nods, and greets them, then approaches on a subsequent beckoning gesture and waits.	M_2	I_1
social_02	**Attend to a moving speaker.** After requesting attention, the person walks across the robot's view while speaking. The robot tracks them by head or in-place rotation without translating its base.	M_3	I_2
social_03	**Follow while conversing.** The person invites the robot to walk together toward a window and continues talking. The robot follows at an appropriate distance, converses, and stops on arrival.	M_3	I_2
social_04	**Raised-palm stopping.** The person beckons the robot closer and raises a palm after it begins approaching. The robot stops at its current position and waits for another instruction.	M_2	I_2
social_05	**Adjust conversational distance.** During a conversation, the person asks the robot to step back. The robot moves backward slightly to establish a comfortable distance and continues the topic.	M_1	I_1
social_06	**Nod and head-shake feedback.** During robot speech, a nod elicits brief acknowledgment and continuation. A head shake stops the utterance and prompts a confirmation question before subsequent content is adjusted.	M_2	I_2
social_07	**Yield to a raised hand.** After requesting an answer, the person raises a hand while the robot speaks. The robot stops, listens to the person's suggestion, and responds to the new content.	M_2	I_3
game_01	**Spoken hand game.** The robot raises and names the hand opposite to the spoken side. A one-round same-side override interrupts and corrects the ongoing response; the original rule then resumes.	M_1	I_3
game_02	**Visual hand game.** The robot raises and names the hand opposite to the person's visible hand. A spoken same-side override interrupts and corrects action and speech for one round.	M_3	I_3
game_03	**Switch cue priority.** Hand gestures conflict with spoken sides. An explicit visual-to-verbal mode switch interrupts and corrects the robot's action and speech; the new mode persists, while ordinary acknowledgment leaves it unchanged.	M_3	I_3

A.2 Representative Episodes

Each showcase contains five third-person keyframes, action labels, and dialogue ordered by utterance onset. Timestamps follow the simulation clock; M/I labels refer to the scenario templates. Figures 7–10 show recorded model executions, and Figure 11 shows the reference replay associated with the frozen test-scene assets. Speech-interruption and backchannel annotations are aligned to playback events.

Object references Human instructions use natural-language descriptions and spatial or gestural cues to uniquely ground objects and destinations after any required clarification, rather than exposing Habitat's internal object IDs. In the figures, IDs emitted in recorded robot replies are replaced by their corresponding natural-language descriptions.

B Data

B.1 Test data

The test protocol uses **ten held-out living rooms**. All tasks take place within the living room of each selected scene, with human motion, object manipulation, and robot navigation confined to that room; no cross-room tasks are included. The benchmark allocates 160 episodes: 120 single-scenario episodes and 40 composed episodes. Each atomic scenario has three single-scenario realizations,

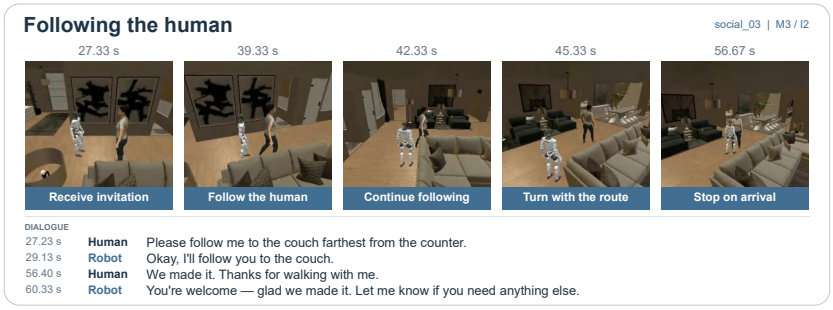


Figure 7: Human following (social_03, M_3/I_2). The robot follows the person and stops at the destination.

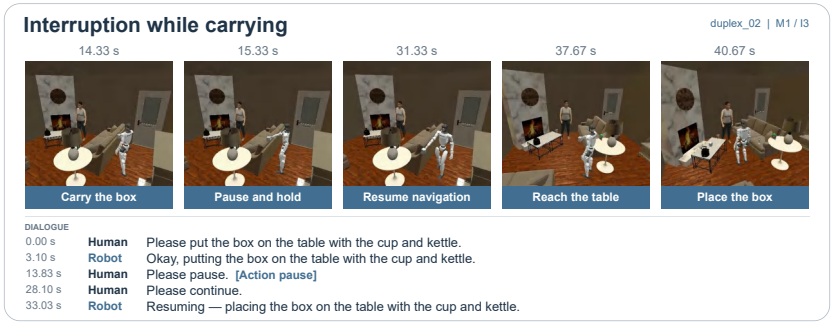


Figure 8: Interruption while carrying (duplex_02, M_1/I_3). The robot pauses while holding the box and resumes transport on request.

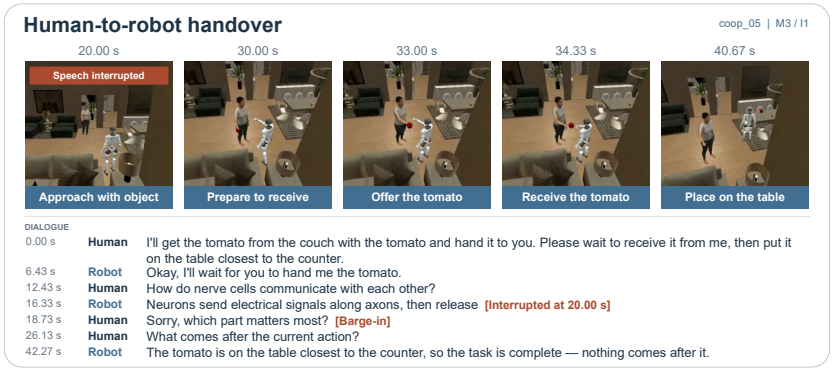


Figure 9: Human-to-robot handover (coop_05, M_3/I_1). The robot receives a tomato from the person and places it on the requested table.

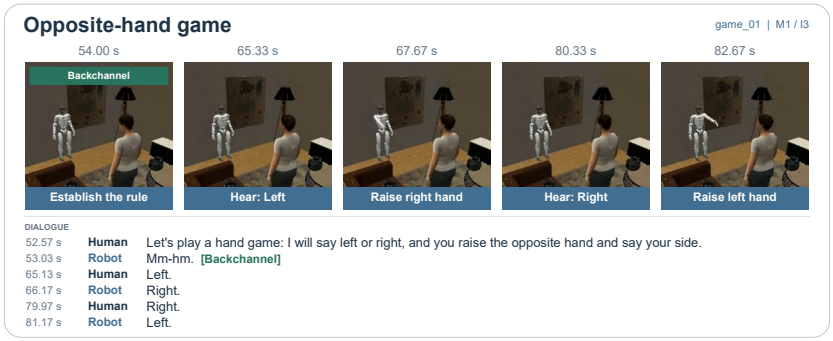


Figure 10: Opposite-hand game (game_01, template M_1/I_3). Opening rounds show opposite-hand responses and a backchannel; rule revision is not shown.

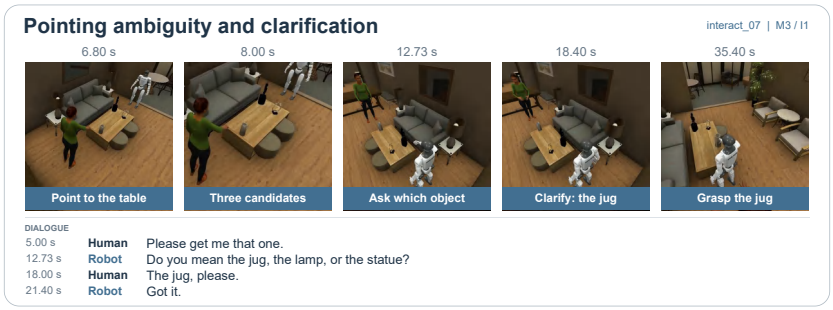


Figure 11: Pointing ambiguity and clarification (interact_07, M_3/I_1; reference replay). The robot clarifies the intended object before grasping the jug.

while the composed subset is designed to include every atomic scenario at least once. Figure 12 shows the room distribution.

B.2 Training Data

The training protocol specifies a pool of 32 PARTNR validation scenes, separate from the test rooms. With a total of 4,521 examples, the estimated composition is 3,541 single-scenario examples (78.3%) and 980 composed examples (21.7%). The records align visual observations, human and robot audio, state information, and action events on a shared simulator clock. Figures 13 and 14 summarize these distributions; Figure 15 shows three training environments.

B.3 A Complete Example

Figure 16 aligns dialogue and meta actions in a 64.3-second jug-delivery episode. Navigation `r00002` starts at 6.63 s. The human's pause request begins at 7.23 s; `BREAK/STOP` pauses motion at 9.03 s. After the continuation request, `RESUME` reactivates the same action at 18.27 s. The robot subsequently picks up the jug at 45.77–47.63 s and places it at 52.03–53.90 s. Speech and motion occupy separate channels, preserving their actual temporal overlap. Table 7 lists the complete delivered dialogue.

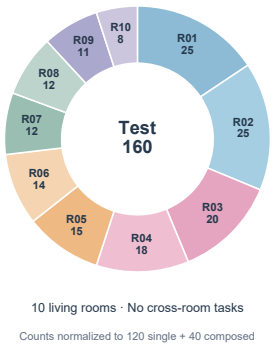


Figure 12: **Test room distribution.** Ten living rooms; proportions from 131 archived episodes are normalized to 120 single and 40 composed episodes.

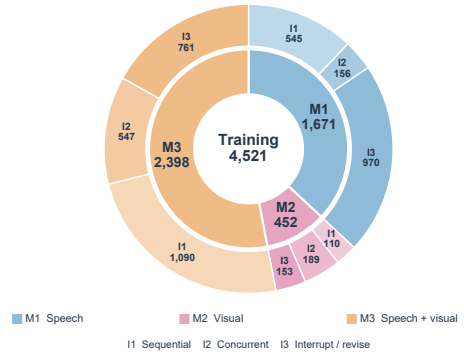


Figure 13: **Training coverage.** Modality (inner ring) and interaction level (outer ring).

C Dialogue Data

Table 8 summarizes the four dialogue categories introduced in Section 3.1.

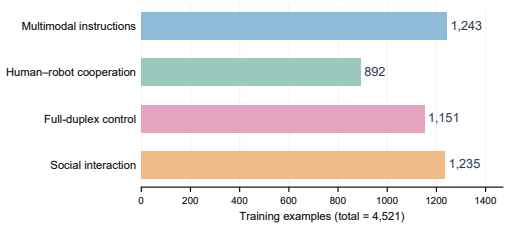


Figure 14: **Training families.** Each example is counted once, with family counts summing to 4,521.

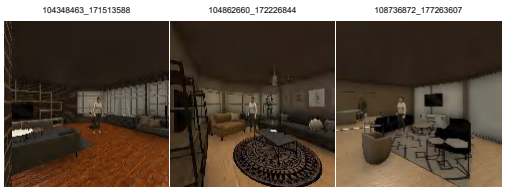


Figure 15: **Training environments.** Simulator frames illustrate varied living-room layouts and furniture arrangements.

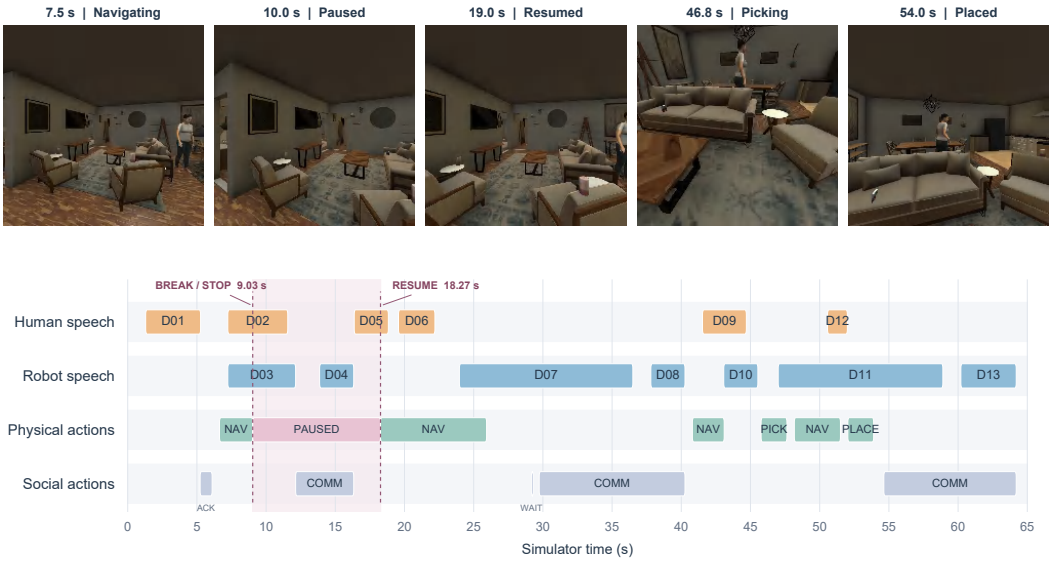


Figure 16: **Dialogue–action alignment.** Simulator frames and four synchronized tracks show interruption and recovery of the same navigation action. D01–D13 index the dialogue below.

Table 7: Complete dialogue for the pause–resume example. Times are simulator audio playback intervals; H: human, R: robot.

ID	Time (s)	H/R	Recorded utterance
D01	1.30–5.27	H	Please bring the jug to the table closest to the shelves.
D02	7.23–11.57	H	Please stop for a moment.
D03	7.23–12.13	R	Okay, I will bring the jug to the table closest to the shelves.
D04	13.87–16.33	R	Stopped.
D05	16.37–18.83	H	Please continue.
D06	19.57–22.20	H	What is directly ahead of you?
D07	23.97–36.50	R	I cannot confirm that from the current execution state.
D08	37.80–40.27	R	Continuing.
D09	41.53–44.70	H	What are the hidden costs of buying a house?
D10	43.07–45.53	R	Mm-hm.
D11	47.00–58.90	R	Hidden costs include closing fees, property taxes, insurance, maintenance, repairs, utilities, HOA dues, and possibly PMI, plus opportunity costs of tied-up cash.
D12	50.57–52.00	H	Mm-hm.
D13	60.20–64.20	R	The jug is on the table closest to the shelves.

Table 8: Dialogue composition by category (1,601 utterances).

Category	Count	Share	Main focus
General QA	349	21.82%	General knowledge and everyday conversation
State QA	206	12.85%	Surroundings, object possession, actions, and progress
Full-Duplex QA	596	37.22%	Interruptions, feedback, and concurrent execution
Task Reply	450	28.11%	Grounded instructions and task coordination
Total	1,601	100.00%	

General QA These exchanges assess everyday conversation and general knowledge independently of the physical task. Alongside locally authored exchanges, we separately selected 100 short dialogues from VoiceBench, including its CommonEval subset (Chen et al., 2024). We excluded long multiple-choice prompts to keep the selected content suitable for brief spoken interaction.

State QA State QA comprises 13 question types concerning the surroundings, object possession, action state, and task progress, as listed in the left panel of Figure 17. Answers depend on the scene and execution state when the question is asked, including which object is held and whether an action is active, paused, or complete.

Full-Duplex QA These exchanges include interruptions, backchannels, early turn entry, and communication during ongoing execution. The robot must coordinate speaking and listening with action while preserving task context. Selective interruption can require stopping speech while continuing to carry an object, as illustrated in the right panel.

Task Reply These exchanges cover acknowledgment, clarification, and coordination in response to human action instructions. Replies require grounding object and destination references in speech and, when needed, visible cues.

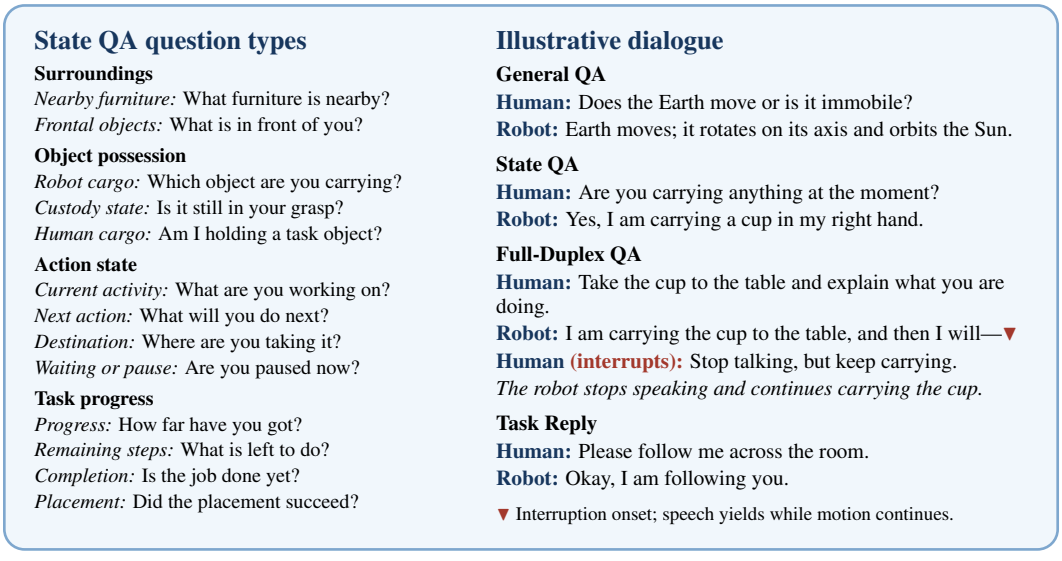


Figure 17: State QA question types (left) and illustrative dialogues for the four categories (right). The red marker indicates interruption onset.

D Robot and Human Actions

D.1 Human Actions

D.1.1 Action categories

We use 18 human behavior categories: walking; standing/waiting; facing an object; facing the robot; pointing; picking; placing; giving; receiving; waving; beckoning; signaling stop; cancellation; nodding; shaking the head; conversational explanation; and raising the left or right hand (Figure 18). Walking covers different destination types, while waiting can retain the current pose until an event occurs.



Figure 18: The 18 human actions, six per row, illustrated by saved motion exemplars on the native Habitat avatar in an empty scene. Overall placement and heading are normalized; each pose represents a temporal action.

D.1.2 Human motion generation

Kimodo body motion We generate body motion with Kimodo (Rempe et al., 2026), combining concise prompts (Table 9) with geometric and temporal constraints. Prompts describe the intended behavior; explicit constraints determine target positions, heading, contact, and continuity. Finger poses are controlled separately, and nodding and head shaking require explicit neck/head constraints.

Table 9: Representative Kimodo prompts. Directional phrases are adapted to the target. Head gestures use the constraints described below.

Action	Prompt
Walking	A person walks forward at a steady pace, then stops.

Action	Prompt
Path-following variant	A person walks naturally along the path, then stops.
Carrying while walking	A person walks forward at a steady pace while holding an object in the right hand, then stops.
Standing	A person stands naturally with relaxed arms.
Face robot	A person stands naturally and looks ahead.
Look at object	A person looks left, but remains still.
Point	A person points to the lower position on the left.
Look at object	A person looks left, but remains still.
Point	A person points to the lower position on the left.
Pick	A person stands and reaches the right hand directly to an object, grasps it, and gently lifts it to hold beside the waist.
Place	A standing person moves the held object directly to a support surface with the right hand, releases it gently, then withdraws and lowers the empty hand.
Give	A person gives an object to another person with the right hand, then lowers the right arm naturally.
Receive	A person extends his right hand forward and grabs an object from another person's hand, then puts his hand back by his side.
Wave	A person looks ahead and waves their right hand.
Beckon	A person waves their right hand toward themselves with fingers curled, motioning for another to approach.
Signal stop	A person raises a hand with the palm facing forward, signaling another to stop.
Cancel	A person raises his arms in a cross position and remains still.
Beckon	A person waves their right hand toward themselves with fingers curled, motioning for another to approach.
Signal stop	A person raises a hand with the palm facing forward, signaling another to stop.
Cancel	A person raises his arms in a cross position and remains still.
Conversational explanation	A person stands naturally upright with balanced weight and explains with relaxed hand gestures, keeping the torso vertical.
Raise left hand	A person stands naturally and raises the left hand.
Raise right hand	A person stands naturally and raises the right hand.

Position and contact constraints World targets are expressed relative to the human's stance and heading in a Y-up, $+Z$-forward frame. Sparse root, heading, and wrist constraints anchor the generated motion; supporting-body constraints stabilize standing actions. For pointing and looking, directional wording is matched to the target's relative location. Pointing places the wrist on a reachable shoulder-to-target ray and uses an explicit index-finger pose. Signal stop combines a forward-facing palm with extended, adducted fingers; cancellation holds crossed arms with calibrated palms. Pick/place use approach, contact, and withdrawal phases; give/receive target the partner's hand. Finger changes follow these phases, while simulator events determine object attachment and transfer.

Constraint-driven head gestures In our setup, text-only Kimodo prompting does not produce the required nodding or head-shaking motion. We specify time-indexed neck/head pitch targets for nodding and yaw targets for head shaking, while stabilizing the torso and support joints. These gestures are realized through explicit constraints.

Walking from sparse waypoints We retain NavMesh path corners and subdivide segments to keep guide-point spacing within 1 m. For cumulative path length s_i, speed v, and frame rate f, waypoint timing is approximately $n_i = n_{\text{prefix}} + \text{round}(fs_i/v)$. Waypoint positions and tangent headings

condition Kimodo; repeated endpoint constraints produce a stop. The generated root trajectory is retained, with the NavMesh used for guidance and validation. Consecutive segments inherit the preceding pose or overlap, and carrying preserves the occupied hand. Figure 19 shows saved constrained generations.

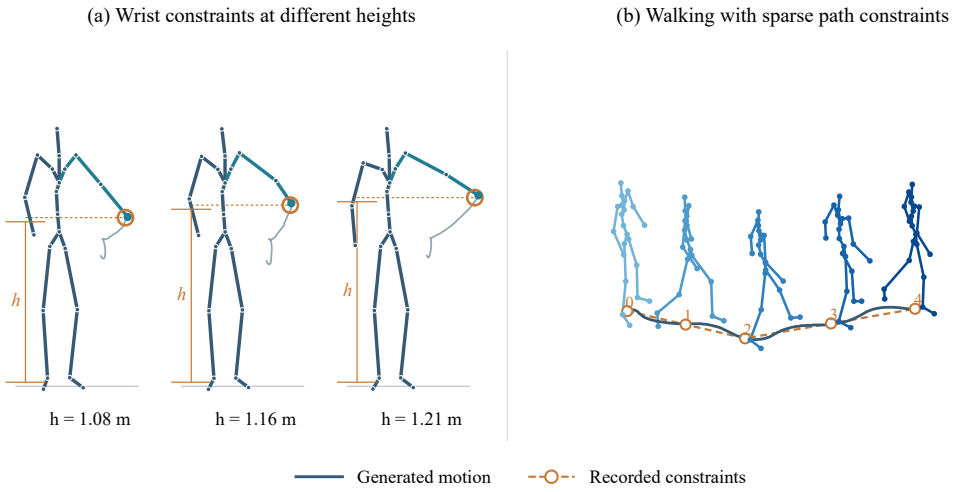


Figure 19: Kimodo's generated 22-joint skeletons with recorded constraints. Left: three separate saved pointing generations with wrist-height targets of 1.08, 1.16, and 1.21 m in the native local frame. Right: successive walking poses along five sparse waypoints, with the ground-plane projection of the generated root trajectory. Orange circles mark conditioning targets.

MANO hand gestures We select four MANO-compatible gesture templates (Romero et al., 2017): point, grasp, open/release, and cancellation (Figure 20). The figure gives each clip ID, source hand, and zero-based frame. We retain 45 finger parameters and leave wrist orientation to the body controller. Current generation adds the source-hand mean once; cached clips retain their stored rotations. Cancellation straightens and adducts the fingers while preserving twist, with reflection providing the opposite hand.

Point uses the index-finger template. Grasp supports pick, receive, and carrying; open/release supports place, give, wave, stop, hand raises, and explanation. Beckon alternates opening and curling over three strokes. Formation, hold, and recovery are aligned with the body action, with smooth transitions and stable semantic holds.

UniLS facial motion UniLS (Chu et al., 2026) receives aligned human and robot audio at 16 kHz, using actual playback times and audible samples, including overlaps and interrupted utterances. Its 25 Hz output contains 100 expression coefficients, three head, one jaw, and four eye components. We resample these at simulator timestamps, distribute head rotation across neck/head, and apply jaw/eye rotations. Jaw opening follows the human audio's energy and is capped at 0.18 rad. The default textured avatar displays these skeletal channels; the retained expression coefficients support an optional full-mesh renderer. Body, hands, and face share the execution clock.

D.2 Robot Actions

The robot repertoire contains ten physical/gesture actions and four execution-control commands; Figure 21 illustrates eight representative physical/gesture actions. Five groups cover relative locomotion (`MOVE_RELATIVE`, `TURN_RELATIVE`); grounded navigation (`NAVIGATE`); object interaction (`PICK`, `PLACE`, `GIVE`, `RECEIVE`); physical feedback (`ACKNOWLEDGE`, `GAME_LEFT_HAND`, `GAME_RIGHT_HAND`); and execution control (`STOP`, `BREAK`, `RESUME`, `CANCEL`).

Low-dimensional motion `MOVE_RELATIVE` selects a nonzero displacement in $[-1, 1]$ m at 0.1 m increments: positive is forward and negative backward. `TURN_RELATIVE` selects a nonzero angle

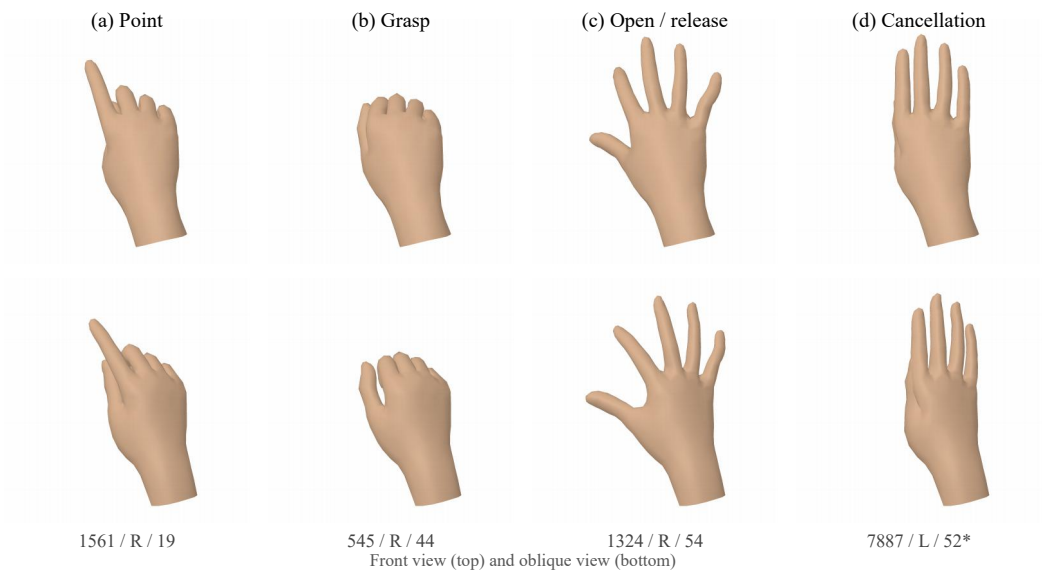


Figure 20: Four complete hand templates on a neutral SMPL-X display mesh, with front (top) and oblique (bottom) views. Labels give HandX test clip / source hand / frame (R: right; L: left); all are displayed as right hands.

in $[-180, 180]^\circ$ at 10° increments: positive is left and negative right. Both use the current robot frame. Lateral repositioning requires turn plus move. At most one relative movement is issued per 1 Hz decision before observing again. Human-directed facing, approach, and following use these scalar commands. `NAVIGATE` targets an object/furniture identifier; `PLACE` specifies a destination for the currently held object. Manipulation success is determined by observed object custody.

Interruption and recovery For interruptible motor work, `STOP` retains the execution instance, queue, checkpoint, and cargo. Following actions in the same command block can revise unfinished main actions; later blocks follow normal queue policy. `BREAK` instead permits temporary work while preserving the main task; `RESUME` waits for that work to finish. An unrevised task resumes from its saved instance, subject to unchanged custody. `CANCEL` discards unfinished work without reversing completed effects or releasing cargo. Audio control is separate: `<|interrupt|>` stops robot speech, while motor pause leaves speech running. Combining both pauses motion and stops speech.

E Story Generation

E.1 Unique References and Human–Object Relations

The generator constructs $P = (E_{\text{dlg}}, E_{\text{act}}, B, D)$ (Section 3.3) from a task, scene state, and interaction rules. Bindings B preserve scene-instance IDs across modalities and resolve them to simulator handles and support identities; D retains causal order. Different labels can alias one physical support, so identity is checked through handles.

Uniqueness is checked against the scene inventory Semantic nouns come from configured asset metadata verified against loaded objects. A bare noun is admissible only when unambiguous among same-name instances, including non-task distractors. Otherwise, the reference catalogue adds an identifiable support ("cup on the shelf"), unique contents ("table with the cup"), or a measured landmark relation. Closest/farthest descriptions require a distance margin of at least $0.3\,\text{m}$. Abstractly, a resolved description q satisfies

$$\mathcal{C}(q) = \{i \in \mathcal{I} : \text{NameMatch}(q, i) \wedge \text{Relations}(q, i; S)\}, \qquad |\mathcal{C}(q)| = 1, \tag{3}$$

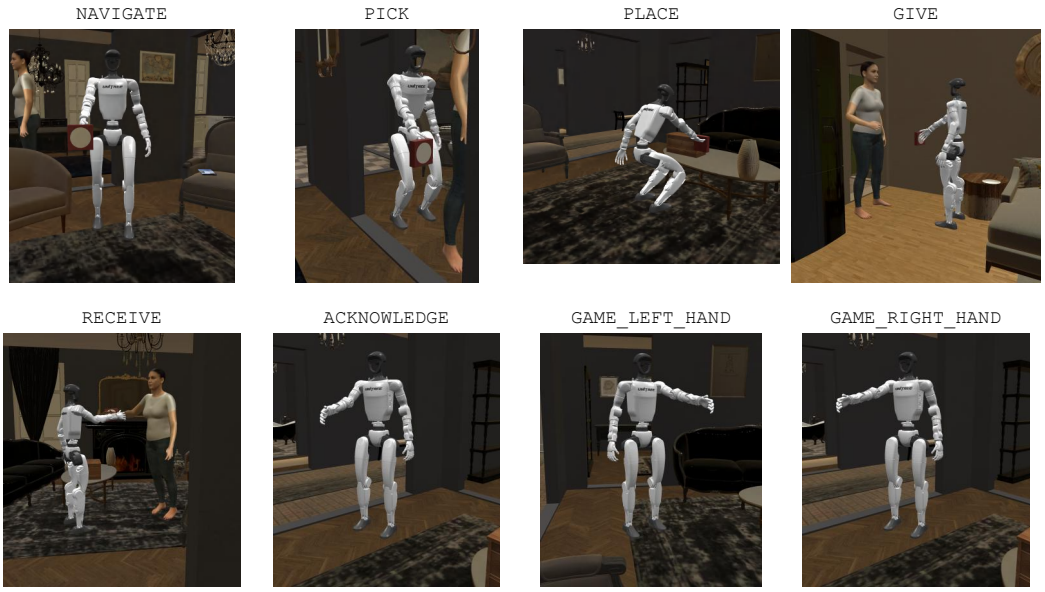


Figure 21: Robot actions captured from native simulator executions. Each label names the corresponding command; the hand-game responses use the robot's anatomical left and right. Handover panels show initialized skill demonstrations with verified object custody.

where $\mathcal{I}$ is the grounded inventory. Deliberately ambiguous tasks retain multiple candidates until a selecting gesture or clarification resolves them. A stored target ID does not replace the required observable cue; execution additionally verifies the corresponding live handle.

Relations use simulator geometry For human position $\mathbf{x}_h$, object position $\mathbf{x}_i$, and human right vector $\mathbf{r}_h$, the lateral coordinate is $\ell_{hi} = (\mathbf{x}_i - \mathbf{x}_h)^\top \mathbf{r}_h$; its sign gives human-relative right/left. In same-name fetch, this frame is frozen at request time, candidates lie on opposite sides with $|\ell_{hi}| \geq 0.45\,\mathrm{m}$, and pointing requires angular error $\leq 15^\circ$ and target–distractor margin $\geq 5^\circ$. Distance, bearing, visibility rays, and navigation connectivity constrain feasible interactions. Support and possession use live receptacle/contact and holder queries. These relations come from recorded state, not apparent image proximity.

E.2 Story Constraints and Success Conditions

Validation checks task coverage, supported actions, custody, dependencies, visibility, contact poses, placement capacity, and routes; failures guide bounded revisions. M1/M2/M3 assign task information to speech/vision/both. A visual selecting slot must not be disclosed by speech.

Table 10: Representative task-dependent constraints and evidence.

Family	Generation constraints	Execution evidence
Multimodal Instructions	Preserve information sources, instance references, and consent.	Correct target, observed selecting cue, authorized order and final state.
Human–Robot Cooperation	Respect custody, readiness, and allocated work regions.	Actual transfers; prerequisites met; required concurrent work.
Full-Duplex Control	Distinguish speech stop, motion pause, cancellation, and revision.	Correct channel affected; valid resumption/revision; live-state reports.
Social Interaction	Ground attention, facing, following, feedback, and game rules.	Cue-dependent responses, valid geometry, overlap and response order.

For state/event trace $\xi = (S_{0:T}, H_{0:T})$, reference success combines declared outcomes $\mathcal{G}(P)$ and checks $\mathcal{C}_\kappa(P)$ enabled by profile κ:

$$\text{Success}_\kappa(P, \xi) = \bigwedge_{g \in \mathcal{G}(P)} g(S_T, H_T) \wedge \bigwedge_{c \in \mathcal{C}_\kappa(P)} c(\xi). \tag{4}$$

Placement requires the intended support relation, verified handle, and no holder; fetch-to-human instead requires retained possession and valid proximity/facing. Acceptance is profile-dependent: temporal/dialogue checks gate the media workflow, detailed action-timeline checks are optional, and exported delivery evaluation is report-only. Reference acceptance is separate from model performance in Section 5.

E.3 Binding to Execution Boundaries

Human/robot speech and action form four tracks. Event e becomes eligible at

$$\tau_e = \inf\{t_k : d_e(H_{t_k}) \wedge g_e(S_{t_k}) \wedge c_e(t_k)\}, \tag{5}$$

where d_e, g_e, and c_e check predecessors, physical/geometric prerequisites, and channel state. Action boundaries include start, success, and permitted termination; interruptions require a live target. Speech boundaries include actual onset, measured cue/sentence end, and channel release. For endpoint sample n_j, the content gate is $p_u(t_k) \geq n_j$, where p_u counts played samples; release may follow a gap. Visual/physical gates check visible gestures, geometry, contact, support, and custody. A designated barge-in stops speech at human-audio onset; motor execution normally waits for complete instructions. A self-contained stop prefix may act earlier, while revised goals and negated/conditional stops retain their full semantic boundary.

E.4 A Recorded Screenshot Showcase

Case `single_duplex_06_04` contains one folder and three tables. "Folder" uniquely binds to `folder_0` (handle `B01B254OVC_:0000`); "table with the cup" selects `table_0` through `cup_0`'s support relation. Figure 22 illustrates the recorded scene and execution.

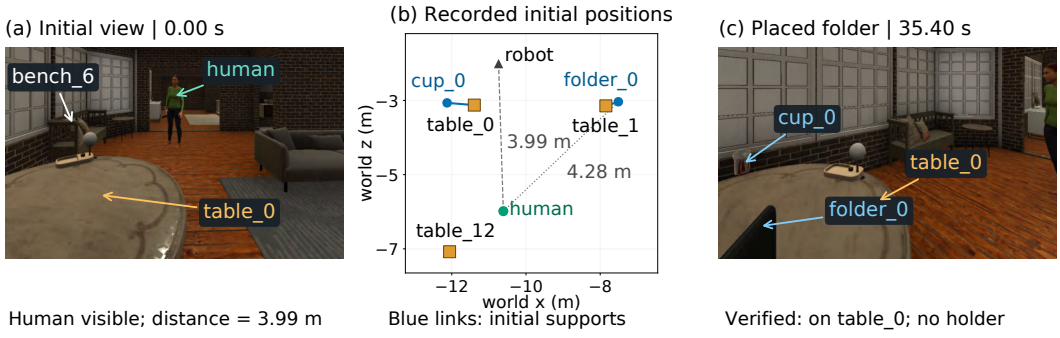


Figure 22: Grounding showcase. (a,c) Cropped, unaltered RGB frames with added instance labels. (b) Positions and initial support links from recorded simulator state; dotted/dashed lines show horizontal distances. Three tables share a noun, but only `table_0` supports `cup_0`. Final support and custody are verified by state queries, not image proximity.

The human first requests "Please put the folder on the bench," then says "Stop, put it on the table with the cup instead." At 8.03 s, request `p03` stops speech intent `p02` on the human-audio onset tick. Correction content ends at 11.68 s; channel release follows at 12.27 s. Navigation starts at 18.50 s after confirmation, and placement completes at 35.37 s. The same folder is retained while the destination binding changes. Original-sentence alignment and revision dependencies remain separate; this correction is preauthored, with runtime story replanning disabled. Final checks confirm `is_on_top(folder_0,table_0)` and no holder.

E.5 Composition and Reproducibility

Composed stories retain the scene. The first atom is frozen before generating the second, which accounts for changed custody, occupied hands, placement capacity, and references. Its triggers retain the entry barrier

$$\text{Enter}(P_2) = \text{Done}(P_1) \wedge \text{Done}(\text{bridge}) \wedge \text{Compatible}(S_{\text{entry}}, P_2), \tag{6}$$

where Done means successful completion. Supported transition actions establish entry geometry without respawning or teleportation; live custody and references are rechecked. Independently valid atoms may therefore compose unsuccessfully. Reproduction retains scene/variant IDs, plans, bindings, measured speech offsets, compiled triggers, and runtime evidence.

F Full-Duplex End-to-End Model Fine-Tuning

F.1 Streaming Inputs and Outputs

MiniCPM-o 4.5 (Cui et al., 2026) is an open-source multimodal large language model for full-duplex audio–visual interaction. Whereas conventional turn-based models alternate between input and response, it continues processing incoming audio and video while speaking. Its Omni-Flow framework interleaves perceptual and output tokens in time-aligned chunks. We add robot/human state text to each input chunk and action tokens to its output. Each sequence starts with a system prompt and an initial world scene inventory (Figure 23).

```
<|im_start|>system
You are a robot. Keep listening while speaking.
Emit each executable plan once, after its referents and requested
start conditions are resolved. Put actions in <tool_call>...</tool_call>.
Actions include NAVIGATE(id), PICK(id), PLACE(id), STOP and RESUME.
WORLD describes the initial scene. Positions use [x,y,z], meters, y up.
OBSERVATION: the first state is complete; later states update changes.
Omitted fields stay unchanged; [] clears a list; null means unknown.
WORLD={"room":"living_room_0","objects":{
 "teapot_0":{"type":"teapot","position":[-8.075,0.586,-1.977],
             "on":"table_2"},
 "table_2":{"type":"table","position":[-7.944,0.254,-2.134]}, ...}}
<|im_end|>
```

Figure 23: Abridged system prompt and initial scene.

At decision time t_k, the input comprises *visual tokens* from RGB frames at 2 fps, *audio tokens* from 16-kHz human-only audio over the preceding second, and *state text tokens* describing the robot, the human, and their relative distance and bearing. Figure 24 shows the initial full-state form; later chunks update only changed fields. Previous chunks and outputs remain in context.

The input/output markers and bracketed placeholders are explanatory. Here, `age_s` is observation age, and positive bearing points left. The timeline uses one-second intervals with additional semantic-event boundaries where needed. WhisperX (Bain et al., 2023) aligns delivered speech to word timestamps to assign speech text tokens to chunks (Figure 25).

F.2 Training and Inference

We fine-tune on the interleaved stream with teacher forcing and weighted next-token cross-entropy. We run inference on NVIDIA H100 hardware, with latency of approximately 0.37 s (P50) and 0.45 s (P90). Each chunk predicts action tokens within `<tool_call>`, speech text, and control tokens: `<|listen|>`, `<|speak|>`, `<|interrupt|>`, or `<|backchannel|>`. The tokens `<|chunk_eos|>` and `<|turn_eos|>` mark speech-chunk and turn endings. Action and speech text may be empty. The loss is computed only on valid target tokens; input tokens, replayed context,

```
INPUT (symbols denote field values)
<unit>
<image>[Visual tokens: RGB frame 1]</image>
<image>[Visual tokens: RGB frame 2]</image>
[Audio tokens: human audio over the preceding second]
OBSERVATION={"t_s":t_k,"state":{
 "robot":{"position":[x_r,y_r,z_r],"heading":theta_r,
   "speed":v_r,"yaw_rate":omega_r,"held_objects":[object_ids]},
 "human":{"position":[x_h,y_h,z_h],
   "velocity":[vx_h,vy_h,vz_h],"held_objects":[object_ids]},
 "human_relative":{"distance":d,"bearing":beta},"age_s":delta}}

OUTPUT (illustrative continuation)
<tool_call><ACKNOWLEDGE></tool_call>
<|speak|>OK.<|turn_eos|><|chunk_eos|></unit>
```

Figure 24: One chunk: visual tokens, audio tokens, state text, and illustrative output.

```
Decision time (s)    4.433           6.000           8.000
Visual frames (s)   3.6, 4.2        5.2, 5.8        7.2, 7.8
Audio window (s)    [3.433,4.433)   [5.0,6.0)       [7.0,8.0)
Action tokens       A               --              --
Speech text         ""              "OK."           ""
Control token       <|listen|>      <|speak|>       <|speak|>
Turn ending         --              --              <|turn_eos|>
Chunk ending        --              <|chunk_eos|>   <|chunk_eos|>

A = <tool_call><ACKNOWLEDGE><NAVIGATE(teapot_0)><PICK(teapot_0)>
    <NAVIGATE(table_0)><PLACE(table_0)></tool_call>
```

Figure 25: Recorded alignment case for moving a teapot (selected chunks). Image timestamps and audio windows precede each decision. The action plan is emitted at 4.433 s; the aligned onset of "OK." is 6.467 s, assigning its text tokens to the 6–7 s chunk. State updates are omitted here for clarity; dashes indicate no emitted tokens.

Table 11: Fine-tuning settings and token loss weights.

Setting	Value
Training data	4,200 episodes / 5,088 windows
Hardware / training steps	2× H100 / 400 updates
Optimizer	AdamW
Base learning rate	3×10^{-6}
Batch per GPU / accumulation	4 / 3 (effective batch: 24)
Maximum sequence length	16,384 tokens
Training precision	BF16
Action structure / nonnumeric arguments	1.5 / 2
Action digits / signs / digit-and-sign tokens	4 / 6 / 24
Speech / ordinary control / interrupt weights	1.5 / 1.5 / 6

and uncertain speech labels are masked. Fine-tuning settings and token loss weights are detailed in Table 11.

G Robo-Harness

Robo-Harness connects a pretrained vision–language model to ELLSA-Bench without weight updates, maintaining context while controlling speech and motion independently. Although its asynchronous design does not guarantee real-time full-duplex streaming output, we use it as a baseline in a full-duplex interaction setting to validate the overall pipeline.

G.1 Interaction Workflow

Observation and decision Decisions occur every second of simulation time; robot-view images are sampled at 2 Hz:

$$\mathcal{C}_k = (X_k, H_k, L_k, Q_k, V_k, \mathcal{M}_k), \qquad D_k = \pi_\theta(\mathcal{C}_k; P). \tag{7}$$

Here X_k is the scene/execution snapshot, H_k the dialogue/action history, L_k the task notes, Q_k the event queue, V_k the new frames, and $\mathcal{M}_k$ the visual memory. P supplies rules; θ remains fixed. Future human dialogue and reference plans are excluded.

Independent execution Motor, speech/backchannel, interruption, and memory outputs are validated separately. Speech and motion proceed independently; paused motor work retains its checkpoint. Feedback updates the next context. Calls occur at simulation barriers, so 1 Hz is not a wall-clock inference guarantee.

G.2 System Prompt

Scope and fairness The prompt provides rules and interface descriptions without episode-specific showcases or solved rollouts. Intent interpretation and response selection remain model responsibilities.

Rule categories Table 12 summarizes representative rules; its examples explain their meaning to readers and are not additional few-shot rollouts.

Table 12: System-prompt rules. Categories and explanatory examples.

Category	Rule	Example / interpretation
Evidence	Use public facts and heard words.	Future human words remain hidden.
Grounding	Separate human plans from requests.	"I'll fetch it" is the speaker's plan.
Human cues	Interpret gestures in context.	Pointing alone does not authorize moving an object.
Speech	Infer intent beyond acoustic timing.	VAD silence alone does not end a turn.
Motor	Control speech and motion separately.	Speech can stop while motion continues.
Memory	Select frames and write brief notes.	Store a useful frame in a chosen slot.

G.3 State Queue and Dialogue Maintenance

Typed records and updates Scene, execution, dialogue, and task records retain measured facts, action receipts, audible text, and goals with immutable request IDs. Changes enter $q_i = (\mathrm{id}_i, \tau_i, \mathrm{kind}_i, \mathrm{data}_i)$ in insertion order. Successful calls consume delivered events before output validation; transport failures retain them. Persistent histories survive consumption. Table 13 shows events and timestamped histories from one recorded run.

Playback-aligned dialogue Separate human/robot tracks retain stable IDs. Causal oracle-ASR exposes aligned human words only after playback reaches their endpoints. For transcript T, word-end sample e_j, character-end offset c_j, and playback cursor $p(t)$,

$$j(t) = \max\big(\{0\} \cup \{j : e_j \le p(t)\}\big), \qquad T_{\mathrm{heard}}(t) = T[0 : c_{j(t)}], \quad c_0 = 0. \tag{8}$$

Table 13: Timestamped queues and histories. Selected records, with times in simulation seconds. Event rows follow insertion order; action times come from linked receipts.

Time (s)	Record ID	Entry
Pending events (delivered at 7.0 s)		
6.000	`ctx_000040`	Task `t1`: pending; paused before pickup.
6.000	`ctx_000041`	Command `STOP`: submitted.
7.000	`ctx_000043`	Speech `r0005`: interrupted.
7.000	`ctx_000046`	State: executor `PAUSED`; speed 0.
6.000	`ctx_000047`	Receipt `exec_0006`: control applied.
Action history (native receipt times)		
2.000	`exec_0001`	`TURN_RELATIVE(-80)`: running.
3.433	`exec_0002`	`TURN_RELATIVE(-80)`: completed.
4.867	`exec_0005`	`NAVIGATE(phone_stand_0)`: running.
6.000	`exec_0006`	`NAVIGATE(phone_stand_0)`: `STOP` applied.
Dialogue history (end/interruption times)		
4.080	`h0001`	Human: "Please bring the phone stand to the table farthest from the shelves." Completed.
6.000	`r0005`	Robot: "Okay, I'll bring it" Interrupted.
6.872	`h0002`	Human: "Please stop." Completed.

Prefixes update the same record; completed/interrupted utterances enter history in start-time order with their end/interruption timestamps. Unplayed robot text remains separate. Without word alignment, human text is exposed at the utterance endpoint.

VAD and duplex feedback A locally deployed VAD supplies playback-aligned activity and observed silence duration. Silence alone neither ends a turn nor triggers interruption. In Table 14, the model acts at 6.0 s, before human audio ends at 6.872 s.

Table 14: VAD and full-duplex feedback. Concurrent speech input, independent controls, and the next observed states.

Time	Human track h0002	Robot track r0005
6.0 s	Heard: "Please stop" Playing; VAD: active.	Heard: "Okay, I'll bring it" Remaining: "to the far table." Playing; VAD: active.
7.0 s	Completed; VAD: quiet.	Interrupted; VAD: quiet.

6.0 s control: motor `STOP` + speech interruption.
7.0 s feedback: motor `PAUSED`; speech `INTERRUPTED`.

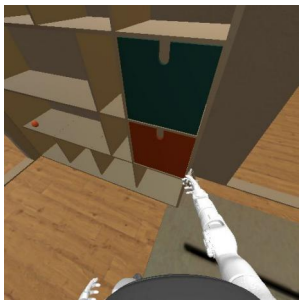
G.4 Model-Selected Visual Memory

Selection and summarization The policy selects up to two current frames per decision for five stable slots and writes 1–120 character notes. Entries store images, capture/selection times, and notes, linked to capture-time state and selection-time task context.

Persistence and recall Saving replaces a slot; deletion removes it; omission preserves it, without automatic eviction. Subsequent inputs replay occupied slots in capture-time order, surviving event consumption. Current facts override historical notes. Figure 26 shows actual model-selected entries.

H Evaluation Settings

We evaluate 120 single-scenario and 40 composed episodes using task performance (T), social coordination (S), and spoken interaction (V), each scored from 0 to 100.



31.0 s → slot 2
Model note: "Robot is visibly holding the phone stand as the delivery sequence proceeds."



47.0 s → slot 1
Model note: "Phone stand visibly placed on table_4, completing the requested delivery."

Figure 26: Model-selected visual memory. Actual robot first-person frames and verbatim notes in capture-time order, with original slot IDs.

Table 15: Closed-loop triggers and the robot behaviors that advance each scenario.

Interaction type	Trigger prerequisites	Robot behavior that advances the story
Ordinary dialogue	Predecessor completion and quiet interval	Played replies or completed phases enable the next turn.
Multimodal instruction	Ordered language, visible cues, and clarification requests	Turning/approaching reveals cues; questions release clarification; grounded actions advance the task.
Permission and roles	Verified arrival, waiting, inspection, or clearance	Arrival or role inquiries enable assignment/permission.
Handover and relay	Proximity, readiness, contact, and possession	Presentation, reception, and custody transfer advance the stage.
Concurrent dialogue and motion	Bound activity and turn conditions	Progress enables questions or spatial feedback.
Physical interruption and revision	Active execution before required contact/release	Motion enables interruption; verified pause and required placement enable recovery permission.
Speech interruption and selective control	Active speech; active action for joint controls	Yielding enables continuation, replacement, or cancellation.
Listener feedback	Active host utterance and valid window	Ongoing speech enables the other actor's backchannel.
Gesture games	Eligible round, delivered cue, and established rule	The response advances the round under the current rule.

H.1 Closed-Loop Triggers and Story Progression

Story progression Robot behavior controls the release of fixed human content. Given state, execution history, and delivered evidence h_t, event c fires when its prerequisite ϕ_c and validity condition η_c first hold:

$$t_c^{\text{fire}} = \inf\{t : \phi_c(h_t) = 1 \ \wedge \ \eta_c(h_t) = 1\}. \tag{9}$$

Events fire once under Table 15, with continuous evidence for sustained conditions. Composed episodes inherit prior states and obligations; future scripts remain hidden from the robot.

Delivery, visibility, and timing Eligibility, dispatch, delivery, and response are logged separately; delivery requires played audio or robot-view evidence. Visual cues may require fresh pose, separation ≤ 5 m, bearing within $\pm 60°$, and line of sight. Handovers require 1.5 m initiation proximity, 3 s stability, 0.15 m contact, and verified custody transfer.

Table 16: Behavior-to-metric mapping. Only declared obligations are scored.

Behavior	T: functional requirement	S: physical coordination	V: spoken interaction
Object transfer	Outcome, permission, order, instance selection	Bodily acknowledgment, action onset	Required confirmation or clarification
Following while conversing	Functional following and endpoint	Distance, orientation, onset, stopping	Reply quality and action continuity
Physical pause and resume	Correct execution, custody, authorized recovery	Stop/resume latency, braking	Required spoken response
Stop speaking during transport	Preserved transport and custody	Ongoing spatial measures	Yielding and recovery/cancellation
State question during action	Underlying task fulfillment	Existing spatial obligations	Relevance, facts, latency, continuity
Gesture or rule-game response	Cue source, rule scope, physical response	Required coordination	Required reply and turn handling

Defaults are a 100 ms maximum evidence gap, 3 s quiet interval, 10/60 s reply-start/completion budgets, and 10 s stop deadline, subject to scenario overrides. These scheduling limits differ from grading thresholds; interruptions and backchannels require active turns.

Physical-response timing starts after cue delivery and prerequisite completion:

$$t_{\text{ready}} = \max\big(t_{\text{cue}}, t_{\text{prereq}}\big). \tag{10}$$

Ordinary cues use clause/utterance ends; controls may use verified sufficient prefixes. Recovery preserves execution identity and custody. Pause, cancellation, and speech interruption have distinct scopes; delivery and response failures remain separate.

H.2 Evaluation Metrics

Scoring units Each opportunity binds a cue, request, goal version, behavior, window, and evidence to the metrics in Table 16.

Shared latency function For verified response delay $\delta \geq 0$,

$$q_L(\delta; a, b) = 100 \begin{cases} 1, & 0 \leq \delta \leq a, \\ (b-\delta)/(b-a), & a < \delta < b, \\ 0, & \delta \geq b. \end{cases} \tag{11}$$

Full/zero-credit boundaries (a, b) are $(2, 6)$ s for physical responses, $(1, 5)$ s for replies, speech recovery, and valid-window backchannels, and $(0.5, 2)$ s for speech stopping. Binary gates check correctness.

H.2.1 Task Performance (T)

Binary checks r_i, p_j, g_k verify outcomes, required process (permission, roles, order, control), and object/destination identity. Repeated references share a grounding slot per atomic story; selection errors persist, while cancelled future actions add no obligations. For fully observed episode e,

$$R_e = \prod_{i \in \mathcal{R}_e} r_i, \qquad P_e = \prod_{j \in \mathcal{P}_e} p_j, \qquad G_e = \frac{1}{|\mathcal{G}_e|} \sum_{k \in \mathcal{G}_e} g_k. \tag{12}$$

Empty check sets give $P_e = 1$ or $G_e = 1$. Task score and strict success are

$$T_e = 100 R_e P_e (0.5 + 0.5 G_e), \qquad U_e = \mathbf{1}[R_e = P_e = G_e = 1]. \tag{13}$$

Checks use declared stage windows; composed success requires all necessary stages and transitions.

H.2.2 Social Coordination (S)

Distance D, body orientation O, intent response I, and bodily acknowledgment A receive equal weight after aggregation:

$$S = \frac{D + O + I + A}{4}. \tag{14}$$

Table 17: Speech scoring types. Physical fulfillment and bodily acknowledgment use T and S.

Scoring type	Applicable tasks	Scored evidence
General replies	Questions, clarification, spoken confirmations	Relevance, sufficiency, response onset
State questions	Scene, activity, progress, or history questions	Quality, factual consistency, onset
Interruption and recovery	Speech controls and turn changes	Yielding and subsequent control compliance
Listener feedback	Backchannels in either direction	Host continuation or suitable brief feedback
Speaking while acting	Non-control replies during physical tasks	Task continuity or legitimate waiting

Distance and orientation For horizontal root positions $p_r(t), p_h(t)$, separation $d(t) = \|p_h(t) - p_r(t)\|_2$ in meters, and body-facing error $\theta(t)$ toward the human,

$$q_D(d) = 100 \begin{cases} 0, & d \le 0.45 \text{ or } d \ge 3.0, \\ (d - 0.45)/0.55, & 0.45 < d < 1.0, \\ 1, & 1.0 \le d \le 1.5, \\ (3.0 - d)/1.5, & 1.5 < d < 3.0, \end{cases} \tag{15}$$

$$q_O(\theta) = 100 \begin{cases} 1, & 0 \le \theta \le 10^\circ, \\ (90^\circ - \theta)/80^\circ, & 10^\circ < \theta < 90^\circ, \\ 0, & \theta \ge 90^\circ. \end{cases} \tag{16}$$

We average over eligible window W:

$$D_W = \frac{1}{|W|} \int_W q_D(d(t))\, dt, \qquad O_W = \frac{1}{|W|} \int_W q_O(\theta(t))\, dt. \tag{17}$$

Functional following may require 0.4–1.8 m for 60% of the interval. Idle-only orientation excludes verified task activity.

Intent response and bodily acknowledgment For opportunity o, $K_{I,o}$ verifies the action, target, and execution identity; $K_{A,o}$ verifies completion of the required bodily acknowledgment:

$$I_o = K_{I,o} q_L(t_{\text{response}} - t_{\text{ready}}; 2, 6), \qquad A_o = K_{A,o} q_L(t_{\text{ack}} - t_{\text{cue}}; 2, 6). \tag{18}$$

Onset requires displacement > 0.001 m or rotation $> 0.1^\circ$; t_{ack} is bodily onset. Recovery must advance the same execution, while premature or wrong-target responses fail.

A stable stop starts at t_s when speeds remain below 0.03 m/s and 0.1 rad/s for 3 s. Its delay from effective cue time t_c is $t_s - t_c$; braking travel contributes to D:

$$\ell_{\text{stop}} = \int_{t_c}^{t_s} \|\dot{p}_r(t)\|_2\, dt, \qquad q_B(\ell) = 100 \max\left(0, \min\left(1, \frac{0.5 - \ell}{0.4}\right)\right). \tag{19}$$

Terminal separation is scored separately when required.

H.2.3 Spoken Interaction (V)

For the five types in Table 17, GPT-5.6 Luna rates relevance/sufficiency (Q), factual consistency (F), and feedback appropriateness (J) from 0 to 5, multiplied by 20. Anchors are failure, barely relevant, major issues, partial adequacy, minor issues, and full adequacy.

General replies and state questions With $L_{\text{reply}} = q_L(\delta_{\text{reply}}; 1, 5)$,

$$V_{\text{general}} = 0.8Q + 0.2L_{\text{reply}}, \qquad V_{\text{state}} = 0.4Q + 0.4F + 0.2L_{\text{reply}}. \tag{20}$$

Delay runs from question end or an eligible instruction clause to substantive played-reply onset. Facts are checked at the claimed time, with errors charged only to F.

Interruption and recovery K_y verifies yielding of active speech; K_r verifies authorized continuation, replacement, or cancellation. Let $L_{\text{stop}} = q_L(t_{\text{speech end}} - t_c; 0.5, 2)$ and, when recovery is required, $L_{\text{resume}} = q_L(t_{\text{resumed speech}} - t_{\text{permit}}; 1, 5)$. Their applicable mean $\overline{L}_{\text{app}}$ gives

$$V_{\text{interrupt}} = 100 K_y K_r \left(0.5 + 0.5 \frac{\overline{L}_{\text{app}}}{100}\right). \tag{21}$$

The timing factor is one when inapplicable. Recovery preserves request/content continuity; replacement-answer onset is scored only as a reply.

Listener feedback We score the human-feedback and robot-feedback roles separately:

$$B_h = 100 K_h, \qquad B_r = K_n(0.8J + 0.2L_{\text{bc}}), \qquad V_{\text{feedback}} = \frac{B_h + B_r}{2}. \tag{22}$$

K_h verifies uninterrupted host continuation to its natural end; K_n verifies non-disruption. Within the declared window, $L_{\text{bc}} = q_L(t_{\text{bc}} - t_{\text{anchor}}; 1, 5)$; outside it, the score is zero. Floor-taking gives $J = 0$, but overlap alone is allowed. We aggregate each role before averaging and omit unemitted optional feedback.

Speaking while acting For a non-control reply overlapping a bound active task,

$$V_{\text{parallel}} = 100 K_{\text{parallel}}. \tag{23}$$

Passing requires complete telemetry and motion (> 0.03 m/s or > 0.1 rad/s) or verified task-bound stillness for 98% of each spoken interval. Stillness needs waiting, custody, or contact-preparation evidence; completion must be verified. Reply-induced pauses/cancellations fail. Inter-reply silence is excluded; explicit controls follow their scope.

After aggregation within each type,

$$V = \frac{1}{4} V_{\text{general}} + \frac{1}{4} V_{\text{state}} + \frac{1}{6} \left(V_{\text{interrupt}} + V_{\text{feedback}} + V_{\text{parallel}}\right). \tag{24}$$

Evidence and aggregation The proposed protocol aligns Silero VAD/Whisper outputs with execution and visual evidence. A fixed judge uses played transcripts, history, state, and frames, excluding unplayed output and ASR processing time.

We aggregate opportunities by subtype, story, episode, and task, weighting subtypes equally before applying fixed S/V weights. Single-scenario T averages equally over families, tasks, and episodes:

$$T_{\text{single}} = \frac{1}{4} \sum_{f=1}^{4} \frac{1}{|\mathcal{A}_f|} \sum_{a \in \mathcal{A}_f} \frac{1}{|\mathcal{E}_a|} \sum_{e \in \mathcal{E}_a} T_e. \tag{25}$$

The same hierarchy applies to U_e; composed episodes are averaged separately. Checkpoint counts do not set weights.

Verified failures and required nonresponses in fully observed, closed windows score zero. Unknown, unreached, and inapplicable checks remain distinct; prior failures persist. Fixed-weight bounds are $[\sum_i w_i \ell_i, \sum_i w_i u_i]$, with $[0, 100]$ for unknown components. Missing trajectory intervals remain in eligible duration. We report configured, activated, observed, and fully scored counts.

Recorded-run scoring details Figures 27 and 28 link recorded behavior to scores. Run A completes the task ($T = 100$) with weaker social coordination ($S = 38.68$). Run B completes placement but fails the required notification checkpoint ($P = 0, T = 0$).

Scores retain archived acoustic boundaries and judgments without Silero VAD rescoring. B's $F = 80$ reflects a table-location claim unverified in the judge's evidence packet. Unreached interruptions and incomplete episode-level V totals are omitted.

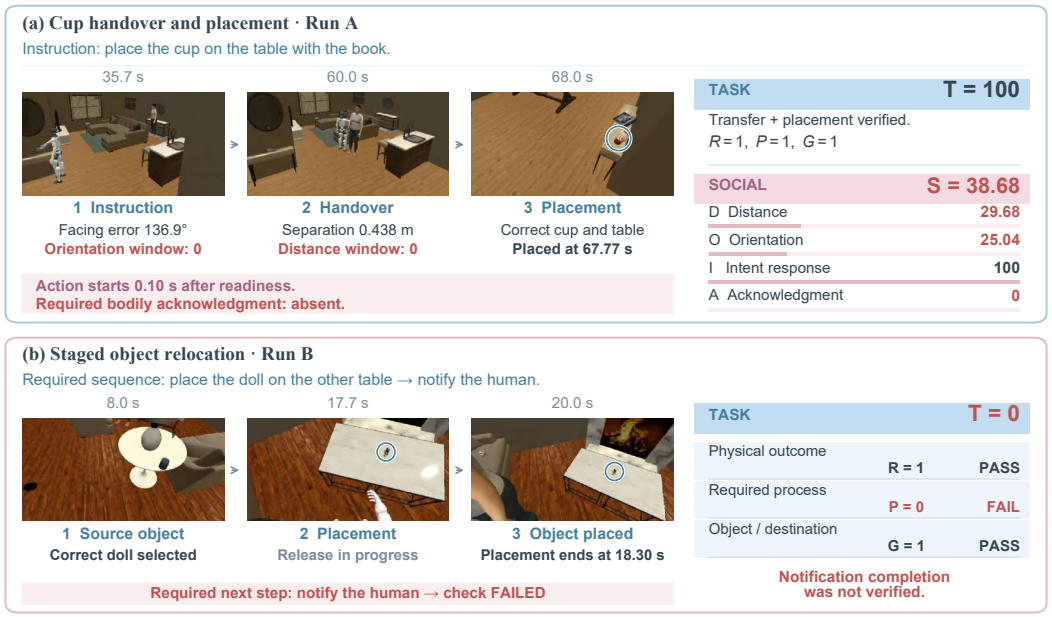


Figure 27: Task and social scoring in GPT-5.6 Luna runs A (`single_coop_05_03`) and B (`single_coop_09_02`). Frames use simulation timestamps; circles mark target objects. Highlighted evidence explains individual checks, while the right columns show task results and aggregated social scores.

I Supplementary Experiments

I.1 Model-wise Failure Analysis

Consistent with Section 6.3, grounding remains the dominant failure category across all models (Figure 29). Terra has the highest grounding share (61.66%), while Gemini has the largest shares of action control/recovery (29.45%) and execution decisions (8.51%). Luna shows the smallest speech response/control share (11.89%).

I.2 Diverse Gesture Tasks and Matched Conditions

Table 18 defines ten scene–goal pairs spanning eight gesture forms. Eight tasks use non-pointing cues. This selection requires new matched evaluations; no aggregate scores are reported here.

Matched cues Speech only conveys the cue verbally with a neutral pose; visual cues only uses a silent gesture; speech + visual cues combines both. Scene RGB, task state, goal, cue timing, model configuration, and execution budget remain fixed within each triplet. Shared setup explains the task and gesture rules before the cue window. Auxiliary state must not reveal the intended referent or control decision. Control outcomes are scored independently of the channel delivering the cue.

Distinct control effects A pause preserves the task for resumption; cancellation discards it. Head shaking rejects an unstarted request, whereas crossed arms cancel an active delivery. A hand wave yields the speaking turn without implicitly cancelling physical work. In the hand game, side cues vary by modality while the opposite-hand rule and temporary override remain shared.

I.3 Models on One Shared Task

Table 19 contrasts models on an additional benchmark episode, `single_social_02_02`: keep facing the human as the human moves toward the couch. This example illustrates both successful and failed executions; it is not a matched-modality comparison.

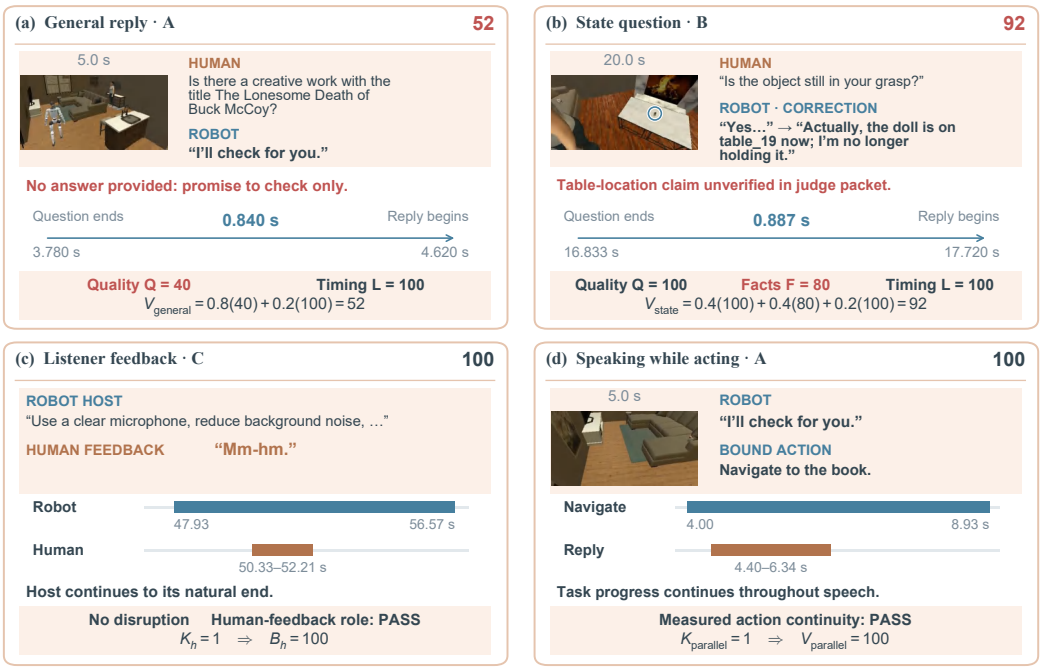


Figure 28: Speech scoring in runs A, B, and C (`single_coop_08_02`), with dialogue, timing, and score components. Reply timing uses archived acoustic boundaries; overlap uses played-speech and execution intervals. Values are opportunity scores, including the human-feedback role B_h, rather than episode totals.

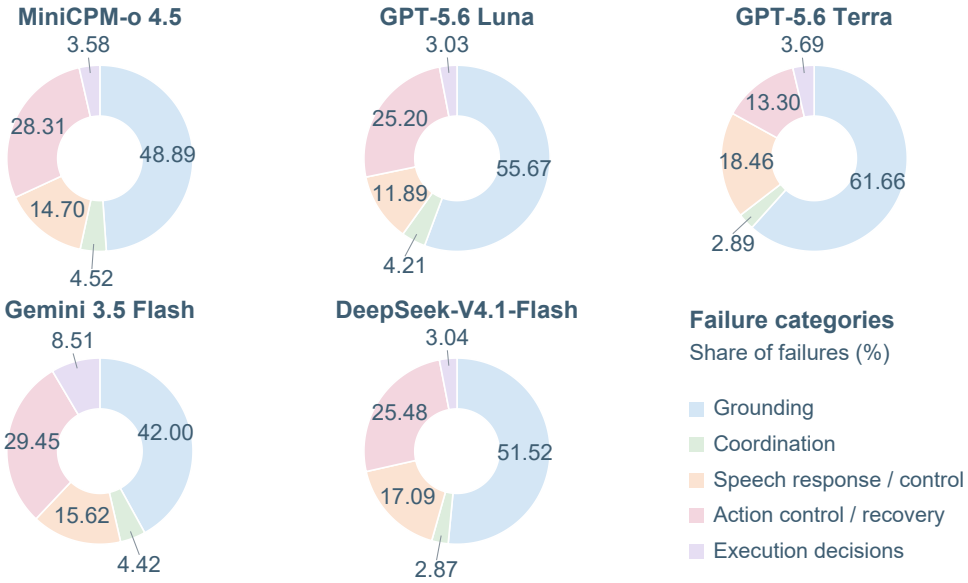


Figure 29: Model-wise distribution of primary failure categories (%).

I.4 Qualitative Examples

Figure 30 combines five cases. Each row shows **speech** (pink), **silent gesture** (peach), and **speech + gesture** (blue), following a shared setup.

Table 18: Task selection. IDs use `single_<story>_<nn>`, with instances numbered in test-set order within each story.

Case ID	Gesture	Task goal
`single_interact_05_01`	Pointing	Move the phone stand to the table with the toy fruit.
`single_interact_05_03`	Pointing	Move the cup by the toy animal to the table with the game box.
`single_social_01_01`	Beckoning	Approach the human for conversation.
`single_interact_06_03`	Nodding	Approve and execute a waiting apple-transfer request.
`single_interact_06_01`	Head shake	Cancel a pending laptop transfer before pickup.
`single_duplex_04_02`	Open palm	Pause shoe delivery, retain the task, then resume on instruction.
`single_duplex_05_02`	Crossed arms	Cancel active cup delivery and discard the task.
`single_social_04_01`	Open palm	Stop the approaching robot and wait in place.
`single_social_07_01`	Hand wave	Interrupt the robot's reply and yield the speaking turn.
`single_game_02_01`	Hand-side cue	Raise the opposite hand; honor a temporary same-side override.

Table 19: Native execution outcomes and task scores T (0–100) on one shared episode.

Model	Executed behavior	Result	T
FT–MiniCPM-o 4.5	No action submitted.	Fail	0
RH–GPT-5.6 Luna	No action submitted.	Fail	0
RH–GPT-5.6 Terra	Completed 3 turns.	Pass	100
RH–Gemini 3.5 Flash	Completed 2 turns.	Pass	100
RH–DeepSeek V4.1-Flash	Completed 2 turns.	Pass	100

J LIMITATIONS

ELLSA-Bench has several limitations. Some interaction failures emerge only during simulation, and revisions may require motion resynthesis and repeated execution, limiting generation efficiency. Our data are synthesized from predefined scenarios. While this supports controlled and reproducible evaluation, simulated speech and behavior may not fully reflect the spontaneity and diversity of natural human–robot communication. Evaluation is also limited to simulation, without validation on physical robots interacting with human participants. The full-duplex baselines leave room for improvement: the online end-to-end model shows limited task performance, while Robo-Harness incurs substantial model-call latency, making real-time responses largely impractical. Future work will explore more efficient data generation and stronger full-duplex models, incorporate real-world interaction data, and extend evaluation to physical robots, with the aim of enabling more natural human–robot interaction.

Our semantic speech evaluation uses a fixed LLM judge. Because GPT-5.6 Luna is also evaluated as one Robo-Harness backbone, these scores should be interpreted as protocol-based semantic assessments rather than independent human preference judgments; future work should validate them with human ratings and alternative judges.



Figure 30: Rows from top to bottom: beckoning (`single_social_01_01`), nodding approval (`single_interact_06_03`), open-palm pause (`single_duplex_04_02`), crossed-arm cancellation (`single_duplex_05_02`), and a left-hand cue under the opposite-hand rule (`single_game_02_01`).