AIRIS: BENCHMARKING SOCIAL EMBODIED INTELLIGENCE IN EGOCENTRIC STREAMING VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents operating around people must not only recognize social cues, but also decide when the available evidence is sufficient, whether to intervene, and how to respond without violating social norms. Existing benchmarks typically evaluate these abilities in isolation or with access to completed videos, obscuring where an online social decision fails. We introduce AIRIS, a causal streaming benchmark that formulates social embodied intelligence as a time-anchored chain of Attention, Intention, Reaction, and Interaction (A-I-R-I). AIRIS contains 8,000 expert-curated questions grounded in 357 egocentric clips across 11 context categories and 36 capability dimensions. At each query, models receive only the observation prefix available at that time; decision-aware metrics reward calibrated waiting and complete reactions while penalizing premature commitment. Evaluating a pool of 45 proprietary and open-source general-purpose, embodied, streaming-video, and omni-modal models reveals fragmented and temporally miscalibrated competence. General-purpose VLMs perform strongly in cue grounding and intent recognition, whereas embodied specialists are more competitive on aggregate Reaction; nevertheless, the best Intent Anticipation Score is only 0.237, category-level False Commitment Rates range from 0.26 to 0.51, and no model exceeds 0.147 Reaction Joint-EM. These findings characterize the perception–execution gap as a failure of temporal calibration and reaction composition. Ultimately, reliable social embodied intelligence requires models to actively calibrate anticipation to observable evidence, composing perception, inference, and action into seamless responses that align with human social norms.

1 INTRODUCTION

An embodied agent sharing space with people must do more than recognize objects and execute commands. When a person glances toward an object and begins a gesture, the agent must determine whether the cue concerns it, infer what the person intends, and decide whether to act immediately, wait for more evidence, or refrain from acting. Even a physically feasible response can be socially mistimed, intrusive, or incomplete. Such decisions arise continually in collaborative homes, public spaces, and assistive settings, where successful behavior depends on both task competence and sensitivity to people (Puig et al., 2024; Ramnauth et al., 2024; Zhou et al., 2023). We conceptualize the ability to perceive, interpret, and act appropriately within these evolving interpersonal situations as *social embodied intelligence* (Fong et al., 2003; Dautenhahn, 2007; Lee et al., 2024b).

To date, the community has successfully operationalized distinct facets of this intelligence through specialized, domain-specific tasks, as shown in Table 1. For instance, research in streaming and proactive video evaluates models under continuous input, emphasizing strict observation prefixes and timely intervention (Lin et al., 2026; Niu et al., 2025; Lei et al., 2026; Zhang et al., 2025c). Meanwhile, multimodal social benchmarks dive deep into interpreting gaze, gesture, intent, and latent mental states in natural human interactions (Zadeh et al., 2019; Kong et al., 2026; Lee et al., 2024a; Shi et al., 2025). In parallel, norm-aware embodied evaluations assess whether proposed actions are permissible and compatible with hidden constraints during ongoing human activities (Rezaei et al., 2025; Fang et al., 2026; Zhao et al., 2026; Zhang et al., 2026).

However, these evaluations typically treat social interaction as a set of isolated slices rather than a continuous and interdependent causal chain. Existing benchmarks tend to isolate the interpretation

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.

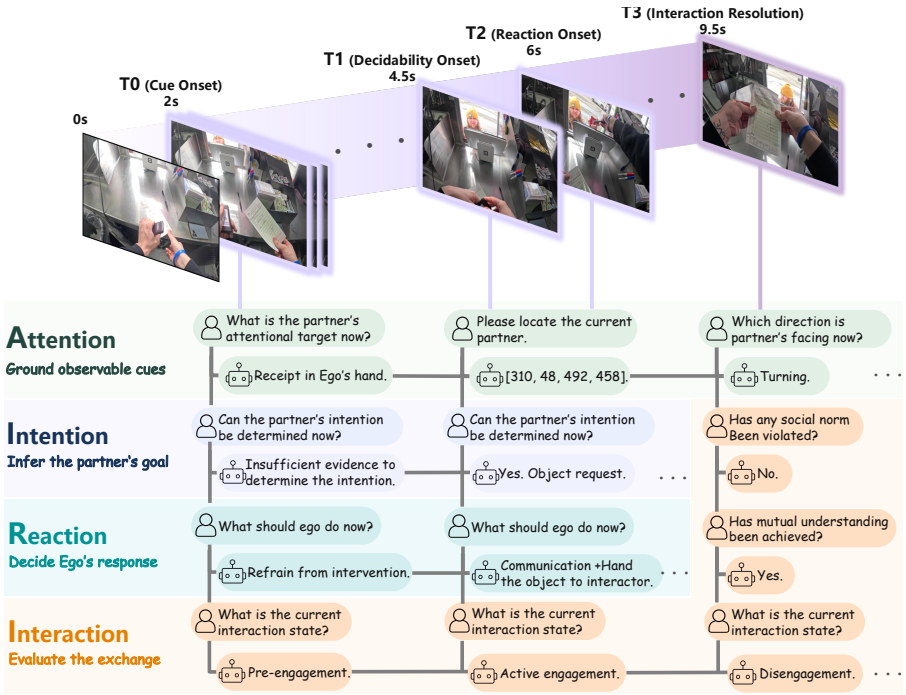


Figure 1: **AIRIS: evaluating social embodied intelligence over time.** An egocentric episode illustrates four linked capabilities: grounding social cues (Attention), inferring the partner's goal (Intention), deciding Ego's response (Reaction), and evaluating the exchange (Interaction). Each query uses only observations up to the query time, without access to future evidence.

of human intent from the subsequent selection of an appropriate ego-agent action. Consequently, when an agent delivers an incorrect final response, tracing the root cause of the failure becomes nearly impossible. Without a unified process, it remains unclear whether the model missed an early non-verbal cue, misunderstood the underlying intent, or simply chose an inappropriate moment to intervene. While prior datasets excel in specific dimensions, none span the complete perception-to-action spectrum required for social embodiment within a single episode (Table 1).

To address this mismatch, we introduce **AIRIS**, a streaming benchmark that evaluates social embodied intelligence as a multi-stage decision process spanning **A**ttention, **I**ntention, **R**eaction, and **I**nteraction, as shown in Figure 1. Specifically, *Attention* evaluates whether the model can localize the interaction partner from the first-person stream and integrate explicit multimodal cues, including gaze, gestures, body posture, emotional states, and speech, into an observable focus. Building on this evidence, *Intention* infers the partner's underlying goal and identifies the earliest moment this goal becomes decidable. *Reaction* then shifts to the ego-agent, assessing its ability to select well-timed navigation, manipulation, communication, or justified inaction in response to the inferred intent to advance the exchange. Finally, *Interaction* tracks the evolving episode to evaluate the overall outcome against task completion, mutual understanding, and social norms. By explicitly linking these stages into an interdependent causal chain, this formulation ensures that models are evaluated not merely on their final actions, but on the complete progression of the exchange from start to finish.

To operationalize this evaluation, AIRIS aligns every social episode along four precise temporal anchors: T_0 (social cue onset), T_1 (when intent becomes inferable), T_2 (ego-agent response onset), and T_3 (interaction resolution). By querying models at specific moments τ using only the causal observation prefix up to that point, we prevent future evidence from leaking into online decisions. To capture the complexity and diversity of real-world social environments, AIRIS encompasses 8,000 questions across 357 egocentric video clips, spanning 11 diverse interaction contexts ranging from retail and market commerce to everyday social encounters and collaborative professional services. These questions probe 36 fine-grained capability dimensions systematically structured across the four A-I-R-I interaction stages. We benchmark 45 representative proprietary and open-source general-purpose, embodied, streaming video, and omni-modal Vision-Language Models (VLMs),

Table 1: Comparison with representative online-video, social-reasoning, and embodied-decision benchmarks. QA counts cover evaluation sets only. The grouped columns instantiate A-I-R-I as interactant/cue grounding, latent-goal/decidability, ego action/non-intervention, and completion/norm compliance. ✓/✗indicate explicit/no explicit coverage; ✓/- indicate partial/unreported human QA review. AIRIS alone spans all columns under both causal-prefix and full-video settings.

Benchmark	Benchmark profile				Input setting		Attention		Intention		Reaction		Interaction	
	Ego view	QA count	QA dimensions	Human review	Causal streaming	Full video	Interactant grounding	Cue-target grounding	Latent goal	Intent Decidability	Ego-action decision	Non-intervention	Completion / understanding	Social-norm compliance
StreamingBench (Lin et al., 2026)	✗	4,500	18	✓	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗
OVO-Bench (Niu et al., 2025)	✓	2,814	12	✓	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗
RTV-Bench (Xun et al., 2026)	✓	4,631	8	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
EgoSAT (Lei et al., 2026)	✓	4,800	6	-	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
ESTP-Bench (Yu et al., 2026)	✓	2,264	14	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗	✗
ProactiveVideoQA (Wang et al., 2025b)	✓	1,427	4	-	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗
IPIBench (Li et al., 2026a)	✓	3,738	9	-	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗
Human-MME (Liu et al., 2026)	✗	19,945	21	-	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
HumanPCR (Li et al., 2026c)	✗	6,618	34	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗
SIV-Bench (Kong et al., 2026)	✓	5,455	10	✓	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗
R^3-VQA (Niu et al., 2026)	✗	5,156	4	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗
GRASP-Bench (Kim et al., 2026)	✗	1,000	16	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
EgoNormia (Rezaei et al., 2025)	✓	5,559	7	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓
RobotEQ (Fang et al., 2026)	✓	6,639	8	✓	✗	✗	✓	✗	✗	✗	✓	✓	✗	✓
AIRIS (Ours)	✓	**8,000**	**36**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

including Gemini, GPT, Doubao, the Qwen and InternVL series, as well as specialized models such as HY-Embodied, RoboBrain, RynnBrain, and JoyAI.

Our evaluation reveals that current social embodied competence is both fragmented across model families and miscalibrated over time. Leading general-purpose VLMs perform strongly in cue grounding and intent recognition, while embodied specialists are more competitive in aggregate Reaction performance; nevertheless, no model closes the full reaction decision chain. The best Intent Anticipation Score is only 0.237, category-level False Commitment Rates remain between 0.26 and 0.51, and Reaction Joint-EM never exceeds 0.147. Thus, the central limitation is not merely recognizing social content, but knowing when the evidence supports an intent and composing that inference into a complete, timely response. Current VLMs remain capable observers and post-hoc judges, but not yet reliably calibrated social agents. Our contributions are threefold:

- **Formalizing the Social Interaction Loop.** We break down multi-party social interaction into an interdependent four-stage decision process (A-I-R-I), introducing a time-anchored framework that bridges passive multimodal perception with proactive embodied action and social norm compliance.
- **First-Person Causal Streaming Benchmark.** We present AIRIS, a benchmark of 8,000 QA pairs across 11 diverse social contexts and 36 capability dimensions, evaluated under a strict causal streaming protocol that mirrors ego-centric sensory constraints and eliminates future-information leakage.
- **Exposing the Perception-Execution Gap.** Across 45 models spanning general-purpose, embodied, streaming-video, and omni-modal families, AIRIS reveals that relatively strong social-cue perception and post-hoc assessment do not translate into evidence-calibrated anticipation or complete embodied reactions. Temporal calibration and end-to-end decision composition emerge as two major unresolved bottlenecks to reliable social embodied intelligence.

2 The AIRIS Benchmark

2.1 Benchmark Formulation

Problem Formulation. We evaluate social embodied intelligence as an online decision process over a first-person stream, spanning Attention, Intention, Reaction, and Interaction. A social episode is a time-bounded interpersonal exchange observed as $S = \{o_t\}_{t=0}^{T_{\max}}$, where each $o_t = (v_t, a_t)$ contains only first-person RGB frames v_t and acoustic signals a_t (raw audio for omni-modal models, or time-aligned ASR transcripts for standard VLMs). At query time τ, a model may access only the causal prefix $O_{\leq\tau} = \{o_t \mid t \leq \tau\}$.

Temporal Anchors. To track the progression of a social exchange, we map the causal chain of each episode to four precise temporal anchors:

- T_0 (*Cue Onset*): the first manifestation of a social cue (e.g., a gaze shift or gesture onset), marking the start of the *Attention* window.
- T_1 (*Decidability Onset*): the earliest moment at which human experts can unambiguously determine the partner's intent.
- T_2 (*Reaction Onset*): the onset of the ego-agent's substantive physical or verbal response.
- T_3 (*Interaction Resolution*): the conclusion of the immediate exchange.

The A-I-R-I Task Suite. To operationalize this causal timeline into measurable capabilities, AIRIS structures the evaluation into four interdependent task layers, encompassing 36 fine-grained capability dimensions, as shown in Figure 1. All tasks are formulated as closed-set multiple-choice to ensure objective and reproducible evaluation. Specifically, the suite unfolds along the interaction timeline as follows:

- *Attention (Understanding observable social cues):* This layer models attention as a sequential, two-step grounding process. First, the model must allocate the ego-agent's focus to localize the potential interaction partner within a complex environment. Second, it must analyze the partner's own attention and observable state, including gaze and pointing targets, as well as gestures, poses, and emotions. By evaluating interactant identification and target anchoring strictly from the $O_{\leq\tau}$ stream, this layer ensures the model correctly isolates the relevant person and their focus before inferring deeper intent.
- *Intention (Inferring the partner's latent goals):* Given the grounded evidence, this layer asks what the partner is attempting to achieve. Intention is therefore a timed inference of a latent goal. First, the model must decide whether the goal is already supportable from $O_{\leq\tau}$. Second, only when $\tau \geq T_1$ must it recover that goal and select the visual, vocal, or verbal evidence that justifies the inference. For prefixes with $\tau < T_1$, "Insufficient evidence" is the explicitly correct answer. This strictly separates calibrated waiting from premature guessing.
- *Reaction (Deciding ego-agent decisions):* This layer shifts from interpreting the partner to deciding what the ego-agent should do now, given the inferred intent and the evidence available in $O_{\leq\tau}$. First, the model must make an engagement decision: whether to act, communicate, keep observing, or refrain. Second, only if engagement is warranted must it specify the action domain, and the concrete sub-action. Waiting and non-intervention remain valid targets when $\tau < T_1$ or when an action would interrupt an ongoing activity.
- *Interaction (Evaluate completeness and norms):* Evaluated post-hoc at $\tau \geq T_3$, this layer assesses the episode's overall state and outcome. First, the model must track the episode state and detect whether a social violation occurs, and if so of what type. Second, at $\tau \geq T_3$ it must judge the completed exchange along task completeness, mutual understanding, spatial etiquette, physical demeanor, communication etiquette, and overall interaction category.

2.2 Dataset Construction

To construct AIRIS under rigorous quality control, we employ a six-layered human-model collaborative pipeline spanning video curation, structured annotation, and benchmark assembly (Figure 2). Quality assurance is incorporated throughout the pipeline: candidate videos undergo VLM screening and human verification, temporal proposals are manually refined, structured labels are reviewed and adjudicated, and every generated question receives expert inspection. This layered design prevents unverified model outputs from propagating directly into the benchmark and preserves a traceable connection from each released question to its source clip, temporal context, and reviewed annotation.

Keyword-guided video retrieval. We first define a keyword inventory covering the target interaction contexts and use it to retrieve candidate egocentric videos from publicly accessible web platforms. At this stage, retrieval serves only to construct a broad candidate pool; the presence of a first-person viewpoint or an interpersonal exchange does not by itself qualify a video for inclusion.

Model-assisted eligibility screening. A proprietary VLM screens candidate videos against five criteria: egocentric viewpoint, temporal continuity, audiovisual quality, content suitability, and interaction validity. These criteria distinguish usable first-person social exchanges from footage with disrupted event sequences, insufficient perceptual evidence, or no meaningful interpersonal

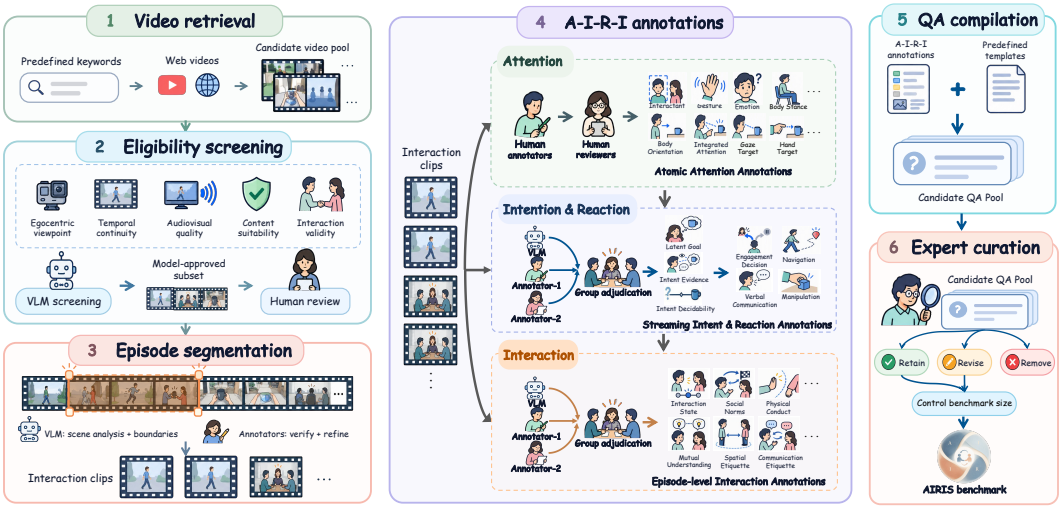


Figure 2: **AIRIS construction pipeline.** Web videos are screened, human-verified, and segmented into episodes. Clips then receive fully manual Attention labels and model-assisted, multi-annotator Intention, Reaction, and Interaction labels with group adjudication. Verified A-I-R-I annotations are templated into a candidate question pool and expert-curated into the final benchmark.

interaction. Human reviewers reassess the model-approved subset using the same criteria; only videos passing both stages proceed to episode segmentation.

Human-refined episode segmentation. For eligible videos, a proprietary VLM analyzes the scene and proposes temporal boundaries for interaction episodes. Human reviewers verify the proposed episodes and refine their boundaries under the same annotation guidelines, preserving the context needed to interpret the exchange. The approved intervals are extracted as egocentric interaction clips.

Layer-specific A-I-R-I annotation. We tailor the annotation protocol for each clip to the distinct evidential demands of the A-I-R-I layers. The Attention layer focuses on perceptually subtle yet directly observable social cues. To minimize missed cues and spatial-grounding errors from automated pre-annotation, we annotate this layer entirely by hand. One annotator identifies the interactant, observable social cues, their targets, and the corresponding spatial grounding, while a second human reviewer verifies the resulting labels. For streaming Intention and Reaction annotations, whose temporal boundaries and semantic interpretations can be ambiguous, a strong proprietary VLM and two human annotators separately propose candidate temporal anchors, partner intentions, and Ego responses based on the admissible causal prefixes. Interaction annotations follow the same model-assisted, multi-annotator protocol, covering both the evolving interaction state and resolved outcomes such as completeness, mutual understanding, and norm compliance. An expert panel reviews the candidate annotations, resolves disagreements against the video evidence, and determines the final labels. Model outputs therefore serve only as annotation proposals and are never accepted as ground truth without human verification.

Template-based question compilation. The verified annotations are converted into a candidate question pool using predefined templates associated with the 36 capability dimensions. Each generated item retains a traceable link to its source clip, supporting annotation, admissible temporal context, and answer key. Streaming items expose only the causal observation prefix $O_{\leq \tau}$. In particular, Intention questions queried before T_1 assign "Insufficient evidence" as the correct response rather than requiring a premature intent prediction. Interaction-state and online-compliance questions likewise use the prefix available at their query time, whereas resolved outcome questions use the episode through T_3. This construction prevents future evidence from entering online decisions while preserving complete episodes for retrospective outcome assessment.

Expert question review and benchmark finalization. Finally, expert reviewers inspect each candidate item together with its video evidence, temporal scope, question, answer options, and proposed answer key. Reviewers retain valid items, revise correctable items, and remove questions with insufficient evidence, ambiguous or redundant options, non-unique answers, inconsistent temporal context, or future-information leakage. This question-level review is conducted separately from

annotation adjudication because a verified structured label does not necessarily yield an unambiguous evaluation item. The resulting review decisions control benchmark size while ensuring that every released question is supported by its designated visual, acoustic, and temporal evidence.

2.3 Evaluation Protocol and Metrics

We evaluate AIRIS using a balanced top-level metric alongside layer-specific probes designed to penalize socially plausible but non-executable or premature predictions. Base items are scored by exact match (single-choice, yes/no, bounding-box) or set F1 (multi-choice), with invalid responses scoring 0. The headline metric is the unweighted mean of the four question-micro layer scores: $\text{Overall} = \frac{1}{4}(S_{\text{Att}} + S_{\text{Int}} + S_{\text{Rea}} + S_{\text{Inter}})$.

Within each layer, we isolate the hardest reasoning bottlenecks. For **Attention**, we evaluate *Semantic target* (which entity) and *Grounding* (where it is) separately, as naming without localizing fails downstream execution. For **Intention**, we propose the *Intent Anticipation Score* (IAS) to evaluate streamed intent recovery. Since "insufficient evidence" is the correct label before T_1, accuracy alone cannot penalize lucky guessing. For models reaching stable correctness at timestamp $p \geq T_1$, we define an event-level anticipation margin $m = \text{clip}((T_2 - p)/(T_2 - t_f), 0, 1)$, where t_f is the first query at or after T_1 (with $m = 0$ if stability is never reached). The event-level score is:

$$\text{IAS}_{\text{event}} = m \cdot (1 - \text{FCR}), \tag{1}$$

where FCR is the fraction of premature pre-T_1 commitments. The final IAS is the macro average across events. For **Reaction**, we report *Engage* as the macro-F1 between active and non-intervention classes to penalize collapsed policies, alongside a strict *Joint-EM* diagnostic requiring simultaneous matches across domain, subaction, and speech. Finally, post-hoc **Interaction** assesses *Compliance* across spatial, physical, and communication etiquette, and *Violation*, which takes "violation present" as the positive class to prevent exploitation of the heavily skewed negative baseline.

Extended diagnostics for all evaluated models are reported in Appendix A.4.2, including nonverbal-state, joint identity-localization, and multimodal-attribution scores for Attention; key-moment, online, evidence-grounding, FCR, and IAS scores for Intention; engagement, action-domain, subaction-plus-speech, Joint-EM, and action-domain breakdowns for Reaction; and state, task-outcome, compliance, and violation scores for Interaction.

2.4 Benchmark Statistics



Figure 3: **AIRIS composition.** The inner ring shows the balanced A-I-R-I hierarchy; the outer ring illustrates the distribution across 11 interaction contexts.

The AIRIS benchmark comprises 8,000 questions grounded in 357 egocentric clips and 276 source videos, totaling 10.91 hours of footage. Clip durations range from 3.0 to 767.3 seconds with a mean of 110.0 seconds, capturing both brief encounters and extended interactions. The dataset averages 22.4 questions per clip and 11.46 sampled frames per question, comprehensively covering 36 capabilities, 11 contexts, and 230 fine-grained subscenes.

These clips span diverse real-world settings as shown in Figure 3. The predominant categories are Market Commerce with 1,324 questions, Store Retail with 1,055, Work & Professional Services with 1,052, and Social Encounters with 976, alongside environments like dining, recreation, and wayfinding. Through rigorous expert curation, AIRIS maintains a perfectly balanced top-level hierarchy that allocates exactly 2,000 questions to each A-I-R-I layer.

To probe different reasoning facets, the benchmark employs four response formats: single-choice at 63.1%, multi-choice at 25.6%, bounding-box selection at 6.4%, and yes/no at 5.0%. Furthermore, 87.7% of questions feature streaming triggers and 41.1% incorporate speech context. To account for heterogeneous question density across clips, we report both per-capability and clip-macro metrics.

Table 2: **Main results on AIRIS benchmark.** Models are grouped by type and sorted by Overall within each group. Scores are in $[0, 1]$ (higher is better); **bold**/underline mark the best/second-best.

Model	Attention			Intention			Reaction			Interaction			Overall
	All	Target	Ground	**All**	Online	IAS	**All**	Engage	Joint	**All**	Comp.	Viol.	
Closed-source models													
Gemini-3.7-Flash (2026b)	**0.936**	**0.951**	**0.948**	**0.569**	**0.569**	**0.237**	0.638	0.546	0.083	**0.760**	0.903	0.525	**0.726**
Gemini-3.1-Pro (2026a)	0.851	0.897	0.820	0.496	0.523	0.126	0.620	0.530	0.111	**0.760**	0.902	0.495	0.682
GPT-5.6-Sol (2026)	0.906	0.925	0.884	0.501	0.457	0.154	0.489	0.261	0.128	0.729	0.906	0.494	0.656
Doubao-seed-2.0-lite (2026)	0.759	0.859	0.742	0.418	0.363	0.066	0.619	0.564	0.092	0.669	0.726	0.572	0.616
Open-source general-purpose models													
Qwen3.6-35B-A3B (2026c)	0.723	0.891	0.646	0.485	0.439	0.122	0.655	0.605	0.064	0.738	0.881	0.577	0.650
GLM-5.3-Flash (2026)	0.801	0.882	0.802	0.501	0.455	0.128	0.590	0.531	0.067	0.638	0.706	0.470	0.633
Gemma-4-31B-IT (2026a)	0.733	0.899	0.619	0.502	0.482	0.167	0.535	0.394	0.123	0.728	0.907	0.495	0.624
DeepSeek-V4.1-Flash (2026a)	0.707	0.829	0.627	0.509	0.451	0.121	0.548	0.456	0.108	0.657	0.737	0.508	0.605
Qwen3.6-27B (2026b)	0.649	0.836	0.528	0.480	0.483	0.119	0.545	0.383	0.128	0.733	0.905	0.501	0.602
Qwen3-VL-32B-Instruct (2025a)	0.682	0.886	0.598	0.472	0.481	0.133	0.479	0.217	0.144	0.723	**0.923**	**0.589**	0.589
Qwen3.8-27B (2026d)	0.685	0.835	0.616	0.395	0.374	0.052	0.475	0.251	0.137	0.632	0.717	0.341	0.547
InternVL3-8B (2025)	0.583	0.778	0.513	0.417	0.375	0.082	0.518	0.271	0.123	0.603	0.724	0.318	0.530
InternVL3.5-30B-A3B (2025a)	0.598	0.727	0.564	0.426	0.458	0.095	0.442	0.134	0.144	0.639	0.754	0.511	0.526
InternVL3.5-8B (2025a)	0.574	0.768	0.490	0.406	0.374	0.094	0.458	0.309	0.119	0.562	0.689	0.351	0.500
LLaVA-OneVision-7B (2024a)	0.500	0.757	0.346	0.387	0.380	0.078	0.379	0.132	0.144	0.602	0.657	0.453	0.467
Qwen3.5-4B (2026a)	0.534	0.715	0.433	0.389	0.442	0.123	0.481	0.453	0.129	0.425	0.523	0.322	0.457
Llama-3.2-11B-Vision-Instruct (2024)	0.391	0.620	0.261	0.384	0.422	0.011	0.463	0.371	0.128	0.455	0.498	0.381	0.423
Qwen3.5-0.8B (2026a)	0.381	0.537	0.278	0.299	0.276	0.036	0.506	0.517	0.044	0.398	0.353	0.329	0.396
Open-source embodied models													
HY-Embodied-VLM-1.0 (2026)	0.596	0.834	0.492	0.396	0.419	0.105	**0.688**	0.485	0.028	0.716	0.873	0.481	0.599
RynnBrain1.1-9B (2026b)	0.593	0.819	0.470	0.417	0.468	0.123	0.557	0.605	0.082	0.736	0.918	0.583	0.576
RynnBrain-30B-A3B (2026)	0.656	0.777	0.596	0.435	0.471	0.091	0.410	0.245	**0.147**	0.674	0.723	0.508	0.544
VeBrain-7B (2025)	0.546	0.801	0.440	0.449	0.494	0.111	0.470	0.154	0.141	0.701	0.828	0.502	0.541
Pelican1.0-VL-7B (2025b)	0.540	0.724	0.425	0.462	0.503	0.099	0.440	0.128	0.144	0.700	0.875	0.470	0.536
MiMo-Embodied-7B (2025)	0.533	0.761	0.391	0.440	0.407	0.114	0.643	0.558	0.056	0.487	0.669	0.466	0.526
RoboBrain2.0-7B (2025)	0.511	0.809	0.349	0.439	0.447	0.054	0.471	0.184	0.146	0.597	0.696	0.379	0.505
Embodied-R1-3B-V1 (2026)	0.475	0.703	0.375	0.408	0.376	0.073	0.468	0.126	0.144	0.569	0.635	0.412	0.480
HY-Embodied-0.5 (2026b)	0.482	0.694	0.464	0.368	0.388	0.022	0.563	**0.606**	0.065	0.429	0.353	0.318	0.461
Open-source streaming video models													
JoyAI-VL-Interaction-Preview (2026)	0.639	0.761	0.628	0.436	0.408	0.113	0.610	0.508	0.029	0.586	0.635	0.417	0.568
AURA-8B (2026)	0.624	0.834	0.574	0.472	0.490	0.113	0.434	0.146	0.144	0.601	0.766	0.510	0.533
StreamingVLM (2026b)	0.456	0.753	0.292	0.400	0.449	0.029	0.536	0.302	0.141	0.655	0.785	0.413	0.512
Flash-VStream-Qwen-7B (2025a)	0.470	0.675	0.367	0.396	0.398	0.065	0.376	0.131	0.142	0.631	0.682	0.436	0.468
VITA-1.5 (2026)	0.460	0.638	0.412	0.388	0.435	0.091	0.420	0.344	0.103	0.477	0.523	0.468	0.436
Open-source omni-modal models													
Qwen2.5-Omni-7B (2025a)	0.507	0.745	0.375	0.453	0.485	0.070	0.449	0.126	0.144	0.705	0.855	0.510	0.529
Qwen3-Omni-30B-A3B-Instruct (2025b)	0.646	0.814	0.590	0.419	0.498	0.153	0.390	0.138	0.141	0.647	0.746	0.415	0.526
OmniVinci (2026)	0.546	0.733	0.456	0.401	0.410	0.116	0.340	0.126	0.144	0.651	0.740	0.454	0.484
Qwen2.5-Omni-3B (2025a)	0.477	0.762	0.338	0.433	0.466	0.065	0.467	0.126	0.144	0.520	0.534	0.200	0.474
Phi-4-Multimodal-Instruct (2025)	0.435	0.541	0.368	0.398	0.426	0.000	0.435	0.133	0.137	0.356	0.172	0.301	0.406

Abbreviations: All = stage Overall; Target = semantic target; Ground = grounding; IAS = Intent Anticipation Score; Engage = engagement Macro-F1; Joint = full Reaction-chain correctness; Comp. = compliance; Viol. = violation.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Evaluated Models. We evaluate 45 models in total to provide a broad landscape of social embodied capabilities. Due to space constraints, the main text focuses on 37 representative state-of-the-art models spanning five categories: closed-source models, open-source general-purpose VLMs, open-source embodied models, open-source streaming video models, and open-source omni-modal models. Results for these 37 models are reported in Table 2, while the complete model list and results for all 45 models are provided in Supplementary Section A.4.1.

Evaluation Harness. Models are evaluated zero-shot on the 8,000-question AIRIS benchmark. For each query at timestamp τ, models receive only the causal video prefix $O_{\leq\tau}$. Acoustic context is supplied as raw audio for omni-modal models and as time-aligned ASR transcripts for standard VLMs. Questions are formatted as multiple-choice, yes/no, or bounding-box selection tasks.

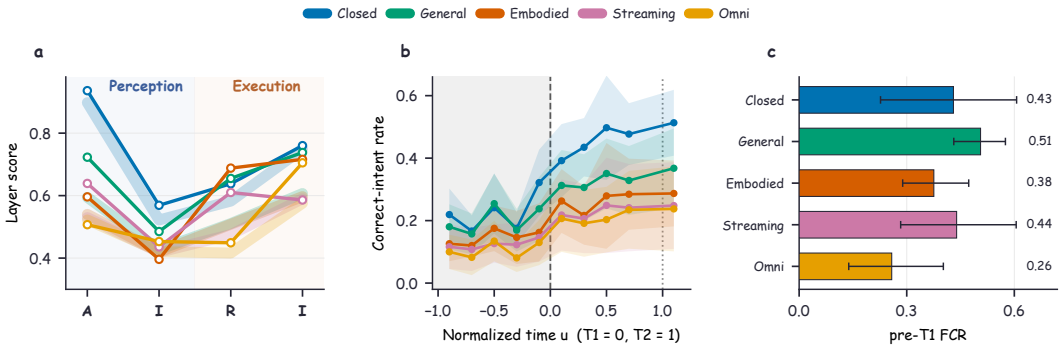


Figure 4: **Fragmented capability across the causal chain.** (a) A-I-R-I performance: category averages (thick, translucent) and best models (thin, markers). (b) Correct-intent rate vs. normalized interaction time.(c) Mean False Commitment Rate (FCR) before T_1.

3.2 Experimental Results and Insights

Table 2 presents the evaluation results of 37 representative models on the AIRIS benchmark. Closed-source models, led by Gemini-3.7-Flash, establish a dominant baseline in perceptual grounding and intent recognition. Gemini-3.7-Flash achieves the highest overall score of 0.726 and significantly outperforms open-source counterparts in Attention at 0.936 and Intention at 0.569. Conversely, open-source models, particularly those fine-tuned for embodied tasks, demonstrate highly competitive or superior decision-making capabilities. HY-Embodied-VLM-1.0 achieves the highest overall Reaction score of 0.688, while Qwen3-VL-32B-Instruct leads in Interaction compliance at 0.923 and violation detection at 0.589.

The Perception-Execution Gap. We observe a pronounced capability divide across the A-I-R-I causal chain (Figure 4a). While general-purpose VLMs excel at passive observation and semantic reasoning (Attention and Intention), their advantage vanishes in the downstream Reaction and Interaction layers. Instead, embodied-specific models take the lead in generating socially appropriate, executable responses. This divergence reveals that "understanding" does not naturally translate to "doing"; generating compliant embodied actions in dynamic social scenarios requires specialized action-oriented alignment beyond standard semantic pre-training.

Miscalibrated Causal Anticipation. Streaming intent anticipation remains a critical bottleneck across all models. Even the leading Gemini-3.7-Flash reaches an Intent Anticipation Score (IAS) of only 0.237, and most models score near 0.1. The anticipation curves in Figure 4b make the temporal structure explicit. The correct intent rate stays low before the decidability onset ($u < 0$), climbs only after T_1, and remains modest even at the reaction onset T_2 (at most $\sim$0.5 for closed source models). Crucially, this early deficit is *not* calibrated waiting. As Figure 4c shows, the False Commitment Rate (FCR) before T_1 remains substantial across categories (0.26 to 0.51). This fraction measures queries that commit to a specific intent when "insufficient evidence" is the correct answer. Individual models range from near abstention (Phi-4-Multimodal, 0.08) to aggressive guessing (GLM-5.3-Flash, 0.64; Qwen3.6-35B-A3B, 0.63; GPT-5.6-Sol, 0.57). Even Gemini-3.7-Flash commits in 42% of early queries. Current models thus fail anticipation in two opposite ways: overconfident premature commitment or blind passivity, rather than exercising the evidence gated inference the task rewards.

No model closes the full reaction decision chain. Our reaction evaluation reports two complementary views: an aggregate score (R-All) crediting each sub-decision (engagement, action domain, and the concrete subaction or utterance) and a stricter Joint-EM requiring all of them to be correct simultaneously for the same event. Each sub-decision is individually non-trivial, yet requiring them to hold jointly makes performance collapse multiplicatively. No model exceeds a Joint-EM of 0.147, which is far below what the component scores alone would predict. Deciding *whether* to engage, or planning a plausible subaction, is therefore not enough on its own; only their correct, end-to-end composition yields a trustworthy social response. The strict lens accordingly re-orders the leaderboard. HY-Embodied-VLM-1.0, Qwen3.6-35B-A3B, and MiMo-Embodied-7B lead on aggregate Reaction (0.688, 0.655, 0.643) but fall to 0.028, 0.064, and 0.056 under the joint constraint. This integrative competence, rather than isolated perception, intent, or action, is the true prerequisite for trustworthy social embodied intelligence, and it remains largely unmet.

Unrealized Architectural Dividends in Social Reasoning. Models specifically designed for native streaming video (e.g., JoyAI-VL, StreamingVLM) or omni-modal audio-visual processing (e.g., Qwen3-Omni) do not exhibit significant advantages over traditional frame-sampled VLMs (e.g., Qwen3.6-35B) in overall social embodied intelligence. The architectural benefits of continuous stream processing appear confined to low-level perception or speech dialogue. For tasks requiring deep Theory of Mind and adherence to complex social norms, these architectural shifts have not yet compensated for the lack of high-level social reasoning capabilities.

Explicit reasoning does not reliably help streaming interaction. Allocating more test-time computation to explicit chain-of-thought is a natural way to attack our causal streaming tasks, yet under identical streaming input, scored with a thinking-aware re-parse that corrects a systematic answer-parsing artifact, it brings no overall benefit: Qwen3.6-27B ties its no-think baseline (0.1 points) and Qwen3.6-35B drops (6.4 points), while Qwen3-Omni gains only marginally (2.9 points). The effect is structured along the A-I-R-I chain (Figure 5a): reasoning tends to help multi-step spatial or decision sub-tasks, such as spatial grounding (A-Ground, +21.8 points for Qwen3.6-27B), but consistently hurts the columns that reward restraint under partial evidence, namely online intent (I-Online) and interaction compliance, across all models. The penalty is legible at the trace level (Figure 5b): accuracy falls with trace length, and although harder items independently elicit longer reasoning (a selection effect), thinking still widens the gap on the hardest Qwen3.6 items, whereas the restrained Qwen3-Omni is nearly unchanged. Traces carrying cognitive-revision markers ("however", "wait", "reconsider"), present in about 77% of Qwen3.6 traces, lose roughly 27 to 30 points of accuracy: when perception is insufficient, extended deliberation fills the visual gap with language and overturns a correct first instinct. Chain-of-thought optimized for static text problems is thus not yet adapted to the real-time, causally restrained demands of streaming social interaction.

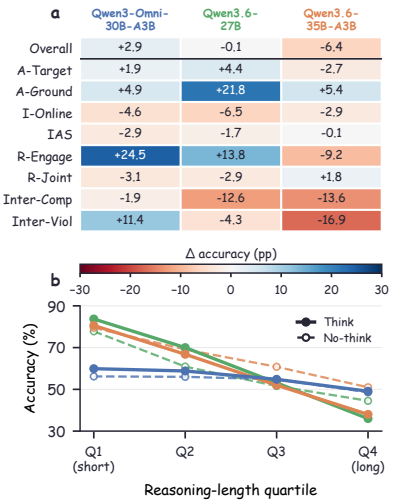


Figure 5: **Think versus no-think.** (a) Accuracy change across AIRIS columns. (b) Accuracy by reasoning-length quartile (solid, think; dashed, no-think). Results use thinking-aware re-parsing.

4 Conclusion

We introduced AIRIS, a first-person causal streaming benchmark evaluating social embodied intelligence as a time-anchored progression across Attention, Intention, Reaction, and Interaction. Across 45 models, strong cue perception and intent recognition do not reliably translate to calibrated anticipation or complete embodied reactions. AIRIS thus shifts the evaluation focus from static social recognition to the dynamic timing of evidence sufficiency, action necessity, and response composition.

These findings highlight key priorities for future models. First, models must preserve calibrated uncertainty before an intent becomes decidable, committing only when supported by observable evidence. Second, perception, inference, and action specification must be optimized as a coupled decision process to ensure isolated strengths produce coherent reactions. Finally, test-time computation alone is insufficient; streaming reasoning must remain grounded in current sensory evidence and adapt its depth to uncertainty, avoiding the replacement of missing observations with unsupported verbal inference.

By making timing, restraint, and decision composition measurable under causal observation, AIRIS establishes a rigorous foundation for evaluating social embodied intelligence. Future work will extend this framework to interactive robot-in-the-loop settings, broader social contexts, and open-ended generation. Ultimately, this progress fosters human-centered embodied agents: ones that not only understand what people are doing, but also know when to wait, whether to act, and how to respond in ways that align with human social norms and facilitate seamless human-robot coordination.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the research methodology and experimental design, and to assist with the interpretation and presentation of experimental results. We also used generative AI tools for literature search and summarization, manuscript drafting and editing, structural revisions to improve clarity, and formatting and submission checks. During dataset construction, generative AI tools were used to produce candidate descriptions and pre-annotations. AI-generated annotations were never accepted directly as ground truth; human experts reviewed, corrected, and adjudicated every candidate incorporated into AIRIS.

All AI-assisted suggestions were reviewed by the authors before being incorporated into this work. Experimental results, quantitative claims, implementation details, and cited literature were checked against the corresponding experiment records, code, or primary sources, as applicable. Generative AI tools did not autonomously make research decisions or replace the authors' verification of the reported findings.

The authors take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

AIRIS is constructed from egocentric videos retrieved from publicly accessible web platforms. Such accessibility is not treated as permission for unrestricted redistribution: the release is limited to benchmark material that passes source-rights and privacy review, while material that cannot be redistributed is represented only through derived annotations and provenance metadata where permitted. The curation protocol excludes violent, gory, or adult content and close-ups of minors' faces. Annotators use non-identifying participant descriptions and are instructed not to infer identities, protected attributes, private relationships, or unobservable mental states. Model outputs serve only as proposals, and final labels are determined through human review and adjudication.

The benchmark is intended to evaluate socially appropriate perception and decision making, not to identify individuals, diagnose mental states, or support surveillance or other high-stakes decisions. Because labels involving intent and social norms can reflect contextual and cultural judgments, AIRIS reports its annotation protocol and should not be interpreted as defining universal social behavior. Benchmark performance alone does not establish that a model is safe for deployment around people. The release documentation will state the applicable licenses, access conditions, known limitations, and a procedure for reporting privacy or removal concerns.

REPRODUCIBILITY STATEMENT

The benchmark formulation, causal input contract, construction pipeline, evaluation protocol, and metric definitions are provided in Sections 2.2 and 2.3. The supplementary material documents the complete capability ontology, expert annotation interface, human-review procedure, and core pre-annotation prompts (Appendix A.3), together with the evaluated model pool and complete quantitative results (Appendix A.4). Upon acceptance, we will release the evaluation harness, scoring code, model configurations, and raw prediction outputs used to produce the reported results, together with benchmark annotations and causal-prefix metadata that can be redistributed under the applicable source terms and privacy constraints. The release documentation will identify any proprietary models or media that cannot be redistributed and specify the corresponding access conditions.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Tongfei Bian, Mathieu Chollet, and Tanaya Guha. Robust understanding of human-robot social interactions through multimodal distillation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5726–5734, 2025.

ByteDance Seed Team. Seed2.0: Model overview. https://seed.bytedance.com/zh/seed2, 2026. Accessed September 25, 2026.

Xu Cao, Pranav Virupaksha, Wenqi Jia, Bolin Lai, Fiona Ryan, Sangmin Lee, and James M Rehg. Socialgesture: Delving into multi-person gesture understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19509–19519. IEEE, 2025.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418. IEEE, 2024.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, et al. Rynnbrain: Open embodied foundation models, 2026.

Kerstin Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480):679, 2007.

Davide Dell'Anna and Anahita Jamshidnejad. Sonar: An adaptive control architecture for social norm aware robots. *International Journal of Social Robotics*, 16(9):1969–2000, 2024.

Kuofei Fang, Xinyi Che, Haomin Ouyang, Shufan Zhang, Xuehao Wang, Qi Liu, Liyi Liu, Chenqi Zhang, Wenxi Cai, Wenyu Dai, et al. Roboteq: Transitioning from passive intelligence to active intelligence in embodied ai. *arXiv preprint arXiv:2605.06234*, 2026.

Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Advances in Neural Information Processing Systems*, 38:75300–75320, 2026.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. Vispeak: Visual instruction feedback in streaming videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21778–21788. IEEE, 2025b.

Google DeepMind. Gemini 3.1 Pro: Model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026a. Published February 19, 2026.

Google DeepMind. Gemini 3.7 Flash: Model card. https://deepmind.google/models/model-cards/gemini-3-7-flash/, August 2026b. Published August 13, 2026.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, et al. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22087–22097. IEEE, 2024.

Junho Kim, Xu Cao, Houze Yang, Bikram Boote, Ana Jojic, Fiona Ryan, Bolin Lai, Sangmin Lee, and James M Rehg. Grasp: Learning to ground social reasoning in multi-person non-verbal interactions. *arXiv preprint arXiv:2605.15764*, 2026.

Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. Siv-bench: A video benchmark for social interaction understanding and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37379–37403, 2026.

Daeun Lee, Subhojyoti Mukherjee, Branislav Kveton, Ryan A Rossi, Viet Dac Lai, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, and Mohit Bansal. Streamgaze: Gaze-guided temporal reasoning and proactive understanding in streaming videos. *arXiv preprint arXiv:2512.01707*, 2025a.

Dong Won Lee, Yubin Kim, Denison Guvenoz, Sooyeon Jeong, Parker Malachowsky, Louis-Philippe Morency, Cynthia Breazeal, and Hae Won Park. Social human robot embodied conversation (shrec) dataset: Benchmarking foundational models' social reasoning. *arXiv preprint arXiv:2504.13898*, 2025b.

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. *arXiv preprint arXiv:2606.10276*, 2026.

Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, and James M Rehg. Modeling multimodal social interactions: new challenges and baselines with densely aligned representations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14585–14595. IEEE, 2024a.

Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, et al. Towards social ai: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*, 2024b.

Yijia Lei, Jinzhao Li, Yichi Zhang, Jiacheng Hua, Yin Li, and Miao Liu. Egosat: A comprehensive benchmark of ego centric s treaming inter act ion understanding. In *European Conference on Computer Vision*, pp. 576–593. Springer, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Jinzhao Li, Yinuo Chen, Wenxuan Song, Yijia Lei, Yichi Zhang, Honglei Yan, Panwang Pan, and Miao Liu. Ipibench: Evaluating interactive proactive intelligence of mllms under continuous streams. *arXiv preprint arXiv:2605.27074*, 2026a.

Kehan Li, Bohan Hou, Minghao Zhu, Tianyi Zhang, Zesen Cheng, Zhikai Wang, Sicong Leng, Xin Li, Xiao Lin, Biying Yao, et al. Rynnbrain 1.1: Towards more capable and generalizable embodied foundation model, 2026b.

Keliang Li, Hongze Shen, Hao Shi, Ruibing Hou, Hong Chang, Jie Huang, Chenghao Jia, Wen Wang, Yiling Wu, Dongmei Jiang, et al. Humanpcr: Probing mllm capabilities in diverse human-centric scenes. In *International Conference on Learning Representations*, volume 2026, pp. 68971–69011, 2026c.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024b.

Xinpeng Li, Shijian Deng, Bolin Lai, Weiguo Pian, James Matthew Rehg, and Yapeng Tian. Towards online multimodal social interaction understanding. *Transactions on Machine Learning Research*, 2026d. ISSN 2835-8856. URL https://openreview.net/forum?id=5P7yVfUEuD.

Xinpeng Li, Bolin Lai, Hardy Chen, Shijian Deng, Cihang Xie, Yuyin Zhou, James M Rehg, and Yapeng Tian. Omni-mmsi: Toward identity-attributed social interaction understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8684–8696, 2026e.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

Yuansen Liu, Haiming Tang, Jinlong Peng, Jiangning Zhang, Xiaozhong Ji, Qingdong He, Donghao Luo, Zhenye Gan, Junwei Zhu, Yunhang Shen, Chaoyou Fu, Chengjie Wang, Xiaobin Hu, and Shuicheng YAN. Human-MME: A holistic evaluation benchmark for human-centric multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4qIK0UV2Nt.

Xudong Lu, Yang Bo, Jinpeng Chen, Shuhan Li, Xintong Guo, Huankang Guan, Fang Liu, Dunyuan Xu, Peiwen Sun, Heyang Sun, et al. Aura: Always-on understanding and real-time assistance via video streams. *arXiv preprint arXiv:2604.04184*, 2026.

Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, et al. Visual embodied brain: Let multimodal large language models see, think, and control in spaces, 2025.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24879–24902, 2025.

Meta. Llama 3.2-Vision model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD_VISION.md, September 2024. Released September 25, 2024.

Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. Multimediate'22: Backchannel detection and agreement estimation in group interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 7109–7114, 2022.

Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, et al. Multimediate'23: Engagement estimation and bodily behaviour recognition in social interactions. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9640–9645, 2023.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913. IEEE, 2025.

Lixing Niu, Jiapeng Li, Xingping Yu, Xinyi Dong, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. Read the room: Video social reasoning with mental-physical causal chains. In *The Fourteenth International Conference on Learning Representations*, 2026.

OpenAI. GPT-5.6 system card. Technical report, OpenAI, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf.

Liangyang Ouyang, Yifei Huang, Mingfang Zhang, Caixin Kang, Ryosuke Furuta, and Yoichi Sato. Multi-speaker attention alignment for multimodal social interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24608–24619, 2026.

Amirreza Payandeh, Daeun Song, Mohammad Nazeri, Jing Liang, Praneel Mukherjee, Amir Hossain Raj, Yangzhe Kong, Dinesh Manocha, and Xuesu Xiao. Social-llava: Enhancing social robot navigation through human-language reasoning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 17192–17198. IEEE, 2025.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *Advances in Neural Information Processing Systems*, 38, 2026.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *International Conference on Learning Representations*, volume 2024, pp. 15306–15336, 2024.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055. IEEE, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026a.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model. `https://qwen.ai/blog?id=qwen3.6-27b`, April 2026b.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. `https://qwen.ai/blog?id=qwen3.6-35b-a3b`, April 2026c.

Qwen Team. Qwen3.8-27B model card. `https://huggingface.co/Qwen/Qwen3.8-27B`, August 2026d. Accessed September 22, 2026.

Rebecca Ramnauth, Dražen Brščić, and Brian Scassellati. Should i help?: A skill-based framework for deciding socially appropriate assistance in human-robot interactions. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pp. 2051–2058. IEEE, 2024.

MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, and Diyi Yang. EgoNormia: Benchmarking physical-social norm understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19256–19283, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.985. URL `https://aclanthology.org/2025.findings-acl.985/`.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 1510–1519, 2025.

Gong Sitong, Tianyu Yan, Caixin Kang, Bo Zheng, Xiang Ruan, Huchuan Lu, Kaipeng Zhang, Yoichi Sato, and Yifei Huang. Vinci2: Providing proactive assistance in continuous egocentric videos. *arXiv preprint arXiv:2607.11523*, 2026.

Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. *IEEE Robotics and Automation Letters*, 10(1):508–515, 2024.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026a.

HY Team, Xumin Yu, Zuyan Liu, Ziyi Wang, He Zhang, Yongming Rao, Fangfu Liu, Yani Zhang, Ruowen Zhao, Oran Wang, et al. Hy-embodied-0.5: Embodied foundation models for real-world agents. *arXiv preprint arXiv:2604.07430*, 2026b.

Tencent Robotics X and HY Vision Team. HY-Embodied-0.5-X: An enhanced embodied foundation model for real-world agents. Official model and code release, 2026. URL https://github.com/Tencent-Hunyuan/HY-Embodied-0.5-X.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models. *arXiv preprint arXiv:2507.09313*, 2025b.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18925–18935. IEEE, 2025c.

Ziyi Wang, Xumin Yu, Yongming Rao, Yonggen Ling, Yunheng Li, Oran Wang, Mingqi Gao, Yuchen Zhou, Yves Liang, Zuyan Liu, et al. Hy-embodied-vlm-1.0: Efficient physical-world agents, 2026.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31219–31229, 2026.

Tianyu Xie, Jinfa Huang, Yuexiao Ma, Rongfang Luo, Yan Yang, Qingchuan Ma, Wang Chen, Yuhui Zeng, Yixuan Zou, Zhiqiang Lu, et al. Socialomni: Benchmarking audio-visual social interactivity in omni models. *arXiv preprint arXiv:2603.16859*, 2026.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4. 1-flash: Pushing the limits of kv cache compression, 2026a.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report, 2025b.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. In *International Conference on Learning Representations*, volume 2026, pp. 61463–61475, 2026b.

Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, et al. Rtv-bench: Benchmarking mllm continuous perception, understanding and reasoning through real-time video. *Advances in Neural Information Processing Systems*, 38, 2026.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, et al. Joyai-vl-interaction: Real-time vision-language interaction intelligence, 2026.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

Xueyang Yu, Cheng Shi, Yang Wang, and Sibei Yang. Eyes wide open: Ego proactive video-llm for streaming video. *Advances in Neural Information Processing Systems*, 38:13420–13463, 2026.

Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, Hongyao Tang, and Jianye Hao. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 8984–9009, 2026.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8799–8809. IEEE, 2019.

Z.ai. GLM-5.3-Flash model card. `https://huggingface.co/zai-org/GLM-5.3-Flash`, 2026.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21059–21069. IEEE, 2025a.

Yi Zhang, Che Liu, Xiancong Ren, Hanchu Ni, Shuai Zhang, Zeyuan Ding, Jiayu Hu, Hanzhe Shan, Zhenwei Niu, Zhaoyang Liu, et al. Pelican-vl 1.0: A foundation brain model for embodied intelligence, 2025b.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12055–12079, 2025c.

Yuedi Zhang, Shuanghao Bai, Wanqi Zhou, Haoran Zhang, Qi Zhang, Zhirong Luan, and Badong Chen. Assistance without interruption: A benchmark and llm-based framework for non-intrusive human-robot assistance. *arXiv preprint arXiv:2605.01368*, 2026.

Shiyun Zhao, Xinwei Song, Tianyu Guo, Xiaomeng Gao, Mingyuan Liu, Xu Han, Yuanyuan Zhang, Zhenliang Zhang, Xue Feng, and Bo Dai. Normact: Benchmarking embodied agents' proactive compliance with unspoken social norms. *eprint arXiv: 2606.27826*, pp. arXiv–2606, 2026.

Yuchen Zhou, Guang Tan, Rui Zhong, Yaokun Li, and Chao Gou. Pit: Progressive interaction transformer for pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):14213–14225, 2023.

Yuchen Zhou, Linkai Liu, and Chao Gou. Learning from observer gaze: Zero-shot attention prediction oriented by human-object interaction recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28390–28400. IEEE, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

A Appendix

A.1 Supplementary Overview

This supplementary document provides extended background, technical formulations, complete empirical results, and qualitative galleries supporting the AIRIS benchmark. Section A.2 expands the literature review across streaming video understanding, multimodal social interaction, and norm-aware embodied decision making. Section A.3 presents granular benchmark specifications, including the full 36-capability A-I-R-I ontology (Section A.3.1, Tables 3-6), the fine-grained metric formulations (Section A.3.2), the expert annotation interfaces (Section A.3.3), and data curation prompt templates (Section A.3.4). Section A.4 documents the complete 45-model evaluation pool across five architectural categories (Section A.4.1) alongside comprehensive quantitative results and diagnostic breakdowns across all four stages (Section A.4.2, Tables 7-10). Finally, Section A.5 visualizes representative streaming episodes, causal observation prefixes, and expert-adjudicated ground truth across Attention, Intention, Reaction, and Interaction (Sections A.5.1-A.5.4).

A.2 Related Work

A.2.1 Streaming and Proactive Video Understanding

Most video-language benchmarks (Li et al., 2024b; Fu et al., 2025a; Wu et al., 2024) assume that a model can inspect a complete, temporally bounded clip before answering, whereas deployed assistants receive observations sequentially. On the model side, VideoLLM-online (Chen et al., 2024) introduced temporally aligned dialogue over continuous video streams and an asynchronous inference pipeline for real-time processing. Subsequent systems target complementary bottlenecks: Flash-VStream (Zhang et al., 2025a) compresses long-running visual context into context and augmentation memories for low-latency query answering, whereas Dispider (Qian et al., 2025) separates perception, decision, and reaction into asynchronous modules so that response generation does not suspend stream processing. More recently, Streamo (Xia et al., 2026) learns explicit silence, standby, and response states within a single autoregressive sequence, coupling response timing with content generation across multiple streaming tasks. On the evaluation side, StreamingBench (Lin et al., 2026) distributes timestamped questions throughout a stream, while OVO-Bench (Niu et al., 2025) organizes online understanding into backward tracing, real-time perception, and forward active responding, including waiting until sufficient evidence becomes available. OmniMMI and SocialOmni (Wang et al., 2025c; Xie et al., 2026) broaden this setting to multimodal interaction, proactive alerting, and turn taking. EgoSAT and StreamGaze (Lei et al., 2026; Lee et al., 2025a) evaluate temporal reasoning from strict egocentric prefixes, with the latter incorporating gaze.

Recent work has also shifted from answering predefined queries to deciding when and how an assistant should respond. ViSpeak (Fu et al., 2025b) extracts visual instructions from gestures and body language in streaming video; ESTP-Bench (Yu et al., 2026) evaluates proactive coherence and just-in-time responses; and ProAssist and EDITH (Zhang et al., 2025c; Lee et al., 2026) use egocentric streams for proactive guidance and robot assistance. IPIBench (Li et al., 2026a) further evaluates proactive monitoring, task modification and cancellation, and interleaved reactive-proactive requests. EgoServe (Sitong et al., 2026) frames proactive assistance as a context-dependent intervention problem over multiple memory horizons. Existing streaming and proactive benchmarks predominantly focus on object-centric task execution or user-directed assistance, treating humans primarily as task performers or commanders. Building on their real-time protocols, AIRIS shifts the paradigm to social embodiment. It evaluates the ego-agent as a social participant, diagnosing its ability to navigate interpersonal dynamics, anticipate social intents, and deliver norm-appropriate responses in shared environments.

A.2.2 Multimodal Social Interaction Understanding

A foundational requirement for social embodied intelligence is the ability to perceive and interpret human behavior from heterogeneous social signals. Social-IQ (Zadeh et al., 2019) frames this ability as multimodal question answering over natural social situations, while Human-MME (Liu et al., 2026) and HumanPCR (Li et al., 2026c) provide broader evaluations spanning fine-grained human perception, comprehension, and reasoning. SIV-Bench (Kong et al., 2026) further structures social video understanding into social scene understanding, state reasoning, and dynamics prediction. Moving beyond answer accuracy, Social Genome (Mathur et al., 2025) evaluates whether social inferences are explicitly grounded in visual, verbal, and vocal evidence as well as relevant external knowledge. A complementary line grounds interaction evidence to relevant regions and participants (Zhou et al., 2024; Ouyang et al., 2026). Robot-centric work models individual attributes, group context, intent, and social actions from onboard observations (Jahangard et al., 2024; Bian et al., 2025). SocialGesture (Cao et al., 2025) targets the recognition, temporal localization, and question answering of natural gestures in multi-person interactions. The MultiMediate challenge series (Müller et al., 2022; 2023) evaluates conversational behaviors such as eye contact, backchannels, agreement, engagement, and bodily behavior in group discussions. The MMSI lineage (Lee et al., 2024a; Li et al., 2026d;e) instead emphasizes identity-attributed social cues, including speaking targets and pronoun references, while GRASP (Kim et al., 2026) connects high-level social reasoning to participant-specific gaze and gesture evidence. At a higher reasoning level, MuMA-ToM (Shi et al., 2025) evaluates goals, beliefs, and nested beliefs in embodied multi-agent interactions. In egocentric settings, EgoGazeVQA (Peng et al., 2026) studies gaze as an indicator of user intent, whereas R^3-VQA (Niu et al., 2026) represents belief, intent, desire, and emotion through mental-physical causal chains. Collectively, these benchmarks substantially advance the perception and interpretation

of other people's attention, references, intentions, and social states. AIRIS treats these capabilities as upstream stages of social embodied intelligence and further evaluates whether and how the ego-agent should respond, when inaction is appropriate, and how its decision shapes the ensuing interaction within the same video stream.

A.2.3 Social Norms and Embodied Decision Making

Understanding others is insufficient for social embodiment: an agent must also determine what it should do, when it should act, and which otherwise feasible actions are socially inappropriate. EgoNormia (Rezaei et al., 2025) grounds this problem in egocentric video by evaluating socially appropriate next actions, their justifications, and acceptable alternatives under potentially competing physical-social norms. RobotEQ (Fang et al., 2026) similarly evaluates action permissibility and spatial grounding from robot-view scenes, with an emphasis on deciding how, when, and whether to act without explicit instructions. NormAct (Zhao et al., 2026) extends norm evaluation to embodied planning, testing whether models can infer hidden social constraints and incorporate them into action sequences while still achieving an explicit task goal. The SHREC benchmark (Lee et al., 2025b) evaluates whether foundation models can identify a robot's social errors and competencies, reason about their effects on interaction progression, and propose corrections. Should I Help? (Ramnauth et al., 2024) challenges the assumption that assistance is always desirable and formalizes whether help is appropriate in terms of human and robot skills together with the social benefits and costs to the recipient. NIABench (Zhang et al., 2026) further formulates non-intrusive assistance as a joint decision over when to act and what action to take while preserving the human's ongoing task flow. Embodied robotics has also operationalized particular dimensions of socially appropriate behavior. Habitat 3.0 (Puig et al., 2024) introduces human-in-the-loop social navigation and collaborative rearrangement tasks in which robots must coordinate with people, maintain distance, and yield shared space. VLM-Social-Nav and Social-LLaVA (Song et al., 2024; Payandeh et al., 2025) use vision-language models for socially aware navigation. SONAR (Dell'Anna & Jamshidnejad, 2024) instead provides an adaptive norm-aware control architecture for social robots, combining explicit agent reasoning with learned contextual norms in human-robot conversations. These studies address complementary aspects of norm recognition, embodied planning, assistance, navigation, and robot-behavior assessment, but generally evaluate them as separate decision problems. AIRIS links norm-aware responses and justified inaction to the social evidence and inferred intent that precede them, enabling process-level diagnosis of both Reaction and Interaction within a single time-anchored streaming episode.

A.3 Supplementary Benchmark Details

This section provides the granular details of the AIRIS benchmark construction and evaluation. It is organized into four parts: the formal capability ontology (Section A.3.1), the supplementary evaluation protocol and metrics (Section A.3.2), the expert annotation interface (Section A.3.3), and the data construction and annotation prompts (Section A.3.4).

A.3.1 A-I-R-I capability ontology

AIRIS operationalizes social embodied intelligence as 36 separately scored capabilities: 19 in Attention, three in Intention, four in Reaction, and ten in Interaction (Tables 3-6). The decomposition is intentionally diagnostic. Attention asks what can be directly grounded in the observed scene; Intention asks what latent goal is supported by that evidence; Reaction asks what Ego should do at the current moment; and Interaction judges how the exchange actually unfolded. Thus, Interaction does not repeat the partner-intent label. Its labels describe episode phase, completion, mutual understanding, and Ego's norm compliance. These labels are adjudicated post hoc using the complete episode, although every model query is still restricted to the released causal prefix. In particular, state and violation polls use $O_{\leq \tau}$ at their query time, whereas resolved episode-level judgments use $O_{\leq T_3}$.

Tables 3-6 organize the ontology into the four consecutive stages of the A-I-R-I process. For every capability, we state its operational meaning, the observation prefix available to the model, and the concrete answer space.

Stage 1: Attention. This stage establishes whom Ego should attend to and what directly observable social cues that person exhibits. It progresses from interactant localization to cue-specific and integrated attentional grounding.

Table 3: Attention capabilities In AIRIS.

ID	Capability	Meaning	Input prefix	Answer space
A1	Interactant Localization	Localizes the person who is socially interacting with Ego. This tests whether the model can isolate the relevant participant before reasoning about finer cues.	$O_{\leq \tau_A}$	Single selection among candidate bounding boxes normalized to $[0, 1000]$.
A2	Interactant Identification and Localization	Jointly identifies the interactant by a short visual description and localizes that person. Both semantic identity and spatial grounding must be correct.	$O_{\leq \tau_A}$	Single choice among person-description and bounding-box pairs.
A3	Gaze Target Recognition	Identifies the object, person, or region indicated by the interactant's gaze or head orientation. It measures semantic gaze grounding.	$O_{\leq \tau_A}$	Single choice among scene-conditioned target descriptions.
A4	Gaze Target Localization	Localizes the target of the interactant's gaze on the current frame. This separates spatial grounding from target naming.	$O_{\leq \tau_A}$	Single selection among candidate normalized bounding boxes.
A5	Gaze Target Identification and Localization	Jointly identifies and localizes the gaze target. The selected option must contain both the correct target description and its location.	$O_{\leq \tau_A}$	Single choice among target-description and bounding-box pairs.
A6	Hand Gesture Recognition	Classifies the interactant's visible hand gesture or hand action. The label captures the form of the cue rather than its target.	$O_{\leq \tau_A}$	Single choice from the hand-gesture ontology, such as pointing, waving, reaching, holding, or no salient gesture.
A7	Hand Target Recognition	Identifies the object or person indicated, touched, or otherwise acted upon by the hand. It tests semantic grounding of manual cues.	$O_{\leq \tau_A}$	Single choice among scene-conditioned target descriptions.
A8	Hand Target Localization	Localizes the target of the observed hand gesture or action. The target is selected on the last visible frame.	$O_{\leq \tau_A}$	Single selection among candidate normalized bounding boxes.
A9	Hand Target Identification and Localization	Jointly identifies and localizes the target of the hand cue. It requires consistency between the semantic label and spatial region.	$O_{\leq \tau_A}$	Single choice among target-description and bounding-box pairs.
A10	Body Pose Recognition	Classifies the interactant's body orientation relative to Ego. This captures whether the person is facing, leaving, or turning toward the interaction.	$O_{\leq \tau_A}$	Single choice among facing Ego, facing away, turning, and side-on orientation.
A11	Body Stance Recognition	Classifies the interactant's visible body configuration. The label summarizes posture independently of the object toward which the body is oriented.	$O_{\leq \tau_A}$	Single choice from the stance ontology, such as standing, seated, leaning, crouching, or arms crossed.
A12	Body-orientation Target Recognition	Identifies what the interactant's torso or body is primarily oriented toward. It grounds body direction in a semantic target.	$O_{\leq \tau_A}$	Single choice among scene-conditioned target descriptions.
A13	Body-orientation Target Localization	Localizes the target toward which the interactant's body is oriented. It evaluates spatial rather than semantic target recovery.	$O_{\leq \tau_A}$	Single selection among candidate normalized bounding boxes.

Continued on next page.

Table 3 continued.

ID	Capability	Meaning	Input prefix	Answer space
A14	Body-orientation Target Identification and Localization	Jointly identifies and localizes the body-orientation target. The task requires agreement between target semantics and position.	$O_{\leq \tau_A}$	Single choice among target-description and bounding-box pairs.
A15	Emotion Recognition	Recognizes the affective state most visibly expressed by the interactant. Only states supported by observable facial or bodily evidence are considered.	$O_{\leq \tau_A}$	Single choice from the affect ontology, such as neutral, happy, confused, concerned, or frustrated.
A16	Overall Attention Target Recognition	Infers the interactant's dominant attentional target by integrating gaze, hand, body, and speech cues. It tests cross-cue attention resolution.	$O_{\leq \tau_A}$	Single choice among scene-conditioned target descriptions.
A17	Overall Attention Target Localization	Localizes the dominant attentional target after integrating all available cues. This is the spatial counterpart of A16.	$O_{\leq \tau_A}$	Single selection among candidate normalized bounding boxes.
A18	Overall Attention Target Identification and Localization	Jointly identifies and localizes the integrated attentional target. Both the semantic and spatial components must agree.	$O_{\leq \tau_A}$	Single choice among target-description and bounding-box pairs.
A19	Dominant Attention Modalities	Identifies which observable modalities provide the primary evidence for the current attentional focus. Multiple cues may jointly support the judgment.	$O_{\leq \tau_A}$	Multiple selection over gaze, hand movement, body posture, and speech.

Stage 2: Intention. This stage moves from observable evidence to the partner's latent goal. Its temporal questions explicitly distinguish justified inference after T_1 from premature commitment before sufficient evidence is available.

Table 4: Intention capabilities in AIRIS.

ID	Capability	Meaning	Input prefix	Answer space
I1	Key-moment Intention Recognition	Recovers the partner's latent social goal at an annotated key moment. The prediction concerns what the partner seeks to achieve, not merely the visible action.	$O_{\leq \tau_k}$	Single choice from the intent ontology, including information seeking, guidance request, object request, action request, social conversation, and other goals.
I2	Temporal Intention Recognition	Determines whether the partner's intent is already supportable at each sampled time and, when it is, recovers that intent. Before T_1, calibrated waiting is represented explicitly.	$O_{\leq \tau}$ sampled around T_1	Single choice from the intent ontology plus "Insufficient evidence to determine the intention."
I3	Intention Evidence Grounding	Identifies the observable modalities that justify the stated intent inference. This tests whether a correct intent is supported by the appropriate evidence.	$O_{\leq \tau_k}$ with $\tau_k \geq T_1$	Multiple selection over gaze, hand movement, body orientation, and speech.

Stage 3: Reaction. This stage converts the inferred social situation into an Ego-centered decision. It first asks whether intervention is warranted and then specifies the response at increasingly concrete levels.

Table 5: Reaction capabilities in AIRIS.

ID	Capability	Meaning	Input prefix	Answer space
R1	Engagement Decision	Decides whether Ego should actively respond or refrain from intervention at the current moment. It directly tests justified action versus justified inaction.	$O_{\leq \tau_k}$	Single choice between active response and refrain from intervention.
R2	Action Domain Selection	Selects the broad response modalities that Ego should employ. More than one domain may be required for a coordinated response.	$O_{\leq \tau_k}$	Multiple selection over navigation, object manipulation, verbal communication, and idle/non-intervention.
R3	Sub-action Selection	Selects the concrete actions that operationalize Ego's response. The labels distinguish, for example, approaching, yielding, grasping, speaking, observing, and taking no action.	$O_{\leq \tau_k}$	Multiple selection from the embodied sub-action ontology, with an additional "Other action" option.
R4	Speech Function Selection	Determines the communicative function of Ego's response when speech is warranted. It evaluates pragmatic response selection rather than free-form generation.	$O_{\leq \tau_k}$	Single choice among greeting, confirmation, clarification question, information response, recommendation, warning, farewell, and empathic response.

Stage 4: Interaction. This stage evaluates the progression and consequences of the exchange. Online state and violation probes use the current prefix, while resolved episode-level attributes assess completion, mutual understanding, and Ego's social compliance from the prefix ending at T_3.

Table 6: Interaction capabilities in AIRIS.

ID	Capability	Meaning	Input prefix	Answer space
N1	Interaction State Recognition	Tracks the current phase of the exchange rather than inferring the partner's goal. It indicates whether social engagement is beginning, active, ending, or absent.	$O_{\leq \tau}$	Single choice among pre-engagement, active engagement, disengagement, and no social engagement.
N2	Online Violation Detection	Detects whether Ego is exhibiting a social violation at the current time. This provides an online binary compliance signal before episode resolution.	$O_{\leq \tau}$	Binary choice between violation present and no violation.
N3	Online Violation Type Recognition	Identifies the type or types of Ego violation visible at the current time. Multiple simultaneous violations can be selected.	$O_{\leq \tau}$	Multiple selection from the spatial, navigation, manipulation, physical-contact, and communication violation ontology.
N4	Episode-level Violation Recognition	Summarizes all Ego violation types that occurred anywhere in the resolved episode. It aggregates event-level compliance evidence without re-inferring partner intent.	$O_{\leq T_3}$	Multiple selection from the complete violation ontology, including "No violation."
N5	Interaction Completeness	Judges whether the exchange formed a complete social and task-level loop. Incomplete episodes distinguish external interruption from unilateral abandonment.	$O_{\leq T_3}$	Single choice among completed, interrupted, and abandoned.
N6	Mutual Understanding	Judges whether Ego and the interactant reached a shared understanding of the request, response, or closure. Completion alone does not guarantee mutual understanding.	$O_{\leq T_3}$	Binary choice between mutual understanding achieved and not achieved.

Continued on next page.

Table 6 continued.

ID	Capability	Meaning	Input prefix	Answer space
N7	Spatial Compliance	Assesses whether Ego's approach, stopping distance, route, and use of shared space were socially compliant. If not, it identifies the principal spatial violation.	$O_{\leq T_3}$	Single choice between compliant behavior and one applicable spatial or navigation violation category.
N8	Physical Compliance	Assesses whether Ego's manipulation and physical-contact behavior was safe, controlled, and respectful of bodily boundaries. If not, it identifies the principal physical violation.	$O_{\leq T_3}$	Single choice between compliant behavior and one applicable manipulation or physical-contact violation category.
N9	Communication Compliance	Assesses conversational and nonverbal etiquette across the exchange. It can capture several communication violations when they co-occur.	$O_{\leq T_3}$	Multiple selection from the communication-violation ontology, including interrupting, prolonged staring, impolite tone, ignored queries, and missing acknowledgement.
N10	Overall Compliance Severity	Assigns the overall severity of Ego's social compliance after considering the complete exchange. This is an outcome-level judgment rather than an intent label.	$O_{\leq T_3}$	Single choice among fully compliant, partial violation, and serious violation.

For evaluation, single-choice, binary, and candidate-box questions use option-key accuracy. Multi-select questions use set F1 as the primary score and additionally report exact match. For Temporal Intention Recognition, the event-level Intent Anticipation Score rewards the earliest stable correct prediction after T_1 and discounts it by the false-commitment rate before T_1. For the six resolved Interaction attributes N5-N10, we additionally report the mean per-segment component score and joint exact match when all six labels are available.

A.3.2 Supplementary Evaluation Protocol and Metrics

Building upon the core evaluation protocol in Section 2.3, this section defines the complete set of metrics used to isolate fine-grained reasoning bottlenecks across the 36 capabilities of AIRIS. These metrics are reported in the detailed evaluation tables (Section A.4.2) to prevent surface-level accuracy from masking deeper failure modes.

Scoring and aggregation. Base items are scored by exact match (single-choice, yes/no, bounding-box) or set F1 (multi-choice). Empty, unparseable, and out-of-vocabulary responses score 0. The headline metric is the unweighted mean of the four question-micro layer scores: $\text{Overall} = \frac{1}{4}(S_{\text{Att}} + S_{\text{Int}} + S_{\text{Rea}} + S_{\text{Inter}})$.

Attention Metrics:

- *Semantic Target & Spatial Localization: Semantic target* assesses which entity a social cue (e.g., gaze, pointing) refers to. *Spatial localization* evaluates its grounding over candidate boxes.
- *Joint Identity-Localization:* A strict metric requiring both the semantic class and the spatial bounding box to match the reference simultaneously. If either fails, the joint score is 0. This penalizes models that name a referent without localizing it, yielding a non-executable percept.
- *Nonverbal States:* Accuracy on classifying body pose, physical stance, hand gesture, and observable emotion.
- *Multimodal Attribution:* Set F1 score for identifying which observable modalities (gaze, hand movement, body posture, speech) provide the primary evidence for the current attentional focus.

Intention Metrics:

- *Intent Anticipation Score (IAS):* Evaluates streamed intent recovery. Before the earliest supportable moment T_1, "insufficient evidence" is the correct label. Let FCR be the False Com-

mitment Rate (the fraction of pre-T_1 queries prematurely committing to a specific intent). For models reaching stable correctness at timestamp $p \geq T_1$, we define an anticipation margin $m = \text{clip}((T_2 - p)/(T_2 - t_f), 0, 1)$, where t_f is the first query at or after T_1 (with $m = 0$ if stability is never reached). The event-level score is $\text{IAS}_{\text{event}} = m \cdot (1 - \text{FCR})$, and IAS is its macro average.

- *Key-Moment & Online Accuracy: Key-moment* evaluates offline accuracy exactly at the true onset of the partner's substantive response (T_2), or at human-annotated keyframes. *Online* evaluates accuracy across sampled temporal prefixes.
- *Evidence Grounding:* Set F1 score for identifying which specific observable cues (visual, vocal, or verbal) justify the inferred intent.

Reaction Metrics:

- *Engage:* Macro-F1 between active intervention and idle/waiting classes. Chosen over plain accuracy to penalize collapsed policies on heavily imbalanced classes.
- *Action Domain & Subaction: Action domain* and *Subaction* use multi-label F1 to evaluate the broad response modalities and concrete actions Ego should employ. *Speak subtype* uses accuracy for communicative functions.
- *Joint-EM:* A strict diagnostic for active responses requiring the action domain, subaction, and any annotated speech subtype to match the reference simultaneously.
- *Domain-Specific Diagnostics:* Breaks down the Subaction + Speak scores by the required action domain (Communication, Navigation, Manipulation, Refrain/Idle) to isolate modality-specific performance.

Interaction Metrics:

- *State & Task Outcome: State* tracks the current phase of the exchange (e.g., active engagement, disengagement). *Task outcome* judges whether the underlying social exchange succeeded, failed, or was aborted, and whether mutual understanding was achieved.
- *Compliance:* Equally averages spatial etiquette, physical demeanor, communication etiquette, and overall compliance category ($\tau \geq T_3$).
- *Violation:* Equally averages online violation-presence F1, online violation type, and segment-level violation types. "Violation present" is the positive class to prevent exploitation of the skewed negative baseline.

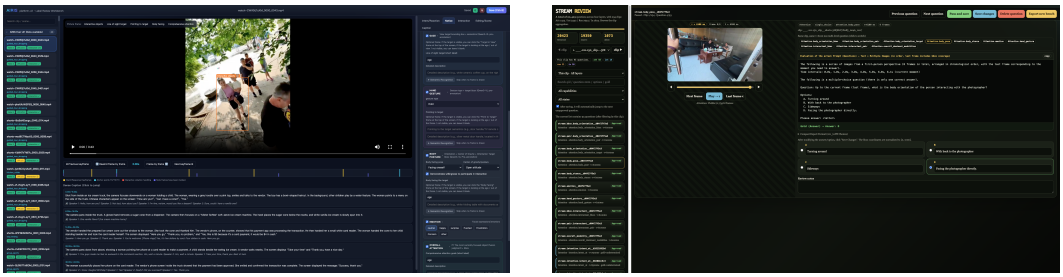
A.3.3 Expert Annotation Interface and Process

To ensure the high quality and reliability of the AIRIS benchmark, the dataset construction relies on a rigorous human-in-the-loop annotation protocol. A total of 11 human experts, who underwent multiple rounds of training and had their capabilities rigorously verified, participated in the annotation and review process. As shown in Figure 6, we developed two specialized interfaces to facilitate this workflow.

The first interface (Figure 6a) is designed for the structured A-I-R-I annotation. It allows annotators to review the video stream, identify critical temporal anchors (T_0 to T_3), and provide structured labels for the partner's intention and the ego-agent's reaction based on observable evidence. The second interface (Figure 6b) is dedicated to question auditing. In this stage, experts review the generated question-answer pairs, distractors, and temporal constraints to ensure that every evaluation item is unambiguous, uniquely answerable from the causal prefix, and free from future-information leakage.

A.3.4 Annotation Prompts

The construction of AIRIS relies on a multi-stage annotation pipeline. To ensure reproducibility and transparency, we provide the core prompt templates used during the data curation and pre-annotation phases. Note that all model outputs serve strictly as initial candidates (pre-annotations); the final dataset relies exclusively on human verification and adjudication.



(a) Structured A-I-R-I annotation interface. (b) Question review and audit interface.

Figure 6: AIRIS annotation interfaces used by human experts during the data annotation process.

Prompt 1: Long Video Audit, Scene Classification, and Interaction Extraction. This prompt is applied to raw, first-person web videos. It acts as the initial gatekeeper, enforcing strict quality and compliance rules (e.g., egocentric perspective, minimum resolution, safe content). For eligible videos, it classifies the scene and extracts continuous social interaction segments, identifying the start and end boundaries as well as the initiator of the interaction.

```
### Role
You are an expert AI Video Auditor and HRI (Human-Robot Interaction) Researcher. Your
     task is to audit video quality and compliance and, if the video is eligible,
     classify the scene and extract all social interaction segments into a single
     structured report.

### Section 1: Quality and Compliance Gatekeeping
Before segmenting the video, evaluate it against all of the following rules:
- [L0-1] Perspective: The video MUST primarily use an egocentric perspective, such as a
     head-mounted, chest-mounted, or equivalent first-person view.
- [L0-2] Editing: Reject the video if its average shot length is shorter than 3 seconds,
     or if subtitles, stickers, transitions, or other overlays occlude more than 30\% of
     its duration.
- [L0-3] Technical quality: Reject the video if its resolution is lower than 720p or its
     frame rate is lower than 24 fps.
- [L0-4] Compliance: Reject the video if it contains violence, gore, adult content, or
     close-ups of minors' faces.
- [L0-5] Relevance: Reject the video if no human-human interaction occurs, such as a pure
     monologue, unboxing, or scenery-only video.

### Section 2: Scene Classification
Classify the whole video using the following fields:
- `primary_category`: Select exactly one value from: `guided_tour_shopping`,
     `kitchen_collab`, `way_finding`, `group_sports`, `home_interaction`, `other`. Use
     `other` only when none of the preceding categories applies.
- `sub_scene`: Create a short snake_case label for the specific setting, such as
     `mall_restroom_wayfinding`.
- `scene_description`: Write one English sentence describing where the interaction
     occurs, who participates, and what happens.

### Section 3: Interaction Extraction
Apply this section only if the video passes all required quality and compliance checks.
1. Continuity principle: Treat interactions with the same group at the same location as
     one continuous segment. Do not over-segment due to brief cuts or topic changes.
2. Start boundary: The interaction starts when joint attention is established, for
     example through a greeting, eye contact, an approach, or an Ego-initiated question.
3. End boundary: The interaction ends when the social contract dissolves, for example
     through a farewell, one party walking away, attention clearly shifting away, or a
     hard cut to a new location.
4. Initiator: For every segment, identify who opened the interaction: `ego`, `other`,
     `mutual`, or `unclear`. Provide short evidence based on observable speech, gesture,
     movement, or a timestamp.
5. Evidence: Use only observable visual or audio evidence. Do not invent intentions,
     relationships, identities, or events that are not supported by the video.

### Output Requirements
Return strictly one valid JSON object. Do not include Markdown fences or explanatory text.
{
  "video_audit": {
    "audit_passed": true,
    "failed_codes": [],
```

```
    "audit_justification": "Detailed explanation of why the video passed or failed the
        rules."
  },
  "scene": {
    "primary_category": "way_finding",
    "sub_scene": "mall_restroom_wayfinding",
    "scene_description": "One sentence describing where, who, and what is happening."
  },
  "interaction_analysis": {
    "total_interactions_found": 1,
    "interaction_segments": [
      {
        "start_time": "00:01",
        "end_time": "00:18",
        "initiator": "ego",
        "initiator_evidence": "Ego says 'Excuse me' at 00:01 before the other participant
            speaks.",
        "participants": "A short, non-identifying description of the visible
            participants.",
        "interaction_summary": "A concise description of the social exchange.",
        "boundary_logic": "Evidence supporting the selected start and end timestamps."
      }
    ]
  }
}
```

Prompt 2: Clip-Level Dense Visual Captioning and Audio Transcription. Once interaction clips are extracted and manually verified, this prompt generates chronological, dense descriptions of the visual scene and verbatim audio transcripts. This provides downstream annotation stages with a rich, text-based grounding of the multimodal context.

```
You are an expert video analyst and professional transcriptionist. Your task is to
    perform dense video captioning and speech-to-text transcription on the provided
    video.

Analyze the video from start to finish and divide it into chronological segments. Create
    a new segment whenever there is a significant visual scene change, a new action
    begins, or a new speaker starts talking.

For every segment, extract the following information:

1. Time range
   - Provide the start and end time in `MM:SS` format.

2. Dense visual caption
   - Describe the setting and environment.
   - Describe visible people using non-identifying terms and report their observable
       actions, gestures, body orientation, gaze direction, and facial expressions.
   - Describe relevant objects and their visible state.
   - Describe notable camera movement or viewing angle.
   - Transcribe visible on-screen text when legible.
   - Report only observable information. Do not infer identity, protected attributes,
       private relationships, or mental state without direct evidence.

3. Audio transcript
   - Transcribe spoken words verbatim when intelligible.
   - Distinguish speakers with neutral labels such as `Speaker 1` and `Speaker 2` when
       possible.
   - Include significant non-speech sounds in brackets, such as `[door closes]`,
       `[background music]`, or `[silence]`.
   - If speech is absent or unintelligible, return `null` or an empty string.

### Output Requirements
Return strictly one valid JSON array. Do not include Markdown fences or explanatory text.

[
  {
    "start_time": "00:00",
    "end_time": "00:15",
    "visual_caption": "Detailed, evidence-based description of the visible scene.",
    "audio_transcript": "Speaker 1: Example transcription. [door closes]"
  }
]
```

Prompt 3: Intention and Reaction VLM-based Pre-Annotation. This prompt identifies the core temporal anchors of the interaction and selects key decision frames. For each key frame, it generates structured candidate annotations for the partner's intention and the normatively appropriate reaction that Ego should take at that moment. The Reaction target is a decision grounded in the causal observation prefix, not a transcription of Ego's observed behavior. The output is strictly evidence-based and serves as a starting point for human adjudication.

```
# Role
You are a senior annotation expert for the social embodied intelligence benchmark. Your
     task is to process a given first-person interaction video clip and in one pass:
1. Locate key temporal anchors of the social interaction;
2. Select key decision frames;
3. Generate structured candidate annotations for P1's intention and what Ego should do at
     each key decision frame.

Your output serves only as an initial pre-annotation and must be reviewed and corrected
     by humans. Therefore, you must judge conservatively, provide observable evidence,
     and avoid fabrication.

# Perspective Conventions
- Ego: The Camera Wearer, treated as the robot's perspective.
- P1: The other person primarily interacting with Ego.
- intention: Always describes P1's intent, i.e., what P1 wants to do or wants Ego to do.
- reaction: Always describes the response Ego should take at the current moment, given
     only the observable evidence available up to that moment.
- Ego's observed response is used only to locate descriptive temporal anchors such as T2.
     It must not be treated automatically as the correct Reaction target.

# I. Temporal Anchors
- T0, Attention Onset: The first moment P1's implicit attention cue relates to Ego, e.g.,
     starting to turn head, orient body, or approach Ego. P1 does not necessarily look
     directly at Ego yet.
- T1, Intent Decidable Point: The first moment P1's intent can be clearly determined
     based on video cues, e.g., the start of an action with clear semantics, or the first
     syllable expressing the intent.
- T2, Observed Ego Reaction Onset: The frame where Ego first begins to make an observable
     reaction to the intent identified at T1. This is a descriptive temporal anchor, not
     a judgment that the observed reaction is normatively appropriate. The reaction can
     be verbal, an action, a posture change, or movement. If Ego never reacts, T2 is null.
- T3, Interaction End: The moment the current interaction event is completely finished,
     e.g., farewell, separating, or a clear shift in attention.

# II. Key Frames
Key frames are moments where Ego needs to make a social decision. They may coincide with
     an observed Ego action, but the Reaction label remains a normative decision based on
     the causal prefix.
Typical trigger types: `state_transition`, `intent_change`, `reaction_switch`,
     `initial_engagement`, `closing`.
Key frame principles:
- Recommend generating 3 to 5 candidate frames per minute.
- Candidate frames should generally fall between T0 and T3.
- Must cover around T0, around T2, and the intermediate stage between T2 and T3. If T2 is
     null, select the frame at which the response decision is most informative; do not
     assume that the absence of an observed response is normatively correct.
- For videos under 15s, recommend 2-4 frames; 15-30s, recommend 3-5 frames; over 30s,
     recommend 1 frame every 6-8 seconds.
- For each key frame at time `tMs`, infer intention and Reaction using only visual and
     audio evidence available at or before `tMs`. Do not use later events to justify
     either label.

# III. Intention Fields
## state
Select one from the following closed enumeration:
- `pre_engagement`: Trying to attract attention or preparing to interact;
- `active_engagement`: Currently interacting;
- `disengagement`: Ending or leaving;
- `idle_ignore`: Ignoring Ego, focusing on self, or passing by.

## sceneIntent
Select one from the following enumeration:
- `ask_information`: Asking for factual information, excluding directions;
- `request_guide`: Asking for directions, routes, or location guidance;
- `request_item`: Asking for or requesting an item to be handed over;
- `request_action`: Requesting Ego to help perform an action;
- `offer_help`: P1 proactively offering help, items, or information to Ego;
- `small_talk`: Greeting, chatting, thanking, or other social confirmations;
```

```
- `complain`: Complaining or expressing dissatisfaction;
- `none_passing`: No interaction intent, just passing by or focusing on self;
- `other`: Does not belong to the above types, needs to be supplemented in
     `sceneIntentFreeText`.

## evidenceModalities
Select at least one observable evidence: `gaze_head`, `hand`, `body`, `speech`.

## rationale
Use 1 to 2 sentences to explain why this intent was determined based on observable
     evidence.

# IV. Reaction Fields
Annotate the socially appropriate response that Ego should take now, using only the
     causal observation prefix ending at the key frame. The target may differ from Ego's
     observed behavior. Do not copy an observed action merely because it occurred, and do
     not use a future observed response as evidence for the current decision.

## domain
Select 1 to 2 action domains: `navigation`, `manipulation`, `communication`, `idle`.
`idle` must appear alone, indicating that Ego should currently wait, continue observing,
     or refrain from intervention.

## subaction
Prioritize using keys from the vocabulary below; only use a free phrase of 10-20 words if
     they cannot cover it.
- navigation: `approach`, `stop`, `yield`, `follow`, `retreat`.
- manipulation: `reach`, `grasp`, `release`, `handover_give`, `handover_receive`,
     `point`, `assist_hold`.
- communication: `acknowledge`, `speak`, `gaze_shift`, `gesture_response`, `silent_wait`.
- idle: `no_action`, `continue_task`, `observe`.

## speakSubtype
Only select from the following enumeration when Ego's reaction includes speaking:
`greeting`, `confirmation`, `clarification_question`, `information_response`,
     `recommendation`, `warning`, `farewell`, `empathy_response`.

## utterance
When speech is warranted, provide a brief example of what Ego should say. Do not
     transcribe the observed utterance unless it is also an appropriate response under
     the current causal prefix; otherwise, omit.

# V. Input Information
[Video Clip] Passed directly as multimodal input.
[Duration] {{duration_sec}} seconds
[Scene Description] {{scene_description}}
[Interaction Summary] {{interaction_summary}}
[Dense Visual Caption] {{dense_caption}}
[Audio Transcript] {{audio_transcript}}

# VI. Output Requirements
Strictly output a valid JSON object, do not output Markdown or extra explanations.
{
  "anchors": {
    "t0": {
      "tMs": 3500,
      "evidence": "Evidence of P1's attention first relating to Ego"
    },
    "t1": {
      "tMs": 5200,
      "evidence": "Evidence of P1's intent first becoming clear"
    },
    "t2": {
      "tMs": 7300,
      "evidence": "Ego's first observable reaction to the intent"
    },
    "t3": {
      "tMs": 11500,
      "evidence": "Observable sign of interaction ending"
    }
  },
  "keyFrameCandidates": [
    {
      "id": "kf_1",
      "tMs": 5200,
      "trigger": "initial_engagement",
      "briefDescription": "One sentence explaining the key event at this frame",
      "recommendedForAnnotation": true,
```

```
      "priority": 1,
      "intention": {
        "state": "pre_engagement",
        "sceneIntent": "request_guide",
        "sceneIntentFreeText": "",
        "evidenceModalities": ["gaze_head", "body"],
        "rationale": "Explain P1's intent based on observable clues."
      },
      "reaction": {
        "domain": ["communication"],
        "subaction": "acknowledge",
        "speakSubtype": "greeting",
        "utterance": "",
        "rationale": "Explain why this is what Ego should do now, based only on the
             causal observation prefix."
      }
    }
  ],
  "initiator": {
    "value": "ego",
    "evidence": "Explain who initiated the interaction, citing observable clues or
         timestamps."
  },
  "confidence": {
    "anchors": 0.0,
    "keyFrames": 0.0,
    "intention": 0.0,
    "reaction": 0.0,
    "overall": 0.0
  },
  "uncertainty_notes": ""
}

# VII. Judgment Rules
1. Anchor order should generally satisfy T0 < T1 < T2 < T3.
2. T2 must occur after T1, as T2 descriptively represents Ego's first observed reaction
     to P1's intent. T2 does not define the correct Reaction label.
3. If Ego has an action before T1, but it is not a reaction to this P1 intent, it must
     not be marked as T2.
4. Determine the anchors in forward temporal order (first T0, then T1, then T2, then T3).
     Fix T1 as the earliest moment at which P1's intent becomes decidable from the
     evidence available up to that moment alone, judging strictly as an online observer
     who does not yet know how the interaction unfolds.
5. If Ego has no reaction throughout, T2 is null, and explain in `uncertainty_notes`.
6. intention can only describe P1; reaction can only describe what Ego should do at the
     current key frame.
7. When involving directions, routes, or locations, prioritize `request_guide` over
     `ask_information`.
8. Select `request_action` when requesting Ego to perform an action; select
     `request_item` when only asking for an item.
9. When `state=idle_ignore`, `sceneIntent` must be `none_passing`, and `reaction.domain`
     must be `["idle"]`.
10. When the appropriate response combines speech with an action such as pointing or
     handing over, two reaction domains can be selected.
11. priority values: 1 for interaction state transition, 2 for Ego response strategy
     switch, 3 for supplementary sampling during state continuation.
12. If there is no substantive interaction in the video, all interaction anchors can be
     null, and only output 1 to 2 keyframes to explain the decision not to respond.
13. When clues are insufficient, lower the confidence and explain the uncertainty, do not
     fabricate.
14. Never infer the Reaction target from events after the key-frame timestamp. If the
     current prefix does not yet justify intervention, waiting, observation, or
     non-intervention should be preferred.
```

Prompt 5: Interaction Multi-Model VLM-based Annotation. This final prompt conducts a post-hoc assessment of the entire interaction. It evaluates whether the exchange was completed successfully and whether the ego-agent's navigation, manipulation, and communication behaviors complied with social norms. It also localizes any identified violations to specific timestamps.

```
# Role
You are a senior annotation expert for the AIRIS social embodied intelligence benchmark.
     Your task is to conduct a post-hoc social compliance assessment on a given
     first-person interaction video clip:
```

```
1. Judge whether the interaction is complete;
2. Judge whether Ego's navigation, manipulation, and communication comply with social
     norms;
3. Mark observable violation time points;
4. Provide a comprehensive evaluation of the entire interaction.

Your output serves only as an initial pre-annotation and must be reviewed and corrected
     by humans. Therefore, you must judge conservatively, provide observable evidence,
     and avoid fabrication.

# Perspective Conventions
- Ego: The Camera Wearer, treated as the robot's perspective.
- P1: The other person primarily interacting with Ego.
- The object of social compliance assessment is Ego's behavior.
- P1's stepping back, dodging, pausing, repeated questioning, or other reactions can
     serve as evidence to judge whether Ego's behavior is compliant, but must not be used
     to infer P1's identity or unobservable mental states.

# Judgment Principles
Judge whether Ego's behavior might make a typical interactant feel uncomfortable,
     pressured, offended, ignored, or unsafe, based solely on observable clues in the
     video and audio.
When clues are insufficient, you should lower confidence and explain the uncertainty, and
     must not fabricate violations or force a judgment of compliance.

# D1: Interaction Completeness
## completeness
Select from the following enumeration:
- `complete`: The interaction naturally starts, progresses, and concludes;
- `interrupted`: The interaction is interrupted midway by external events or one party
     suddenly leaving;
- `abandoned`: One party unilaterally gives up, and the interaction terminates before the
     goal is reached.

## mutualUnderstanding
Judge whether both parties reached a mutual understanding, e.g., P1's need was correctly
     received and responded to by Ego, or both parties clearly know the interaction has
     ended.

# D2: Navigation Compliance
Judge whether Ego's approach, movement, stopping position, and route are compliant.
When `compliant=false`, select from the following violation types:
- `abrupt_approach`: Approaching suddenly or too fast;
- `crowding`: Entering the other person's personal distance without a valid reason;
- `path_blocking`: Blocking the other person's normal path of travel;
- `unsafe_speed`: Movement speed poses a clear safety risk;
- `wrong_stop_distance`: Stopping distance is unsuitable for normal interaction;
- `nav_other`: There is indeed a navigation violation but it does not belong to the above
     types, must be explained in the note.

P1 leaning back, stepping back, dodging sideways, or suddenly interrupting an action can
     serve as potential clues, but must be judged in combination with the full context.

# D3: Manipulation Compliance
Judge whether Ego's reaching, handing over, placing, and other embodied manipulations are
     safe, controlled, and respect bodily boundaries.
When `compliant=false`, select from the following violation types:
- `sharp_toward_other`: Sharp objects or points directed at the other person;
- `unstable_handover`: Unstable handover, with a risk of dropping;
- `startling_grasp`: Grasping or pick-and-place actions are sudden or too forceful;
- `body_space_invasion`: Operating within the other person's body or personal space
     without permission;
- `premature_release`: Letting go too early before the other person has a firm grip;
- `manip_other`: There is indeed a manipulation violation but it does not belong to the
     above types, must be explained in the note.

# D4: Communication Compliance
Judge whether Ego's verbal and non-verbal communication complies with basic etiquette and
     turn-taking norms.
When `compliant=false`, select from the following violation types:
- `interrupting`: Interrupting or talking over before the other person finishes speaking;
- `prolonged_staring`: Staring at the other person for a long time without normal verbal
     or physical communication;
- `impolite_tone`: Tone is clearly impolite, commanding, or lacks necessary politeness;
- `ignoring_query`: Ignoring the other person's explicit question or request;
- `missing_acknowledgement`: Lacking necessary confirmation or response signals;
- `comm_other`: There is indeed a communication violation but it does not belong to the
     above types, must be explained in the note.
```

```
# D5: Overall Compliance Score
## overallScore
From P1's experience perspective, give an integer score from 1 to 5 for the comfort and
     appropriateness of the entire interaction:
- 1: Seriously non-compliant or has obvious safety/offense issues;
- 2: Has relatively obvious violations, poor interaction experience;
- 3: Some aspects are not appropriate enough, but the overall interaction can still be
     completed;
- 4: Overall compliant, with only minor issues;
- 5: Complete, natural, appropriate, and without obvious issues.

## overallCategory
Select from the following enumeration:
- `fully_compliant`
- `partial_violation`
- `serious_violation`
As long as any of D2, D3, or D4 is `compliant=false`, `overallCategory` cannot be
     `fully_compliant`.

# violationPoints
Locate each confirmed violation in D2 to D4 to a specific time point as much as possible.
Each time point contains:
- `id`
- `tMs`
- `domain`: `navigation`, `manipulation`, or `communication`
- `type`: Must belong to the violation enumeration of the corresponding domain
- `note`: Describe the observable violation clues at this time point
If there are no violations, return an empty array.

# Input Information
[Video Clip] Passed directly as multimodal input.
[Duration] {{duration_sec}} seconds
[Scene Description] {{scene_description}}
[Interaction Summary] {{interaction_summary}}
[Dense Visual Caption] {{dense_caption}}
[Audio Transcript] {{audio_transcript}}
[Prior Intent/Reaction Summary] {{prior_anchor_summary}}

# Output Requirements
Strictly output a valid JSON object, do not output Markdown or extra explanations.
{
  "d1": {
    "completeness": "complete",
    "mutualUnderstanding": true,
    "note": "Basis for interaction completeness judgment."
  },
  "d2": {
    "compliant": true,
    "violationTypes": [],
    "note": "Basis for navigation compliance judgment."
  },
  "d3": {
    "compliant": true,
    "violationTypes": [],
    "note": "Basis for manipulation compliance judgment."
  },
  "d4": {
    "compliant": true,
    "violationTypes": [],
    "note": "Basis for communication compliance judgment."
  },
  "d5": {
    "overallScore": 5,
    "overallCategory": "fully_compliant"
  },
  "violationPoints": [
    {
      "id": "vp_1",
      "tMs": 3200,
      "domain": "navigation",
      "type": "abrupt_approach",
      "note": "Describe the observable violation clues at this time point."
    }
  ],
  "rationale": "Use 3 to 4 sentences to comprehensively explain the judgment basis for
       the entire interaction.",
  "confidence": {
```

```
    "d1": 0.0,
    "d2": 0.0,
    "d3": 0.0,
    "d4": 0.0,
    "d5": 0.0,
    "overall": 0.0
  },
  "uncertainty_notes": ""
}

# Judgment Rules
1. All judgments are based on observable clues, do not infer identity, relationship, or
     unobservable mental states.
2. When a dimension is `compliant=true`, its `violationTypes` must be an empty array.
3. `violationPoints.type` must belong to the violation enumeration of the corresponding
     domain.
4. Each violation point must correspond to a violation type already selected in D2, D3,
     or D4.
5. Each selected violation type should in principle correspond to at least one violation
     point; if it cannot be located, it must be explained in the note of that dimension.
6. `nav_other`, `manip_other`, or `comm_other` are only used when existing types cannot
     cover it, and the note is required.
7. `serious_violation` corresponds to 1 to 2 points, `partial_violation` corresponds to 2
     to 4 points, `fully_compliant` corresponds to 4 to 5 points.
8. When there is a clear safety risk, `serious_violation` should be prioritized.
9. When a certain action domain does not occur, mark that domain as compliant, the
     violation array as empty, and you can explain "not applicable" in the note.
10. When there is no substantive interaction, D2 to D4 default to no violation, but it
     must be explained in `uncertainty_notes` that the clip lacks substantive interaction.
11. When clues are insufficient, lower the confidence and explain the uncertainty, do not
     fabricate.
```

A.4 Supplementary Experiments and Results

A.4.1 Model evaluation pool

AIRIS targets a broad model pool spanning proprietary multimodal APIs, general-purpose open-weight models, streaming video-language models, and embodied foundation models. The pool below contains 45 endpoints or checkpoints; the main result table reports only models with completed, validated inference records. Superscript footnotes after each model name give the official provider documentation for proprietary systems and an official upstream Hugging Face repository, or the maintained conversion corresponding to the evaluated checkpoint, for open-weight systems. All links were checked on September 25, 2026. The descriptions distinguish native video or audio support from image-only interfaces, for which temporally ordered frames are required. Model names used by the evaluation harness are retained when they differ from the canonical public identifier.

Proprietary models. These models are accessed through provider APIs and do not have public Hugging Face weights.

- **Gemini 3.7 Flash**[1] (Google DeepMind, 2026b) is Google's natively multimodal, efficiency-oriented reasoning model. Its public API accepts text, image, video, audio, and PDF inputs with a 1,048,576-token input limit.
- **Gemini 3.1 Pro**[2] (Google DeepMind, 2026a) is the higher-compute Gemini endpoint for complex multimodal reasoning over text, images, video, audio, and PDFs.
- **GPT-5.6 Sol**[3] (OpenAI, 2026) is OpenAI's flagship GPT-5.6 API model, with a 1,050,000-token context window. Its public API supports text and image input but not audio or video input, so video evaluation requires an external frame representation.
- **Doubao-Seed-2.0-Lite**[4] (ByteDance Seed Team, 2026) is ByteDance's efficiency-oriented proprietary model for high-frequency production workloads. The updated endpoint natively unifies text,

[1] https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash
[2] https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview
[3] https://developers.openai.com/api/docs/models/gpt-5.6-sol
[4] https://seed.bytedance.com/zh/seed2

image, video, and audio understanding, including joint audiovisual reasoning and temporal event localization, with a 256K-token context window.

General-purpose open-weight models. This group covers native vision-language checkpoints as well as image-language models evaluated through ordered video frames.

- **Qwen3.6-35B-A3B**[5] (Qwen Team, 2026c) is a native image-and-video model with 35B total parameters and 3B activated parameters per token. It combines a vision encoder with a hybrid Gated-DeltaNet and attention MoE language backbone.
- **Gemma-4-31B-it**[6] (Team et al., 2026a) is Google's 30.7B dense instruction-tuned model with text and image inputs. Its official processor handles video as a sequence of frames, while this 31B variant does not include the audio encoder available in smaller Gemma 4 variants.
- **DeepSeek-V4.1-Flash**[7] (Xu et al., 2026a) is a 552B-backbone multimodal MoE with native text-and-image input, no documented native video route, and a context window of up to one million tokens.
- **Qwen3.6-27B**[8] (Qwen Team, 2026b) is the dense 27B member of the Qwen3.6 family. It natively accepts images and videos and provides a 262,144-token context window that can be extended beyond one million tokens.
- **Qwen3.5-35B-A3B**[9] (Qwen Team, 2026a) is a 35B-total, 3B-active MoE trained as a unified vision-language model. Its early-fusion interface supports both image and video inputs.
- **Qwen3-VL-32B-Instruct**[10] (Bai et al., 2025a) is an approximately 33B instruction-tuned VLM designed for spatial and long-video reasoning. It uses a native 256K context and timestamp-aware video positional encoding.
- **Qwen3.5-27B**[11] (Qwen Team, 2026a) is the dense 27B Qwen3.5 checkpoint with a vision encoder. It shares the family's unified image-and-video interface and long-context support.
- **Qwen3.8-27B**[12] (Qwen Team, 2026d) is a dense 27B native vision-language model derived from the Qwen3.5 architecture. Its public checkpoint supports images, videos, and controllable reasoning.
- **Qwen2.5-VL-7B-Instruct**[13] (Bai et al., 2025b) is an instruction-tuned VLM with dynamic image resolution and dynamic frame-rate sampling. Absolute-time alignment in its multimodal positional encoding supports temporal localization in video.
- **InternVL3-8B**[14] (Zhu et al., 2025) is an 8B-class model trained with interleaved text, image, and video data. Its video interface samples frames and packs them as a multi-image sequence with frame identifiers.
- **InternVL3.5-30B-A3B**[15] (Wang et al., 2025a) combines a 0.3B vision component with a 30.5B MoE language component, for 30.8B total parameters and roughly 3B active parameters. Video is processed as a sampled sequence of dynamically tiled frames.
- **InternVL3.5-8B**[16] (Wang et al., 2025a) combines a 0.3B vision encoder with an 8.2B language model. It supports image, multi-image, and sampled-frame video conversations.
- **Qwen2.5-VL-3B-Instruct**[17] (Bai et al., 2025b) is the smaller Qwen2.5-VL checkpoint, retaining dynamic resolution, visual grounding, and variable-frame-rate video input. It provides a resource-efficient comparison to the 7B variant.

[5] https://huggingface.co/Qwen/Qwen3.6-35B-A3B
[6] https://huggingface.co/google/gemma-4-31B-it
[7] https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash
[8] https://huggingface.co/Qwen/Qwen3.6-27B
[9] https://huggingface.co/Qwen/Qwen3.5-35B-A3B
[10] https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct
[11] https://huggingface.co/Qwen/Qwen3.5-27B
[12] https://huggingface.co/Qwen/Qwen3.8-27B
[13] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
[14] https://huggingface.co/OpenGVLab/InternVL3-8B
[15] https://huggingface.co/OpenGVLab/InternVL3_5-30B-A3B
[16] https://huggingface.co/OpenGVLab/InternVL3_5-8B
[17] https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct

- **LLaVA-OneVision-Qwen2-7B**[18] (Li et al., 2024a) couples a Qwen2 language model with the SigLIP SO400M vision encoder. It is trained for single-image, multi-image, and video scenarios and exposes all three through one checkpoint.
- **Qwen3.5-4B**[19] (Qwen Team, 2026a) is a native vision-language model with a 4B language backbone and image and video input. It offers the same hybrid linear-attention and full-attention design as larger Qwen3.5 variants at a smaller scale.
- **Qwen3.5-2B**[20] (Qwen Team, 2026a) has a 2B language backbone and preserves the Qwen3.5 family's unified image-and-video interface for low-compute evaluation.
- **Llama-3.2-11B-Vision-Instruct**[21] (Meta, 2024) is Meta's 10.6B-parameter instruction-tuned image-reasoning model with text-and-image input and text output. It has no native video interface, so videos must be represented through extracted images.
- **Qwen3.5-0.8B**[22] (Qwen Team, 2026a) is the smallest Qwen3.5 checkpoint in the pool, with a 0.8B language backbone. It retains the family's native image and video input interface.
- **GLM-5.3-Flash**[23] (Z.ai, 2026) is a native multimodal MoE with 320B total parameters and 18B activated parameters. Its hybrid sparse- and linear-attention architecture supports text, image, video, and file inputs with text output; the provider does not document native audio input.

Streaming video-language models. These checkpoints are designed for online or long-running video rather than only offline clip-level question answering.

- **JoyAI-VL-Interaction-Preview**[24] (Yao et al., 2026) is an 8B-scale, vision-first interaction model that observes a continuing video stream. At each second it learns to speak, remain silent, or delegate a difficult request to a background model.
- **AURA-8B**[25] (Lu et al., 2026) is a Qwen3-VL-8B-based model for always-on video understanding and proactive assistance. The released system processes a continuous 2-FPS stream and combines the VLM with pluggable ASR and streaming TTS.
- **StreamingVLM**[26] (Xu et al., 2026b) is an 8B-class model fine-tuned from Qwen2.5-VL for effectively unbounded video streams. It bounds inference cost by retaining attention sinks, recent visual tokens, and a longer window of text tokens in a compact KV cache.
- **Flash-VStream-Qwen-7B**[27] (Zhang et al., 2025a) is a Qwen-based video-language checkpoint for real-time understanding and question answering over very long streams. Its Flash Memory mechanism compresses historical visual information while new frames arrive.
- **VITA-1.5**[28] (Fu et al., 2026) is an 8B video-language model developed for real-time vision and speech interaction. The public release includes the main checkpoint together with audio-encoder and speech-generation components.

Omni-modal models. These checkpoints expose native audio-visual interfaces or otherwise combine language, vision, and speech in a single model family.

- **Qwen2.5-Omni-7B**[29] (Xu et al., 2025a) is an end-to-end omni-modal model that accepts text, images, audio, and video and can generate text or natural speech. Its time-aligned multimodal positional encoding synchronizes video and audio streams.

[18] https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf
[19] https://huggingface.co/Qwen/Qwen3.5-4B
[20] https://huggingface.co/Qwen/Qwen3.5-2B
[21] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct
[22] https://huggingface.co/Qwen/Qwen3.5-0.8B
[23] https://huggingface.co/zai-org/GLM-5.3-Flash
[24] https://huggingface.co/jdopensource/JoyAI-VL-Interaction-Preview
[25] https://huggingface.co/aurateam/AURA
[26] https://huggingface.co/mit-han-lab/StreamingVLM
[27] https://huggingface.co/zhang9302002/Flash-VStream-Qwen-7b
[28] https://huggingface.co/VITA-MLLM/VITA-1.5
[29] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

- **Qwen3-Omni-30B-A3B-Instruct**[30] (Xu et al., 2025b) combines a 30B-total, 3B-active MoE Thinker with a 3B-total, 0.3B-active MoE Talker. It accepts text, images, audio, and video and produces text and streaming natural-speech outputs.
- **OmniVinci**[31] (Ye et al., 2026) is NVIDIA's omni-modal model for joint visual, audio, and language understanding. Official inference code accepts images, standalone audio, and video with its audio track and decodes textual responses.
- **Qwen2.5-Omni-3B**[32] (Xu et al., 2025a) is the compact Qwen2.5-Omni checkpoint. It retains text, image, audio, and video perception together with text and speech generation.
- **Phi-4-Multimodal-Instruct**[33] (Abouelenin et al., 2025) is Microsoft's 5.6B model for text, image, and audio input with text output. Its documentation treats video as a sequence of images rather than a native modality.

Embodied foundation models. This group specializes in physical-world perception, spatial grounding, planning, affordance reasoning, or robot-oriented interaction.

- **HY-Embodied-VLM-1.0**[34] (Wang et al., 2026) is Tencent's approximately 30B-total MoE for physical-world agents, with roughly 3B parameters activated per token. It combines a HY3-A3B language backbone with the HY-ViT2 vision encoder for embodied perception and action-centric reasoning.
- **RynnBrain1.1-9B**[35] (Li et al., 2026b) is the dense 9B member of the RynnBrain 1.1 family for embodied perception, localization, reasoning, and planning. This generation adds native 3D-box grounding and instruction-conditioned contact-point prediction.
- **RynnBrain-30B-A3B**[36] (Dang et al., 2026) is a Qwen3-VL-based embodied MoE whose model designation denotes a 30B-class backbone with approximately 3B active parameters. It targets egocentric understanding, spatial-temporal localization, physically grounded reasoning, and planning.
- **VeBrain**[37] (Luo et al., 2025) is the 8B checkpoint associated with the Visual Embodied Brain[38] framework. The framework unifies multimodal understanding, visual-spatial reasoning, and text-form control signals for downstream robot policies.
- **Pelican1.0-VL-7B**[39] (Zhang et al., 2025b) is a 7B embodied multimodal foundation model trained for visual-text reasoning and spatiotemporal understanding. Its stated use cases include grasping, navigation, action planning, and human-robot interaction.
- **MiMo-Embodied-7B**[40] (Hao et al., 2025) is Xiaomi's cross-embodied vision-language model spanning embodied AI and autonomous-driving tasks. Its released checkpoint targets task planning, affordance prediction, spatial understanding, environmental perception, state prediction, and driving planning.
- **RynnBrain1.1-2B**[41] (Li et al., 2026b) is the compact 2B RynnBrain 1.1 checkpoint. It retains the family's 2D and 3D grounding, contact-point prediction, embodied reasoning, and planning tasks.
- **RoboBrain2.0-7B**[42] (Team et al., 2025) is BAAI's 7B embodied VLM that unifies perception, reasoning, and planning. It supports spatial outputs such as points, boxes, affordances, and trajectories together with temporal decision tasks.

[30] https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct
[31] https://huggingface.co/nvidia/omnivinci
[32] https://huggingface.co/Qwen/Qwen2.5-Omni-3B
[33] https://huggingface.co/microsoft/Phi-4-multimodal-instruct
[34] https://huggingface.co/tencent/Hy-Embodied-VLM-1.0
[35] https://huggingface.co/Alibaba-DAMO-Academy/RynnBrain1.1-9B
[36] https://huggingface.co/Alibaba-DAMO-Academy/RynnBrain-30B-A3B
[37] https://huggingface.co/OpenGVLab/VeBrain
[38] https://huggingface.co/papers/2506.00123
[39] https://huggingface.co/X-Humanoid/Pelican1.0-VL-7B
[40] https://huggingface.co/XiaomiMiMo/MiMo-Embodied-7B
[41] https://huggingface.co/Alibaba-DAMO-Academy/RynnBrain1.1-2B
[42] https://huggingface.co/BAAI/RoboBrain2.0-7B

- **RoboBrain2.0-32B**[43] (Team et al., 2025) is the full-scale 32B RoboBrain 2.0 checkpoint. It uses the same embodied interface as the 7B model while allocating greater capacity to spatial reasoning, scene memory, and long-horizon planning.
- **Embodied-R1-3B-v1**[44] (Yuan et al., 2026) is a 3B VLM for general robotic manipulation. It combines a representation based on visual pointing with reinforced fine-tuning to connect perception, grounding, and action.
- **HY-Embodied-0.5**[45] (Team et al., 2026b) is Tencent's embodied foundation-model suite for spatial-temporal perception and prediction, interaction, and planning. The released MoT-2B checkpoint contains about 4B total parameters while activating approximately 2.2B at inference.
- **HY-Embodied-0.5-X**[46] (Tencent Robotics X & HY Vision Team, 2026) is an embodied-focused enhancement of the HY-Embodied-0.5 MoT-2B architecture. It emphasizes fine-grained manipulation, spatial grounding, risk assessment, failure reflection, and long-horizon task planning.

[43] https://huggingface.co/BAAI/RoboBrain2.0-32B
[44] https://huggingface.co/IffYuan/Embodied-R1-3B-v1
[45] https://huggingface.co/tencent/HY-Embodied-0.5
[46] https://huggingface.co/tencent/HY-Embodied-0.5-X

A.4.2 COMPLETE QUANTITATIVE RESULTS

Tables 7-10 report the stage-wise diagnostics for all 45 evaluated models. Models are grouped as in the main-paper table and ordered by global Overall within each group; this shared order keeps rows aligned across tables. Bold marks the best value in each column, including ties; all metrics are higher-is-better except FCR.

Table 7: Complete Attention results.

Model	All	Nonverbal state	Semantic target	Spatial localization	Joint ID-localization	Multimodal attribution
Closed-source models						
Gemini-3.7-Flash	**0.936**	0.901	**0.951**	**0.911**	**0.985**	0.873
Gemini-3.1-Pro	0.851	0.859	0.897	0.718	0.921	**0.912**
GPT-5.6-Sol	0.906	**0.926**	0.925	0.870	0.899	0.859
Doubao-seed-2.0-lite	0.759	0.707	0.859	0.742	0.742	0.799
Open-source general-purpose models						
Qwen3.6-35B-A3B	0.723	0.714	0.891	0.560	0.732	0.875
GLM-5.3-Flash	0.801	0.710	0.882	0.834	0.770	0.824
Gemma-4-31B-IT	0.733	0.781	0.899	0.468	0.771	0.844
DeepSeek-V4.1-Flash	0.707	0.718	0.829	0.675	0.580	0.817
Qwen3.6-27B	0.649	0.709	0.836	0.517	0.539	0.873
Qwen3.5-35B-A3B	0.608	0.573	0.829	0.457	0.552	0.815
Qwen3-VL-32B-Instruct	0.682	0.667	0.886	0.584	0.611	0.829
Qwen3.5-27B	0.668	0.692	0.827	0.511	0.675	0.703
Qwen3.8-27B	0.685	0.705	0.835	0.498	0.734	0.659
Qwen2.5-VL-7B-Instruct	0.507	0.610	0.769	0.296	0.369	0.753
InternVL3-8B	0.583	0.539	0.778	0.426	0.600	0.803
InternVL3.5-30B-A3B	0.598	0.455	0.727	0.575	0.553	0.850
InternVL3.5-8B	0.574	0.524	0.768	0.557	0.422	0.766
Qwen2.5-VL-3B-Instruct	0.495	0.484	0.758	0.254	0.490	0.822
LLaVA-OneVision-7B	0.500	0.524	0.757	0.257	0.436	0.706
Qwen3.5-4B	0.534	0.551	0.715	0.398	0.467	0.631
Qwen3.5-2B	0.477	0.539	0.728	0.343	0.375	0.606
Llama-3.2-11B-Vision-Instruct	0.391	0.394	0.620	0.136	0.387	0.714
Qwen3.5-0.8B	0.381	0.421	0.537	0.292	0.265	0.551
Open-source embodied models						
HY-Embodied-VLM-1.0	0.596	0.577	0.834	0.532	0.453	0.806
RynnBrain1.1-9B	0.593	0.646	0.819	0.400	0.539	0.619
RynnBrain-30B-A3B	0.656	0.659	0.777	0.592	0.601	0.717
VeBrain-7B	0.546	0.507	0.801	0.360	0.519	0.782
Pelican1.0-VL-7B	0.540	0.581	0.724	0.298	0.551	0.782
MiMo-Embodied-7B	0.533	0.577	0.761	0.337	0.446	0.812
RynnBrain1.1-2B	0.602	0.529	0.824	0.335	0.791	0.574
RoboBrain2.0-7B	0.511	0.545	0.809	0.307	0.391	0.829
RoboBrain2.0-32B	0.486	0.566	0.781	0.271	0.358	0.731
Embodied-R1-3B-V1	0.475	0.423	0.703	0.241	0.509	0.830
HY-Embodied-0.5	0.482	0.354	0.694	0.311	0.618	0.462
HY-Embodied-0.5-X	0.461	0.305	0.677	0.257	0.629	0.471
Open-source streaming video models						
JoyAI-VL-Interaction-Preview	0.639	0.522	0.761	0.519	0.737	0.777
AURA-8B	0.624	0.535	0.834	0.522	0.626	0.737
StreamingVLM	0.456	0.507	0.753	0.222	0.361	0.763
Flash-VStream-Qwen-7B	0.470	0.516	0.675	0.334	0.400	0.503
VITA-1.5	0.460	0.425	0.638	0.379	0.446	0.487
Open-source omni-modal models						
Qwen2.5-Omni-7B	0.507	0.531	0.745	0.295	0.455	0.784
Qwen3-Omni-30B-A3B-Instruct	0.646	0.589	0.814	0.501	0.679	0.581
OmniVinci	0.546	0.592	0.733	0.346	0.565	0.557
Qwen2.5-Omni-3B	0.477	0.465	0.762	0.262	0.414	0.805
Phi-4-Multimodal-Instruct	0.435	0.400	0.541	0.336	0.400	0.751

Attention. Table 7 reports Attention Overall in the All column and aggregates the 19 Attention capabilities into five diagnostic families. Gemini-3.7-Flash leads Overall (0.936), semantic target recognition (0.951), spatial localization (0.911), and joint identity-localization (0.985), whereas GPT-5.6-Sol is strongest on nonverbal states (0.926) and Gemini-3.1-Pro on multimodal attribution (0.912). The breakdown also exposes a recurring gap between identifying a target and grounding it: for example, Llama-3.2-11B-Vision-Instruct scores 0.620 on semantic target recognition but 0.136 on

spatial localization, and Qwen2.5-VL-7B-Instruct scores 0.769 versus 0.296. Attention performance is therefore not captured by semantic recognition alone.

Table 8: Complete Intention results.

Model	All	Key moment	Online	Evidence	FCR↓	IAS
Closed-source models						
Gemini-3.7-Flash	**0.569**	**0.549**	**0.569**	0.617	0.421	**0.237**
Gemini-3.1-Pro	0.496	0.425	0.523	0.587	0.220	0.126
GPT-5.6-Sol	0.501	0.520	0.457	0.569	0.574	0.154
Doubao-seed-2.0-lite	0.418	0.409	0.363	0.582	0.686	0.066
Open-source general-purpose models						
Qwen3.6-35B-A3B	0.485	0.501	0.439	0.568	0.634	0.122
GLM-5.3-Flash	0.501	0.513	0.455	0.590	0.644	0.128
Gemma-4-31B-IT	0.502	0.501	0.482	0.554	0.372	0.167
DeepSeek-V4.1-Flash	0.509	0.539	0.451	0.588	0.552	0.121
Qwen3.6-27B	0.480	0.434	0.483	0.578	0.250	0.119
Qwen3.5-35B-A3B	0.448	0.369	0.474	0.565	0.441	0.126
Qwen3-VL-32B-Instruct	0.472	0.409	0.481	0.596	0.397	0.133
Qwen3.5-27B	0.382	0.384	0.336	0.490	0.645	0.063
Qwen3.8-27B	0.395	0.391	0.374	0.459	0.512	0.052
Qwen2.5-VL-7B-Instruct	0.452	0.358	0.499	0.550	0.240	0.102
InternVL3-8B	0.417	0.395	0.375	0.573	0.640	0.082
InternVL3.5-30B-A3B	0.426	0.333	0.458	0.556	0.338	0.095
InternVL3.5-8B	0.406	0.367	0.374	0.579	0.613	0.094
Qwen2.5-VL-3B-Instruct	0.409	0.390	0.357	0.586	0.699	0.069
LLaVA-OneVision-7B	0.387	0.355	0.380	0.476	0.507	0.078
Qwen3.5-4B	0.389	0.287	0.442	0.486	0.475	0.123
Qwen3.5-2B	0.375	0.351	0.356	0.479	0.499	0.037
Llama-3.2-11B-Vision-Instruct	0.384	0.269	0.422	0.549	0.250	0.011
Qwen3.5-0.8B	0.299	0.259	0.276	0.446	0.713	0.036
Open-source embodied models						
HY-Embodied-VLM-1.0	0.396	0.284	0.419	0.595	0.605	0.105
RynnBrain1.1-9B	0.417	0.321	0.468	0.509	0.327	0.123
RynnBrain-30B-A3B	0.435	0.382	0.471	0.461	0.245	0.091
VeBrain-7B	0.449	0.358	0.494	0.541	0.186	0.111
Pelican1.0-VL-7B	0.462	0.372	0.503	0.562	0.142	0.099
MiMo-Embodied-7B	0.440	0.396	0.407	**0.623**	0.640	0.114
RynnBrain1.1-2B	0.396	0.346	0.430	0.424	0.333	0.073
RoboBrain2.0-7B	0.439	0.368	0.447	0.581	0.254	0.054
RoboBrain2.0-32B	0.451	0.423	0.444	0.534	0.199	0.049
Embodied-R1-3B-V1	0.408	0.368	0.376	0.580	0.618	0.073
HY-Embodied-0.5	0.368	0.315	0.388	0.436	0.328	0.022
HY-Embodied-0.5-X	0.373	0.319	0.392	0.443	0.247	0.008
Open-source streaming video models						
JoyAI-VL-Interaction-Preview	0.436	0.407	0.408	0.570	0.529	0.113
AURA-8B	0.472	0.411	0.490	0.563	0.242	0.113
StreamingVLM	0.400	0.258	0.449	0.598	0.149	0.029
Flash-VStream-Qwen-7B	0.396	0.384	0.398	0.415	0.518	0.065
VITA-1.5	0.388	0.324	0.435	0.412	0.320	0.091
Open-source omni-modal models						
Qwen2.5-Omni-7B	0.453	0.378	0.485	0.543	0.137	0.070
Qwen3-Omni-30B-A3B-Instruct	0.419	0.323	0.498	0.439	0.291	0.153
OmniVinci	0.401	0.388	0.410	0.409	0.474	0.116
Qwen2.5-Omni-3B	0.433	0.335	0.466	0.574	0.220	0.065
Phi-4-Multimodal-Instruct	0.398	0.285	0.426	0.587	**0.083**	0.000

Intention. In Table 8, All denotes Intention Overall, FCR is the false-commitment rate (lower is better), and IAS measures timely, stable intent recovery after penalizing premature commitments. Gemini-3.7-Flash obtains the highest Overall score (0.569) and IAS (0.237), but still has an FCR of 0.421. The converse failure is visible for Phi-4-Multimodal-Instruct: it has the lowest FCR (0.083) but an IAS of 0.000. Conservative prediction is thus insufficient by itself; a model must also recover the correct intent early and maintain it.

Table 9: Complete Reaction results with action-domain diagnostics.

Model	Core Reaction metrics					Subaction + Speak by action domain			
	All	Engage	Action domain	Subaction + Speak	Joint EM	Comm.	Nav.	Manip.	Refrain/ Idle
Closed-source models									
Gemini-3.7-Flash	0.638	0.546	0.831	0.486	0.083	0.477	0.417	0.484	0.414
Gemini-3.1-Pro	0.620	0.530	0.838	0.487	0.111	0.487	0.498	0.494	0.443
GPT-5.6-Sol	0.489	0.261	0.829	0.426	0.128	0.438	0.484	0.446	0.466
Doubao-seed-2.0-lite	0.619	0.564	0.792	0.435	0.092	0.483	0.526	0.504	0.573
Open-source general-purpose models									
Qwen3.6-35B-A3B	0.655	0.605	0.818	0.446	0.064	0.466	0.463	0.451	0.530
GLM-5.3-Flash	0.590	0.531	0.810	0.379	0.067	0.403	0.463	0.428	0.384
Gemma-4-31B-IT	0.535	0.394	0.810	0.471	0.123	0.452	0.416	0.390	0.370
DeepSeek-V4.1-Flash	0.548	0.456	0.815	0.401	0.108	0.420	0.470	0.391	0.313
Qwen3.6-27B	0.545	0.383	0.824	0.453	0.128	0.488	0.457	0.472	0.448
Qwen3.5-35B-A3B	0.609	0.563	**0.848**	0.428	0.100	0.421	0.445	0.384	0.447
Qwen3-VL-32B-Instruct	0.479	0.217	0.826	0.425	0.144	0.452	0.489	0.452	0.457
Qwen3.5-27B	0.539	0.446	0.684	0.481	0.123	0.504	0.506	0.474	0.513
Qwen3.8-27B	0.475	0.251	0.651	0.491	0.137	0.541	0.605	0.597	0.567
Qwen2.5-VL-7B-Instruct	0.466	0.195	0.758	0.461	0.144	0.489	0.498	0.459	0.416
InternVL3-8B	0.518	0.271	0.781	0.499	0.123	0.534	0.590	0.484	0.574
InternVL3.5-30B-A3B	0.442	0.134	0.816	0.426	0.144	0.441	0.457	0.414	0.503
InternVL3.5-8B	0.458	0.309	0.752	0.394	0.119	0.388	0.315	0.352	0.406
Qwen2.5-VL-3B-Instruct	0.483	0.142	0.789	0.475	0.144	0.548	0.583	0.589	0.651
LLaVA-OneVision-7B	0.379	0.132	0.704	0.338	0.144	0.350	0.328	0.340	0.438
Qwen3.5-4B	0.481	0.453	0.605	0.413	0.129	0.412	0.395	0.343	0.365
Qwen3.5-2B	0.530	0.489	0.540	0.385	0.036	0.436	0.446	0.454	0.433
Llama-3.2-11B-Vision-Instruct	0.463	0.371	0.638	0.364	0.128	0.426	0.407	0.492	0.519
Qwen3.5-0.8B	0.506	0.517	0.460	0.343	0.044	0.427	0.482	0.493	0.489
Open-source embodied models									
HY-Embodied-VLM-1.0	**0.688**	0.485	0.749	0.522	0.028	0.558	0.575	0.502	0.549
RynnBrain1.1-9B	0.557	0.605	0.760	0.315	0.082	0.300	0.345	0.249	0.244
RynnBrain-30B-A3B	0.410	0.245	0.789	0.329	**0.147**	0.325	0.260	0.222	0.241
VeBrain-7B	0.470	0.154	0.800	0.440	0.141	0.501	0.550	0.517	0.547
Pelican1.0-VL-7B	0.440	0.128	0.734	0.441	0.144	0.500	0.546	0.535	0.478
MiMo-Embodied-7B	0.643	0.558	0.785	0.463	0.056	0.512	0.550	0.536	0.622
RynnBrain1.1-2B	0.467	0.426	0.719	0.334	0.090	0.299	0.257	0.235	0.349
RoboBrain2.0-7B	0.471	0.184	0.782	0.448	0.146	0.488	0.420	0.527	0.516
RoboBrain2.0-32B	0.441	0.229	0.709	0.407	0.136	0.427	0.354	0.403	0.462
Embodied-R1-3B-V1	0.468	0.126	0.784	0.451	0.144	0.518	0.552	0.576	**0.655**
HY-Embodied-0.5	0.563	**0.606**	0.692	0.295	0.065	0.343	0.356	0.308	0.359
HY-Embodied-0.5-X	0.480	0.515	0.671	0.290	0.126	0.333	0.337	0.315	0.363
Open-source streaming video models									
JoyAI-VL-Interaction-Preview	0.610	0.508	0.752	0.421	0.029	0.434	0.455	0.372	0.529
AURA-8B	0.434	0.146	0.812	0.391	0.144	0.389	0.414	0.331	0.404
StreamingVLM	0.536	0.302	0.733	**0.534**	0.141	**0.600**	**0.609**	**0.650**	0.622
Flash-VStream-Qwen-7B	0.376	0.131	0.711	0.341	0.142	0.366	0.356	0.298	0.383
VITA-1.5	0.420	0.344	0.655	0.342	0.103	0.313	0.300	0.251	0.330
Open-source omni-modal models									
Qwen2.5-Omni-7B	0.449	0.126	0.796	0.411	0.144	0.453	0.439	0.479	0.485
Qwen3-Omni-30B-A3B-Instruct	0.390	0.138	0.757	0.343	0.141	0.356	0.369	0.310	0.338
OmniVinci	0.340	0.126	0.713	0.256	0.144	0.265	0.286	0.221	0.263
Qwen2.5-Omni-3B	0.467	0.126	0.769	0.465	0.144	0.524	0.559	0.523	0.594
Phi-4-Multimodal-Instruct	0.435	0.133	0.648	0.435	0.137	0.524	0.502	0.599	0.647

Reaction. Table 9 reports the Overall Reaction performance across all events. For individual metrics, Engage uses an active-versus-non-intervention Macro-F1, action domain and subaction use multi-label F1, and the Speak subtype uses accuracy. For Communication events with a Speak annotation, the domain score averages the subaction and Speak-subtype scores; note that the four domain columns are non-exclusive. HY-Embodied-VLM-1.0 achieves the highest Overall score (0.688) but yields only 0.028 in Joint-EM, whereas RynnBrain-30B-A3B attains the highest Joint-EM (0.147) despite a lower Overall score of 0.410. StreamingVLM exhibits a complementary profile: it leads in Subaction + Speak (0.534) and the Communication, Navigation, and Manipulation breakdowns (0.600, 0.609, and 0.650), yet its Engage score remains low at 0.302. This indicates that strong conditional action specification does not guarantee correct engagement decisions or the successful composition of full reaction chains.

Analysis of the Joint-EM scores explains why eleven otherwise heterogeneous models report an identical value of 0.144. Joint-EM measures the fraction of the 611 complete Reaction events where the entire decision chain is predicted correctly. The ground truth comprises 88 *refrain* (non-intervention) events, which require only a correct engagement decision, and 523 *active* events, which strictly require simultaneous matches across the action domain, subaction, and any annotated speech subtype. The eleven tied models (Qwen3-VL-32B-Instruct, Qwen2.5-VL-7B/3B-Instruct, InternVL3.5-30B-A3B, LLaVA-OneVision-7B, Pelican1.0-VL-7B, Embodied-R1-3B-V1, AURA-8B, Qwen2.5-Omni-7B/3B, and OmniVinci) exhibit policy collapse, outputting a near-constant "refrain" prediction (four models unconditionally predict refrain, while the remaining seven do so on at least 92% of queries). This degenerate policy trivially solves all 88 refrain events but fails all active events at the initial engagement step, fixing the number of fully correct chains at exactly 88 and yielding a Joint-EM of $88/611 \approx 0.144$. Rather than a metric artifact, this phenomenon highlights the exact degenerate behavior the strict joint constraint is designed to expose: a collapsed non-intervention policy paradoxically outscores models that attempt active responses (e.g., Doubao-Seed-2.0-Lite at 0.092, Gemini-3.7-Flash at 0.084, and HY-Embodied-VLM-1.0 at 0.028). Consequently, Joint-EM must be evaluated alongside the Engage Macro-F1 metric, which explicitly penalizes such collapse. Consistently, all eleven collapsed models report poor Engage scores, and even the highest overall Joint-EM (RynnBrain-30B-A3B, 0.147) sits only marginally above this non-intervention baseline.

Table 10: Complete Interaction results.

Model	All	State	Task outcome	Compliance	Violation
Closed-source models					
Gemini-3.7-Flash	**0.760**	**0.441**	0.885	0.903	0.525
Gemini-3.1-Pro	**0.760**	0.368	0.903	0.902	0.495
GPT-5.6-Sol	0.729	0.400	0.880	0.906	0.494
Doubao-seed-2.0-lite	0.669	0.345	0.870	0.726	0.572
Open-source general-purpose models					
Qwen3.6-35B-A3B	0.738	0.293	0.927	0.881	0.577
GLM-5.3-Flash	0.638	0.386	0.823	0.706	0.470
Gemma-4-31B-IT	0.728	0.281	0.812	0.907	0.495
DeepSeek-V4.1-Flash	0.657	0.339	0.828	0.737	0.508
Qwen3.6-27B	0.733	0.235	0.853	0.905	0.501
Qwen3.5-35B-A3B	0.715	0.258	0.900	0.877	0.461
Qwen3-VL-32B-Instruct	0.723	0.299	0.690	**0.923**	**0.589**
Qwen3.5-27B	0.656	0.264	0.870	0.768	0.478
Qwen3.8-27B	0.632	0.339	0.860	0.717	0.341
Qwen2.5-VL-7B-Instruct	0.713	0.310	0.870	0.863	0.483
InternVL3-8B	0.603	0.212	0.925	0.724	0.318
InternVL3.5-30B-A3B	0.639	0.226	0.730	0.754	0.511
InternVL3.5-8B	0.562	0.270	0.728	0.689	0.351
Qwen2.5-VL-3B-Instruct	0.515	0.200	0.903	0.538	0.385
LLaVA-OneVision-7B	0.602	0.232	0.887	0.657	0.453
Qwen3.5-4B	0.425	0.241	0.417	0.523	0.322
Qwen3.5-2B	0.396	0.238	0.657	0.320	0.347
Llama-3.2-11B-Vision-Instruct	0.455	0.249	0.608	0.498	0.381
Qwen3.5-0.8B	0.398	0.281	0.740	0.353	0.329
Open-source embodied models					
HY-Embodied-VLM-1.0	0.716	0.209	0.943	0.873	0.481
RynnBrain1.1-9B	0.736	0.275	0.818	0.918	0.583
RynnBrain-30B-A3B	0.674	0.252	0.927	0.723	0.508
VeBrain-7B	0.701	0.270	0.905	0.828	0.502
Pelican1.0-VL-7B	0.700	0.206	0.877	0.875	0.470
MiMo-Embodied-7B	0.487	0.322	0.215	0.669	0.466
RynnBrain1.1-2B	0.567	0.238	0.807	0.530	0.490
RoboBrain2.0-7B	0.597	0.229	0.943	0.696	0.379
RoboBrain2.0-32B	0.571	0.220	0.920	0.639	0.367
Embodied-R1-3B-V1	0.569	0.197	0.828	0.635	0.412
HY-Embodied-0.5	0.429	0.238	**0.960**	0.353	0.318
HY-Embodied-0.5-X	0.415	0.290	0.905	0.306	0.338
Open-source streaming video models					
JoyAI-VL-Interaction-Preview	0.586	0.287	0.943	0.635	0.417
AURA-8B	0.601	0.342	0.468	0.766	0.510
StreamingVLM	0.655	0.342	0.925	0.785	0.413
Flash-VStream-Qwen-7B	0.631	0.232	0.940	0.682	0.436
VITA-1.5	0.477	0.243	0.542	0.523	0.468

Continued on next page.

Table 10 continued.

Model	All	State	Task outcome	Compliance	Violation
Open-source omni-modal models					
Qwen2.5-Omni-7B	0.705	0.267	0.820	0.855	0.510
Qwen3-Omni-30B-A3B-Instruct	0.647	0.357	0.848	0.746	0.415
OmniVinci	0.651	0.316	0.785	0.740	0.454
Qwen2.5-Omni-3B	0.520	0.258	0.830	0.534	0.200
Phi-4-Multimodal-Instruct	0.356	0.217	0.917	0.172	0.301

Interaction. Table 10 reports Interaction Overall in All. Compliance equally averages spatial, physical, communication, and overall compliance, while Violation equally averages violation-presence F1, online violation type, and segment-level violation types. Gemini-3.7-Flash and Gemini-3.1-Pro tie for the best Overall score (0.760), whereas Qwen3-VL-32B-Instruct leads both Compliance (0.923) and Violation (0.589). Component scores can nevertheless diverge sharply: HY-Embodied-0.5 attains the highest Task Outcome score (0.960) but only 0.429 Overall, with 0.238 on State and 0.353 on Compliance. Accurate resolved-outcome prediction therefore does not imply reliable tracking or assessment of the full interaction.

A.5 Benchmark Visualizations and Qualitative Analysis

A.5.1 Attention cases

Figures 7-25 present representative streaming examples for all 19 Attention capabilities. To balance visual consistency with scenario diversity, adjacent capabilities are grouped within four expertly verified episodes: a professional craft workshop, a retail checkout, an outdoor farmers' market, and a card shop checkout. Each panel displays only the causal observation prefix available at the designated query time, followed by the original options and the verified gold response. For tasks requiring spatial or combined semantic and spatial reasoning, candidate bounding boxes are additionally overlaid on the final visible frame.

A.5.2 Intention cases

Figures 26-31 provide two representative examples for each of the three Intention capabilities: key-moment intention recognition, temporal intention recognition, and intention evidence grounding. The paired cases cover distinct social contexts and evidence conditions, including both supportable intentions and an early prefix for which the calibrated response is insufficient evidence. Each panel shows only the causal video prefix and the ASR available by the marked query time, together with the original answer choices and reviewed gold response.

A.5.3 Reaction cases

Figures 32-38 provide representative examples spanning the four Reaction capabilities: engagement decision, action-domain selection, subaction selection, and speech-function selection. The paired checkout and service-window scenes illustrate how AIRIS progresses from deciding whether Ego should respond to specifying the appropriate response modality, concrete action, and, when speech is warranted, communicative function. Each panel uses only the causal video prefix and any ASR available by the marked query time, and retains the original answer choices and reviewed gold response.

A.5.4 Interaction cases

Figures 39-57 provide 19 representative examples spanning the ten Interaction capabilities. The query-time cases assess the evolving interaction state and whether a social violation is present at the current prefix, whereas the segment-level cases assess completion, mutual understanding, spatial, physical, and communication compliance, overall compliance category, and the violation types observed over the episode. Together, the examples cover both compliant and violating interactions across retail-counter, reception, sales-floor, and clinical-exam settings. Each panel retains the original answer choices and reviewed gold response, and explicitly distinguishes a causal query time from post-hoc review of the supplied interaction segment.

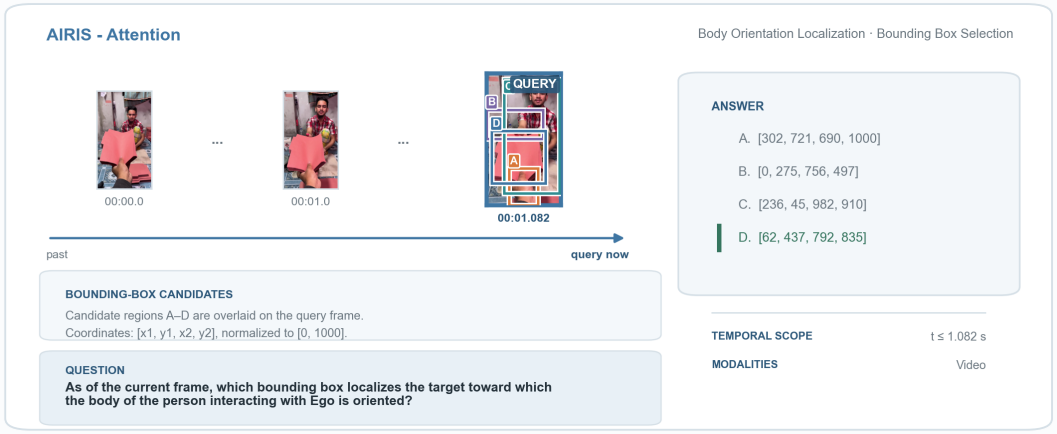


Figure 7: Representative Attention case for body-orientation target localization.

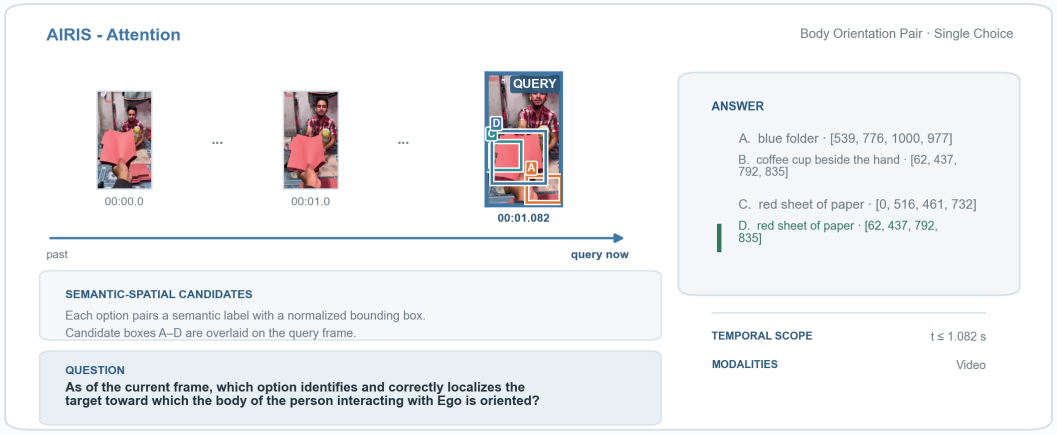


Figure 8: Representative Attention case for joint body-orientation target identification and localization.

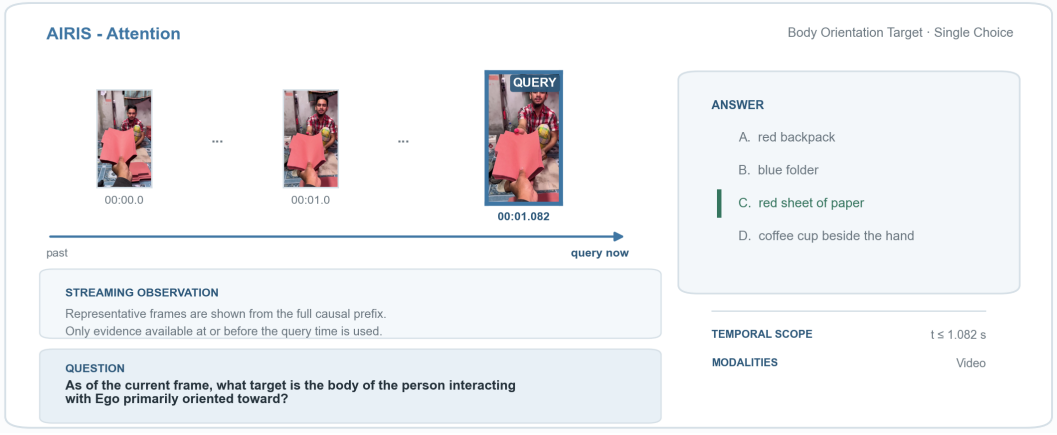


Figure 9: Representative Attention case for body-orientation target recognition.

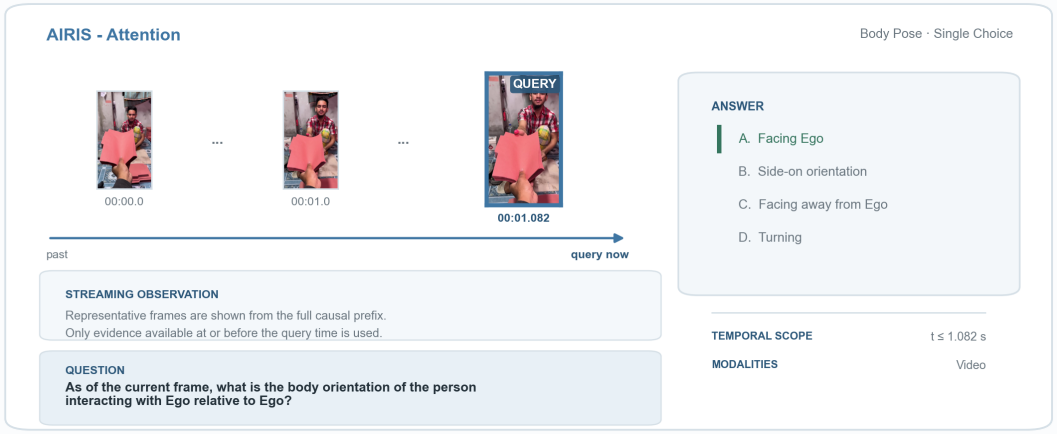


Figure 10: Representative Attention case for body-pose recognition.

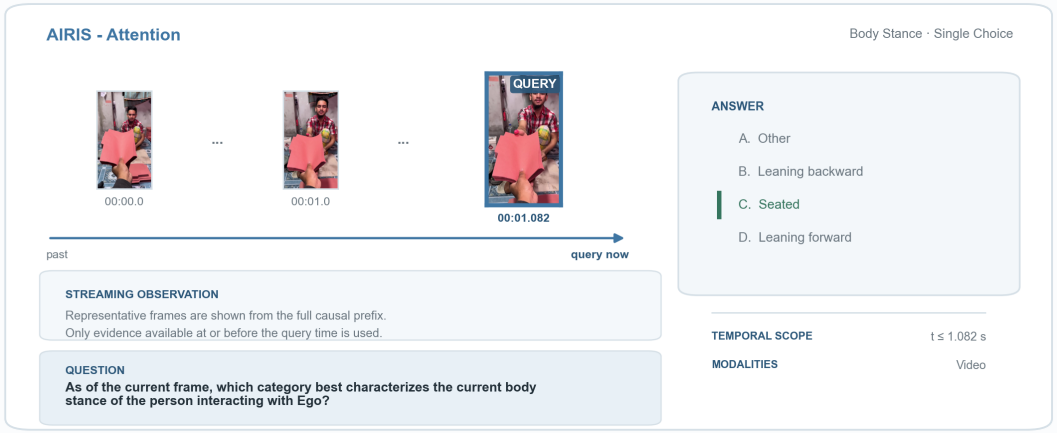


Figure 11: Representative Attention case for body-stance recognition.

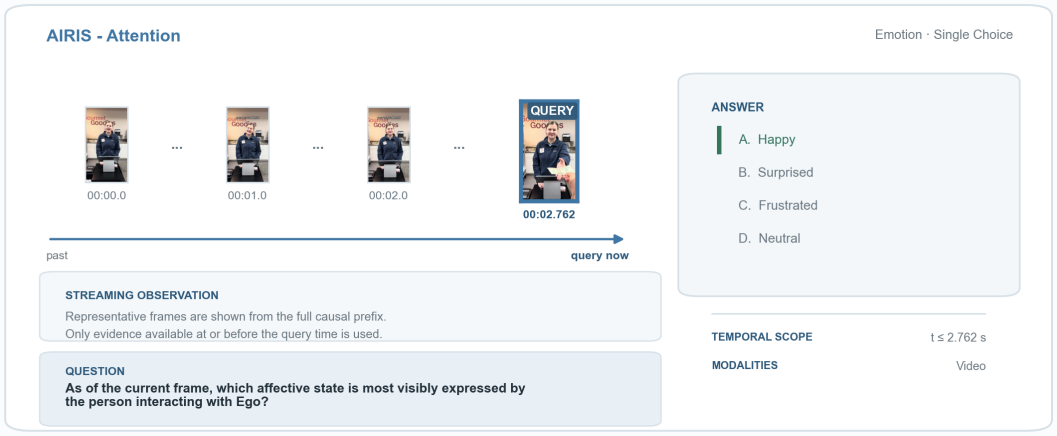


Figure 12: Representative Attention case for emotion recognition.

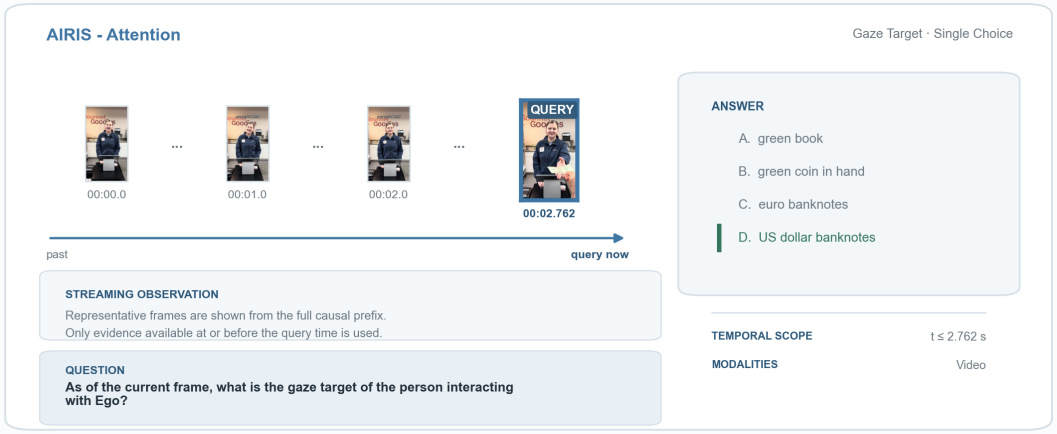


Figure 13: Representative Attention case for gaze-target recognition.

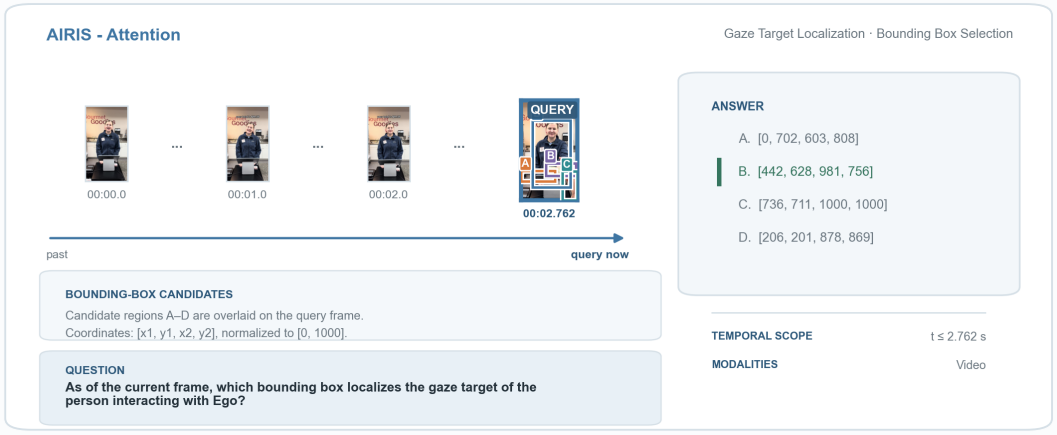


Figure 14: Representative Attention case for gaze-target localization.

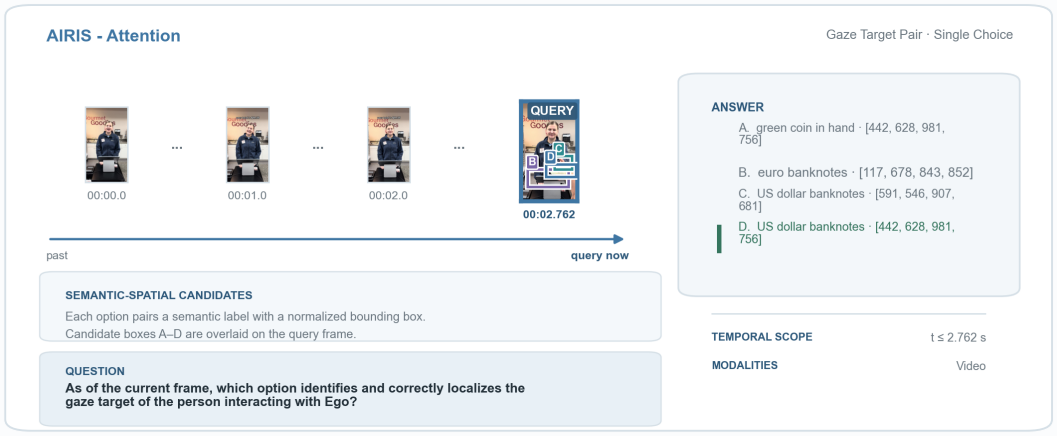


Figure 15: Representative Attention case for joint gaze-target identification and localization.

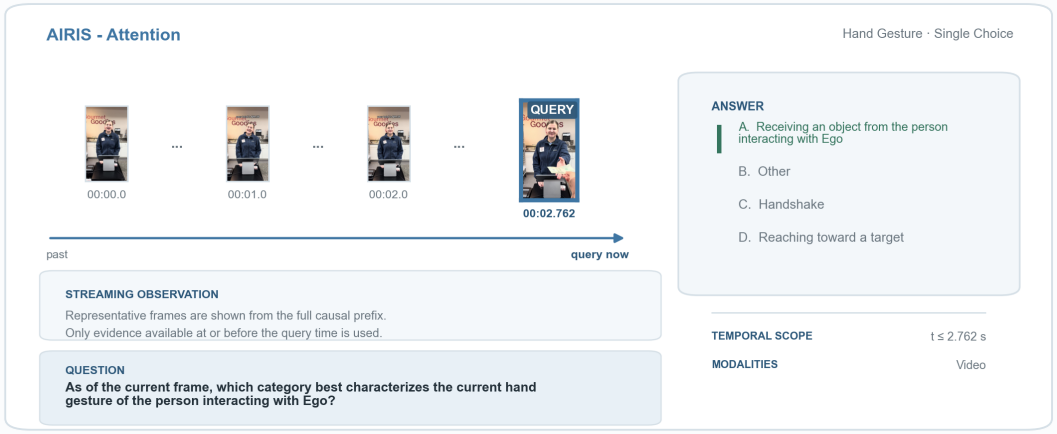


Figure 16: Representative Attention case for hand-gesture recognition.

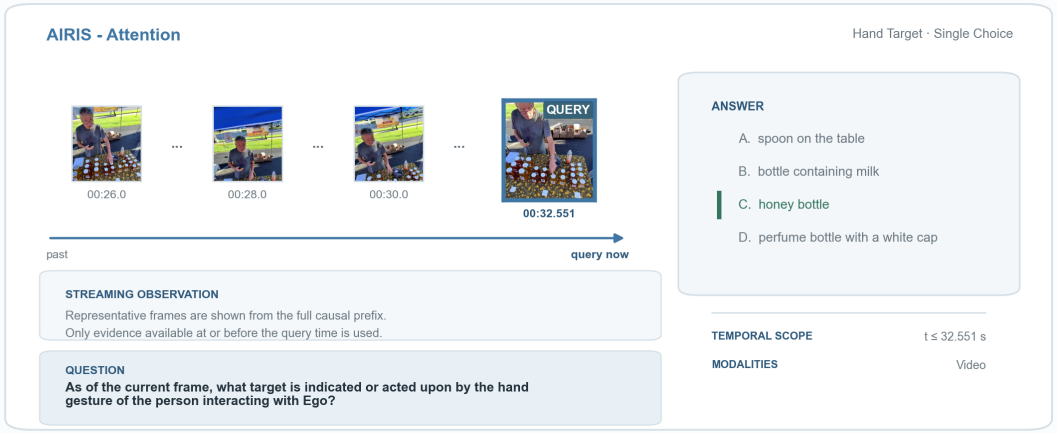


Figure 17: Representative Attention case for hand-target recognition.

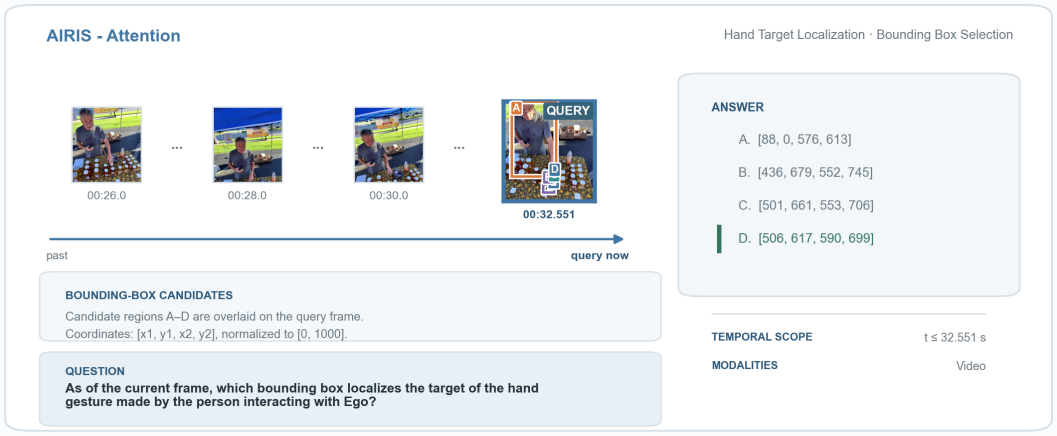


Figure 18: Representative Attention case for hand-target localization.

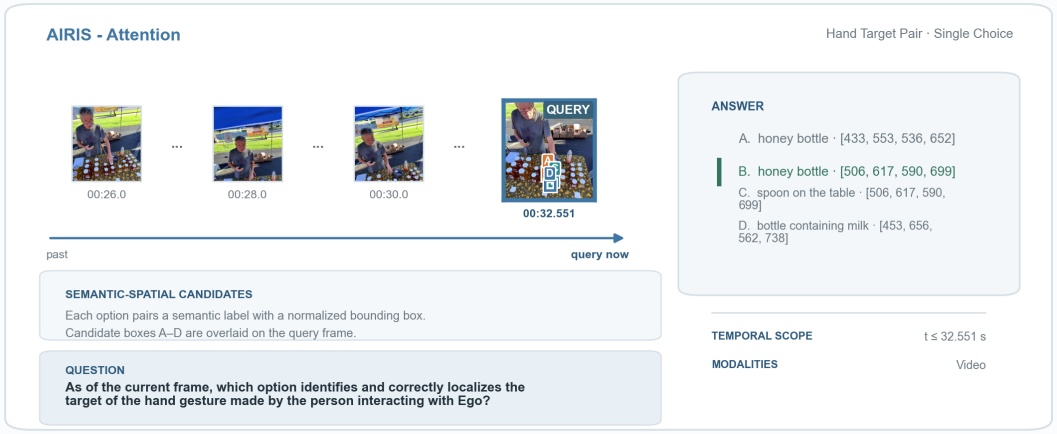


Figure 19: Representative Attention case for joint hand-target identification and localization.

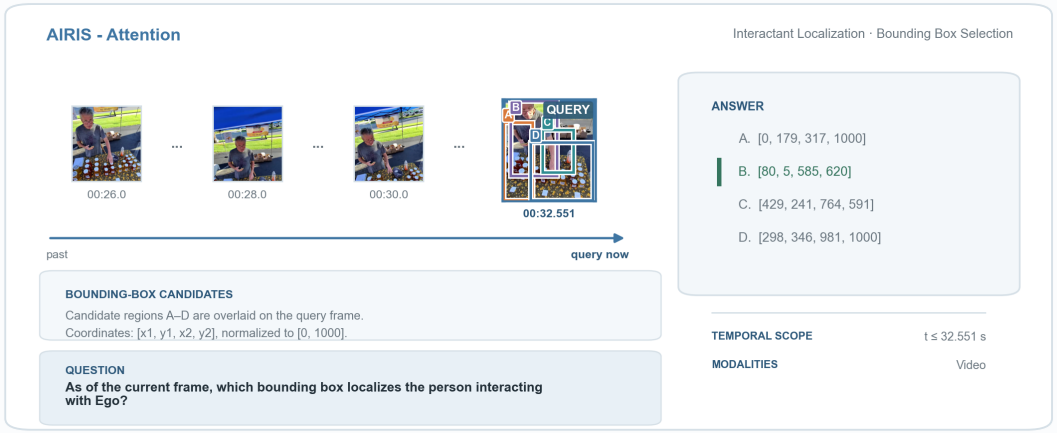


Figure 20: Representative Attention case for interactant localization.

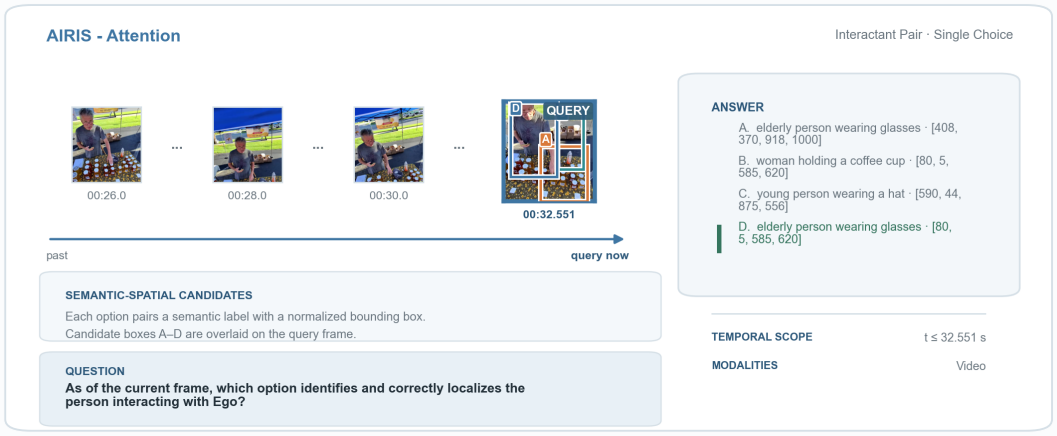


Figure 21: Representative Attention case for joint interactant identification and localization.

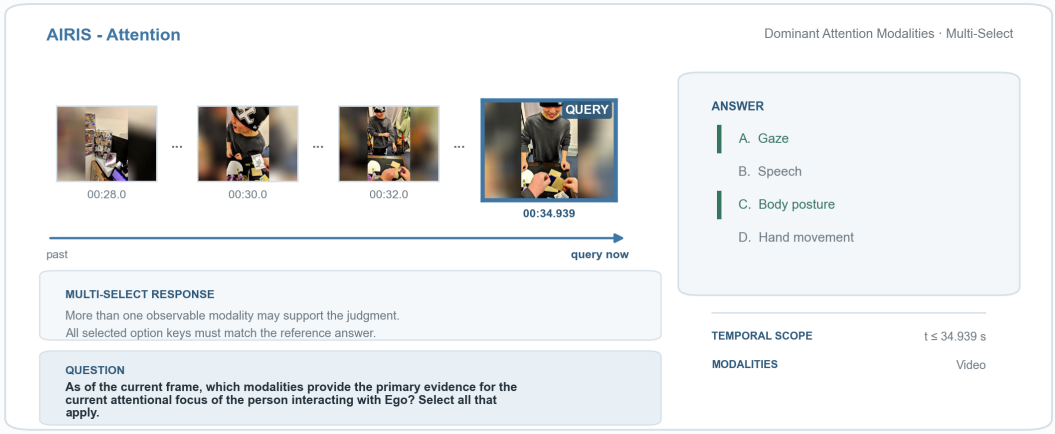


Figure 22: Representative Attention case for dominant-attention modality recognition.

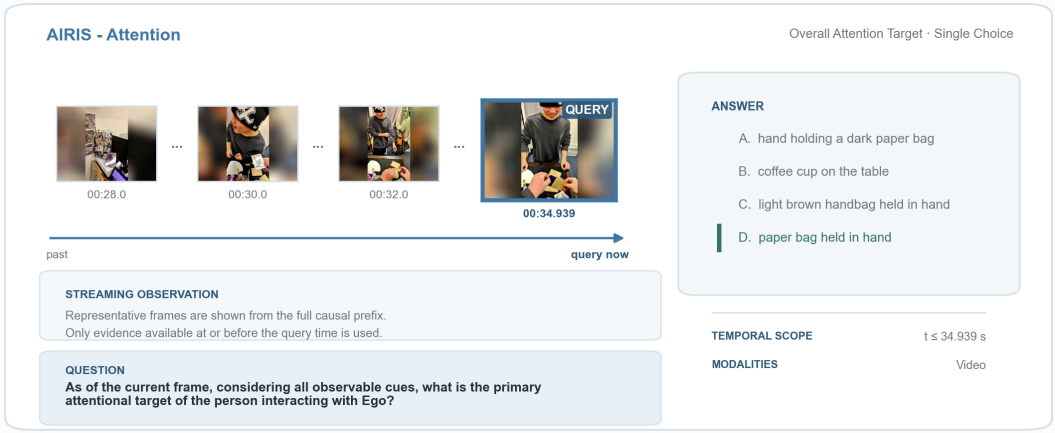


Figure 23: Representative Attention case for overall attention-target recognition.

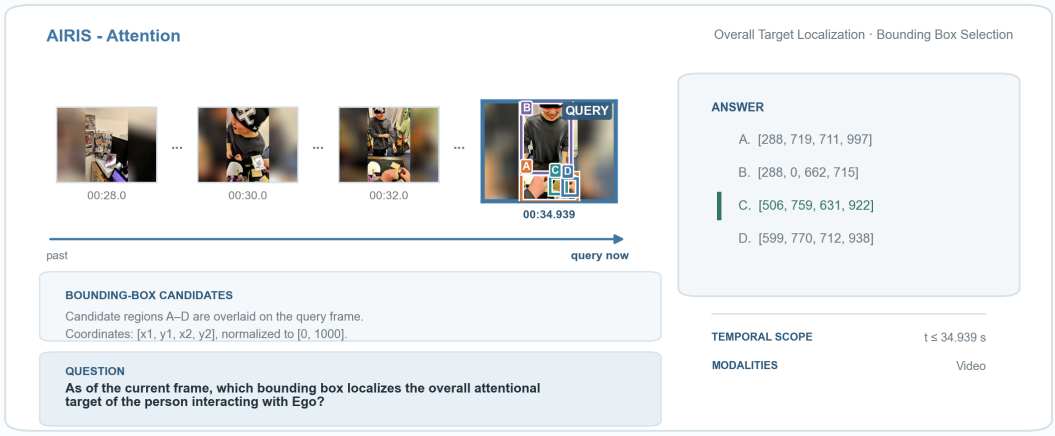


Figure 24: Representative Attention case for overall attention-target localization.

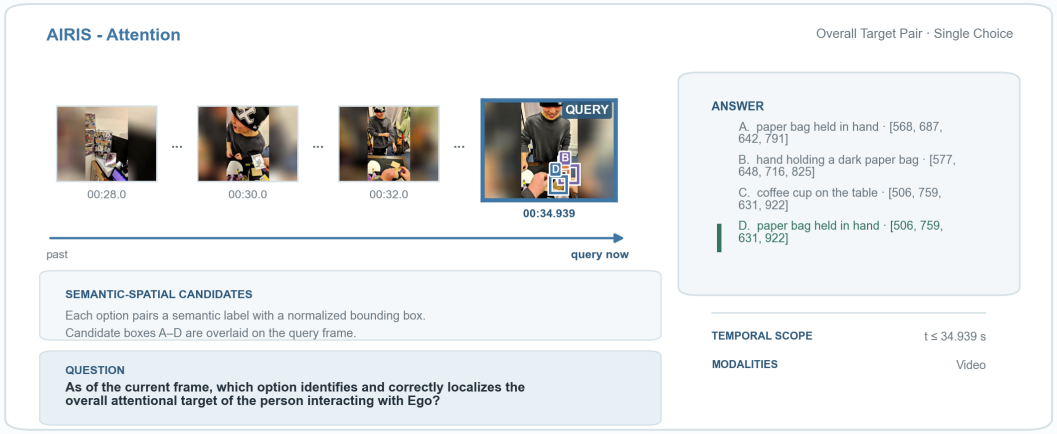


Figure 25: Representative Attention case for joint overall attention-target identification and localization.

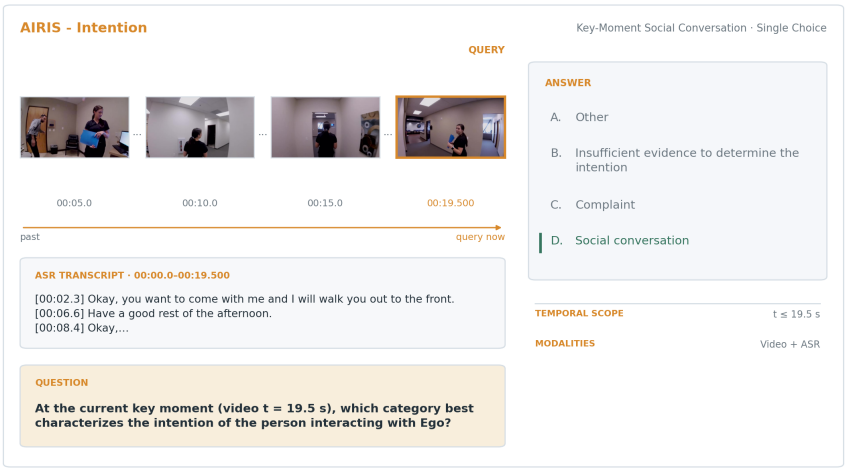


Figure 26: Representative Intention case for key-moment social-conversation recognition.

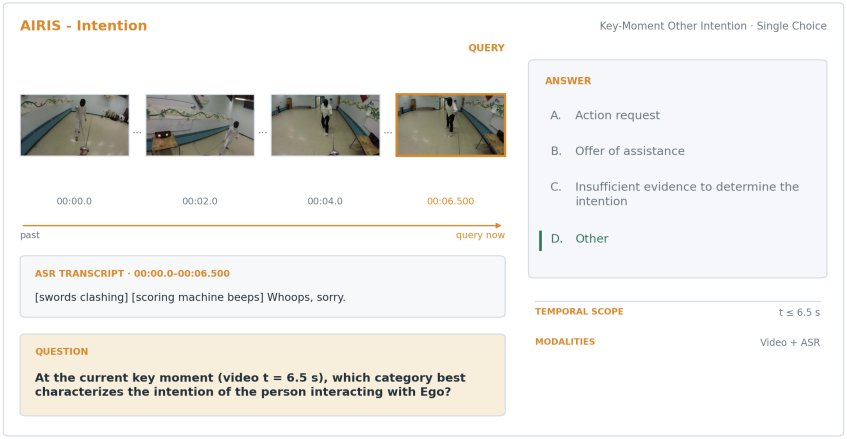


Figure 27: Representative Intention case for key-moment other-intention recognition.

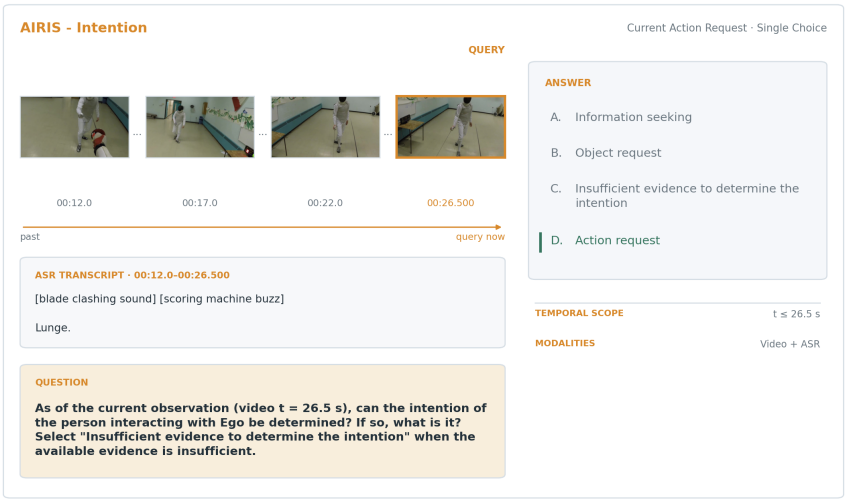


Figure 28: Representative Intention case for temporal action-request recognition.

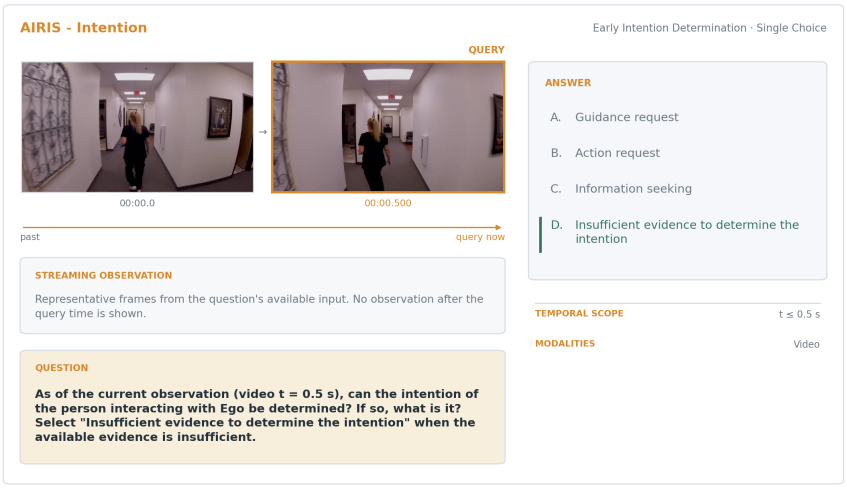


Figure 29: Representative Intention case for early intention determination under insufficient evidence.

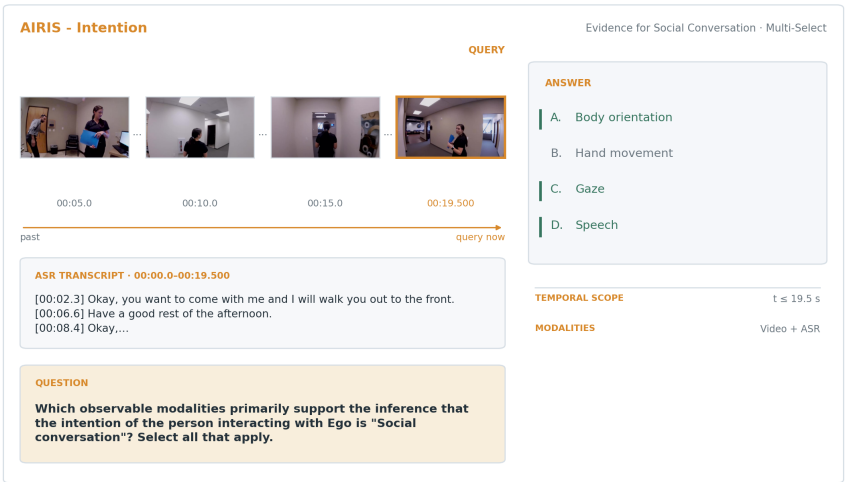


Figure 30: Representative Intention case for social-conversation evidence grounding.

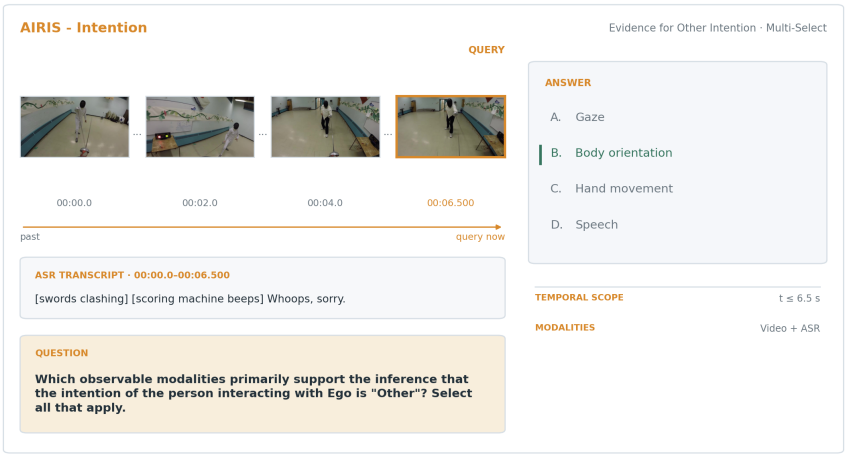


Figure 31: Representative Intention case for other-intention evidence grounding.

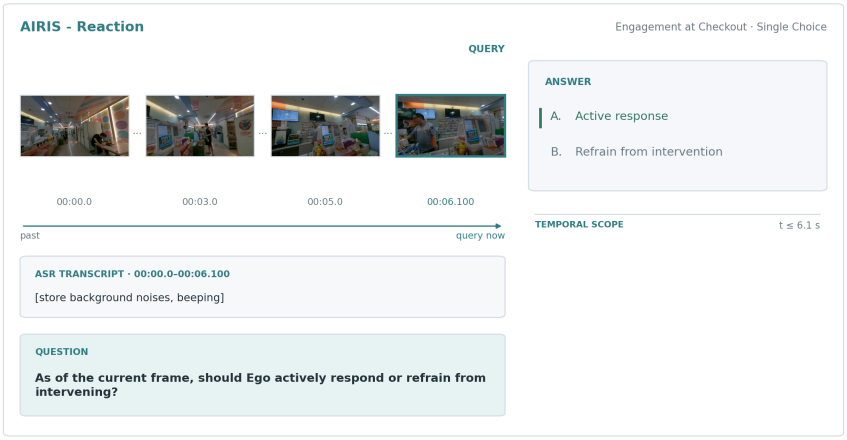


Figure 32: Representative Reaction case for engagement decision at checkout.

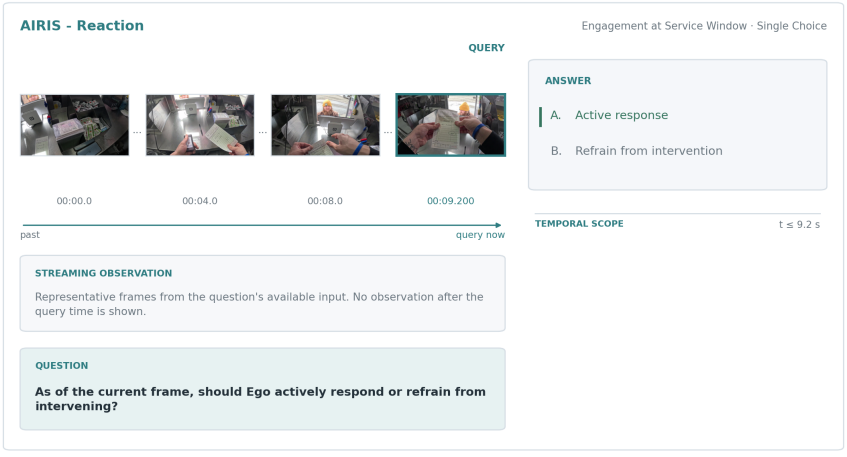


Figure 33: Representative Reaction case for engagement decision at a service window.

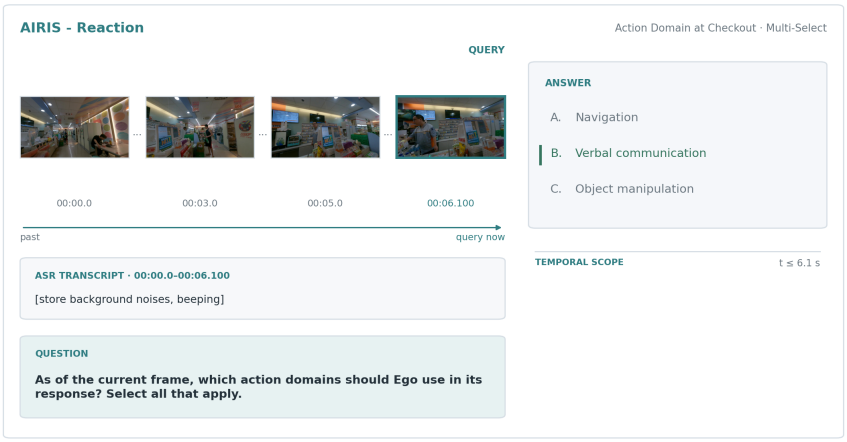


Figure 34: Representative Reaction case for action-domain selection at checkout.

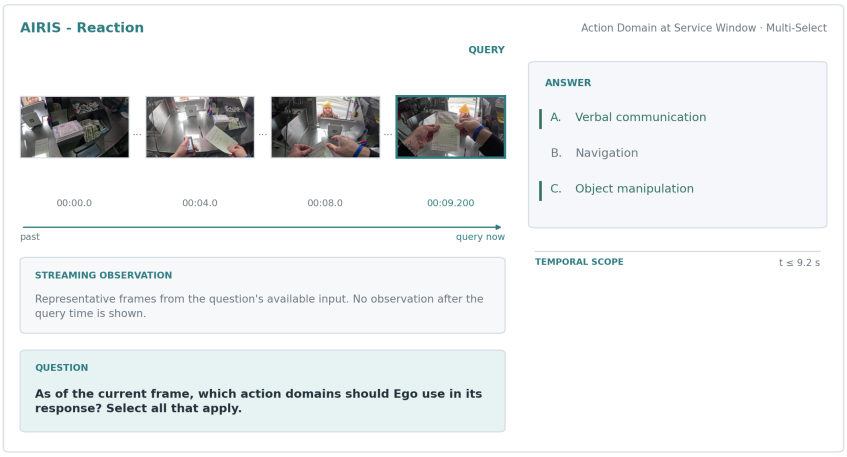


Figure 35: Representative Reaction case for action-domain selection at a service window.

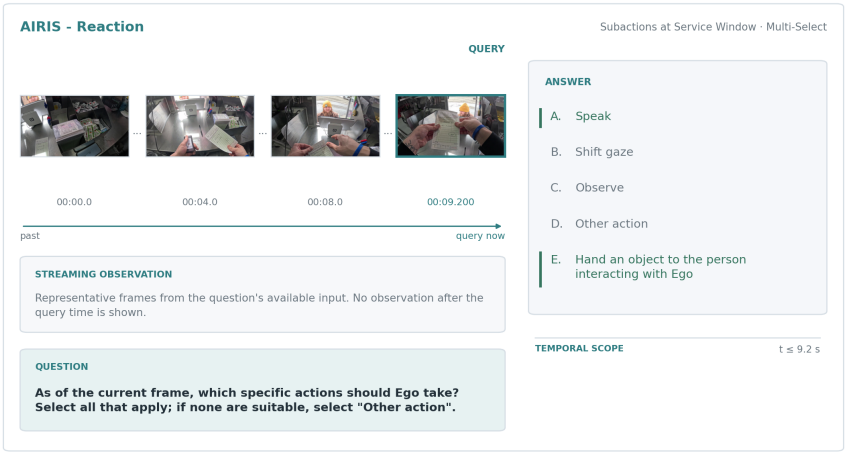


Figure 36: Representative Reaction case for subaction selection at a service window.

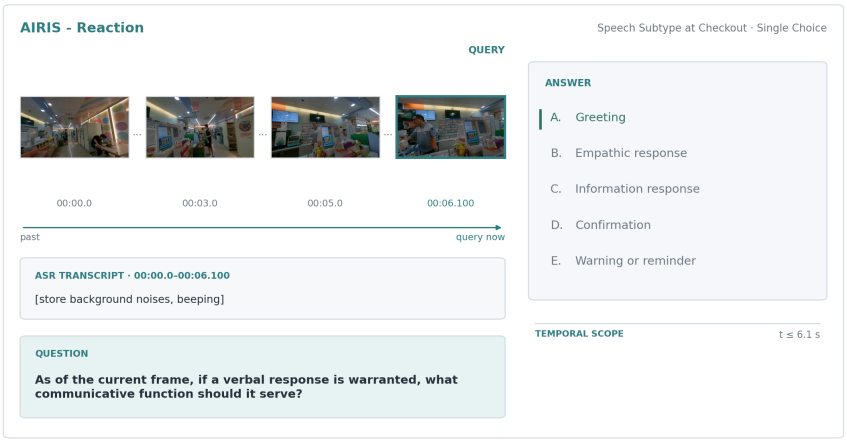


Figure 37: Representative Reaction case for speech-function selection at checkout.

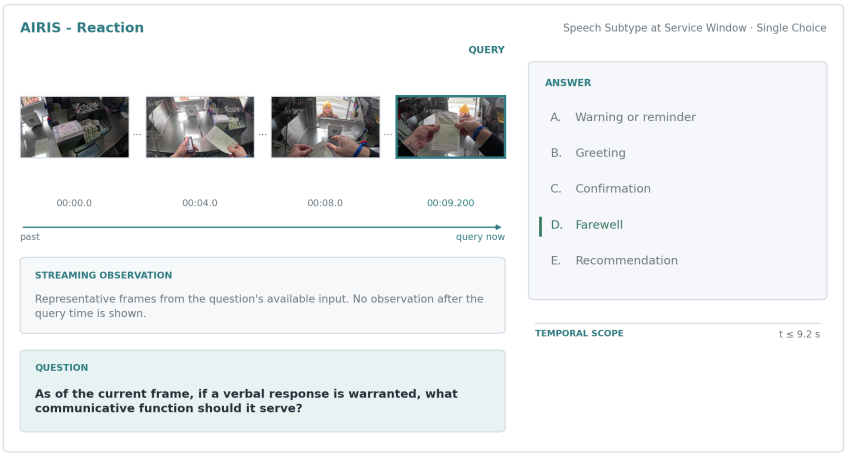


Figure 38: Representative Reaction case for speech-function selection at a service window.

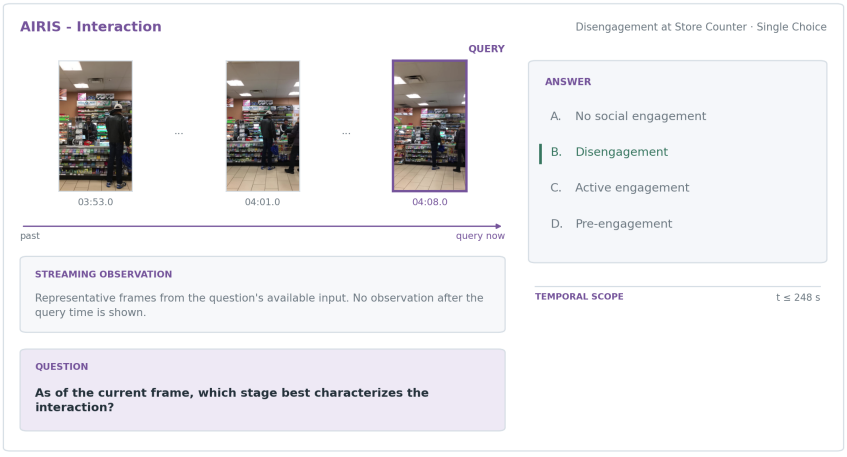


Figure 39: Representative Interaction case for disengagement-state recognition at a store counter.

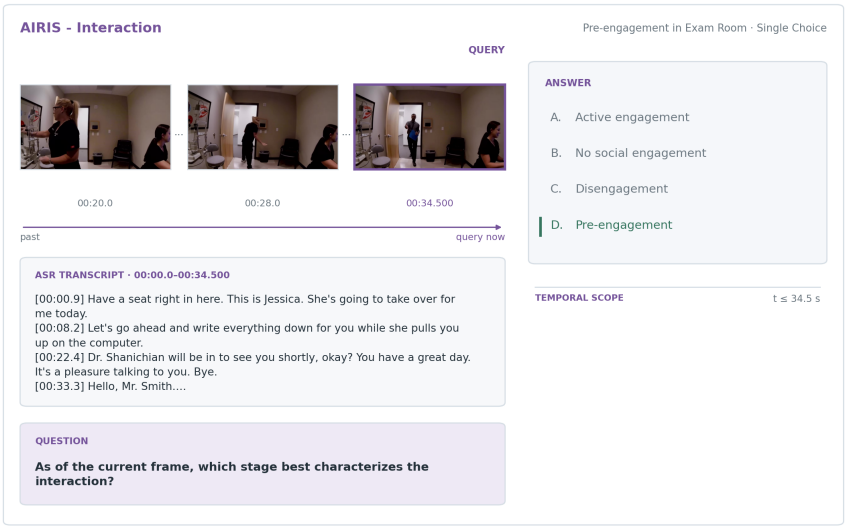


Figure 40: Representative Interaction case for pre-engagement-state recognition in a clinical exam room.

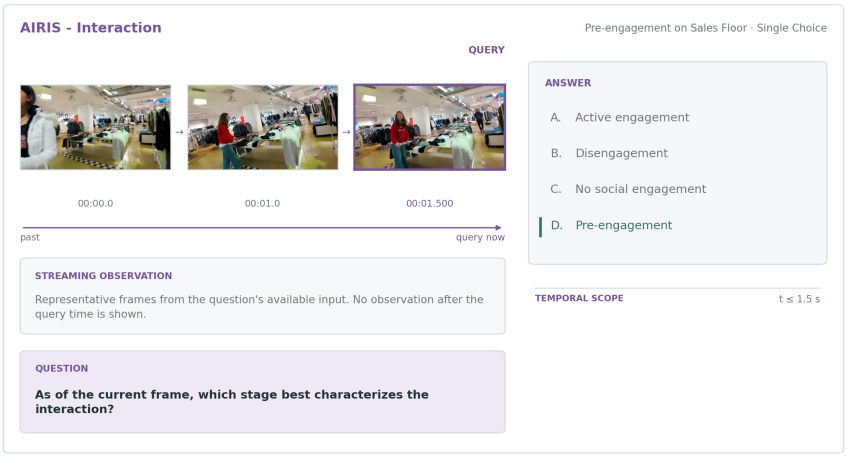


Figure 41: Representative Interaction case for pre-engagement-state recognition on a retail sales floor.

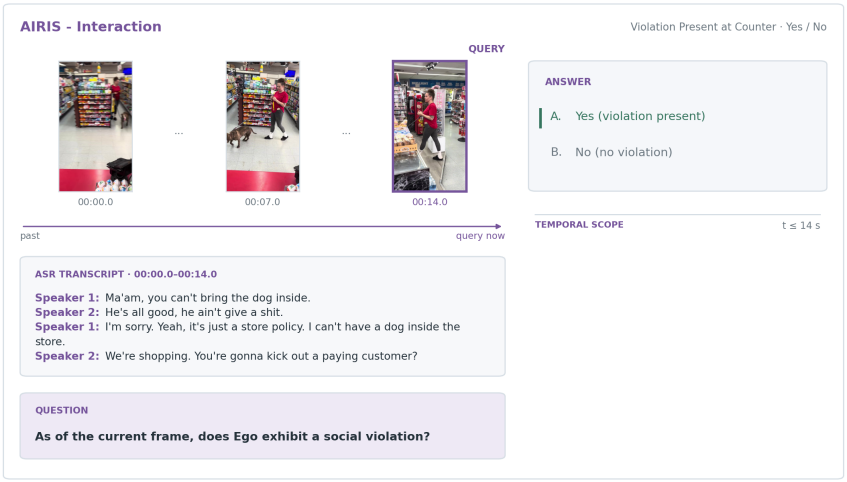


Figure 42: Representative Interaction case for query-time violation detection at a store counter.

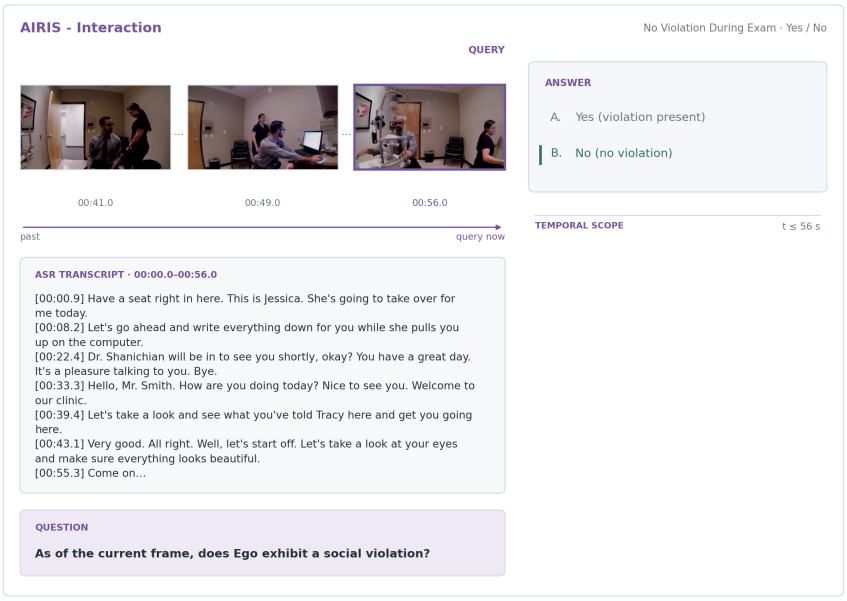


Figure 43: Representative Interaction case for query-time no-violation detection in a clinical exam.

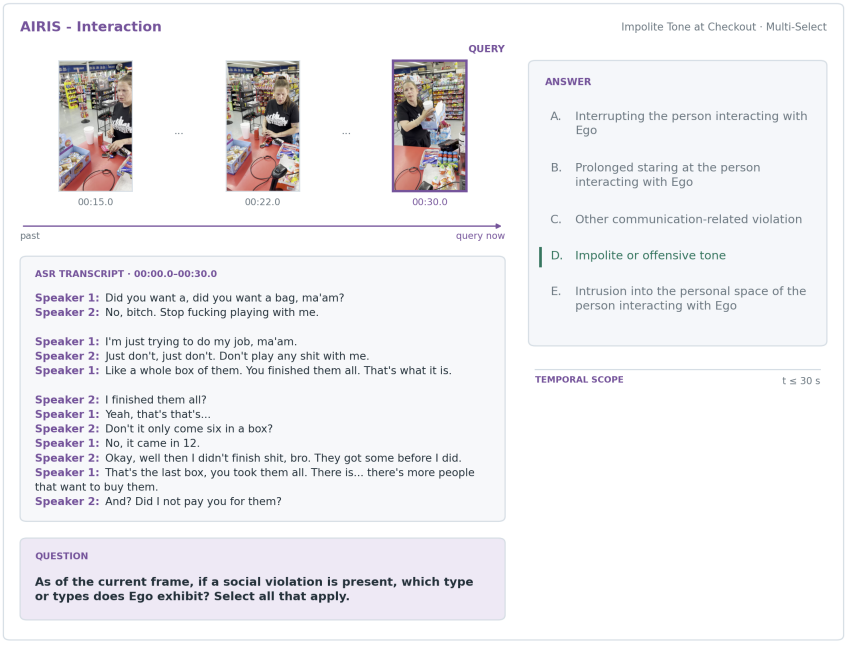


Figure 44: Representative Interaction case for query-time communication-violation typing at checkout.

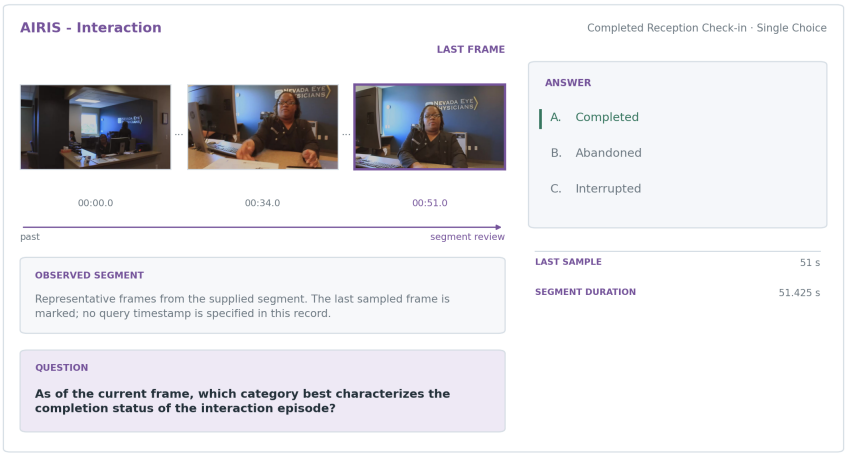


Figure 45: Representative Interaction case for episode-completion assessment at reception.

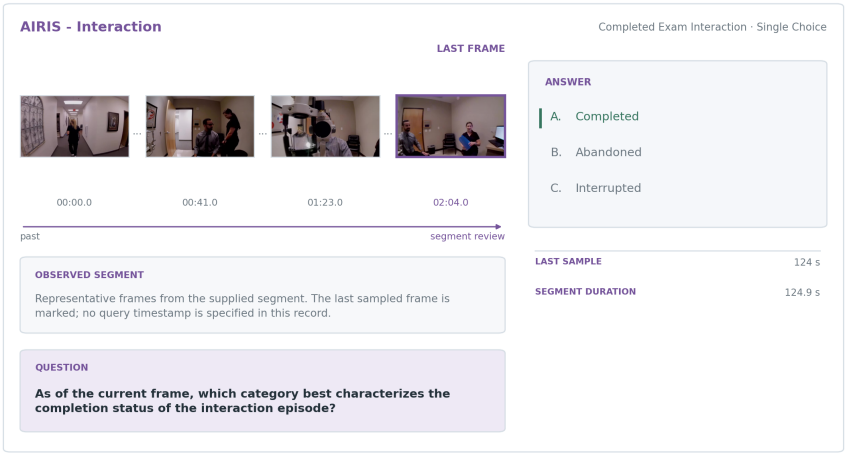


Figure 46: Representative Interaction case for episode-completion assessment in a clinical exam.

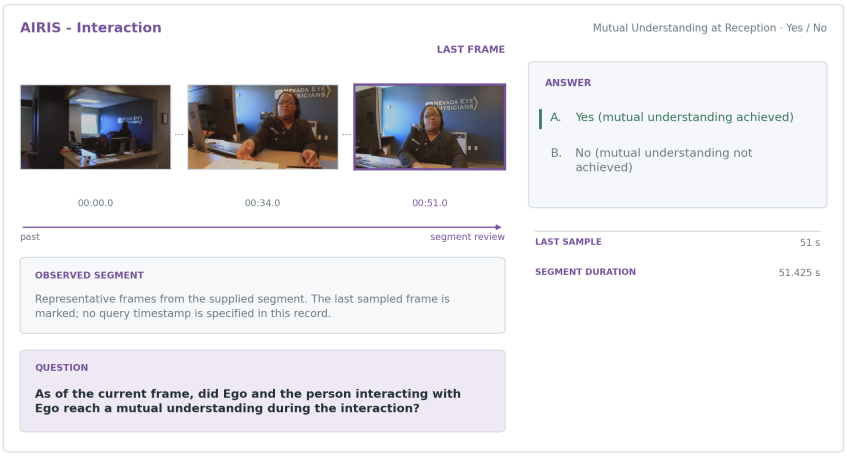


Figure 47: Representative Interaction case for mutual-understanding assessment at reception.

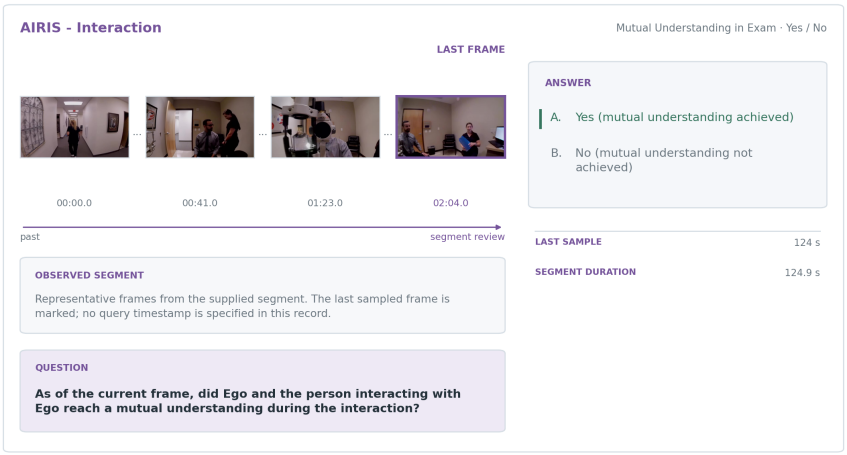


Figure 48: Representative Interaction case for mutual-understanding assessment in a clinical exam.

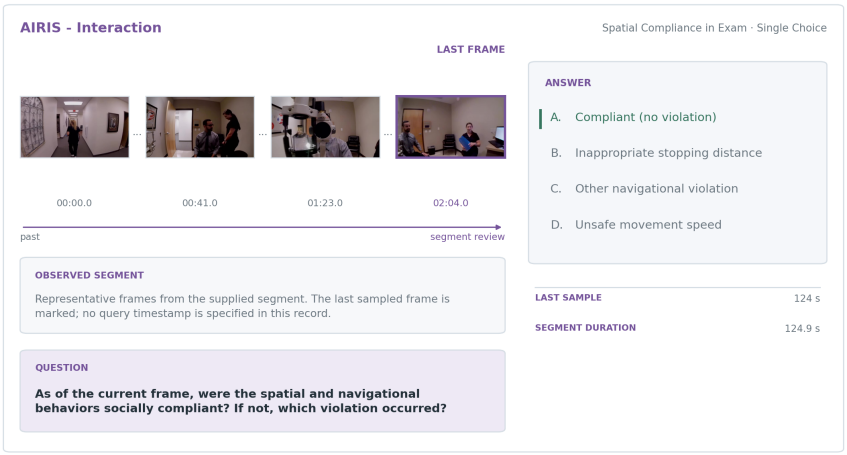


Figure 49: Representative Interaction case for spatial-compliance assessment in a clinical exam.

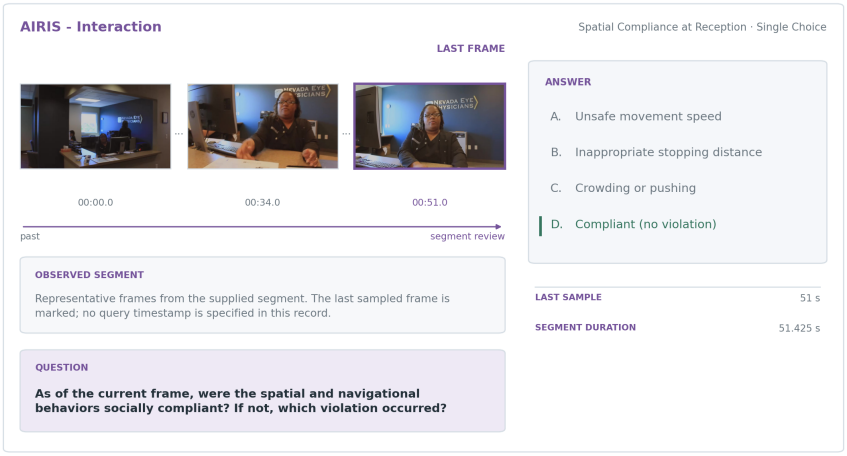


Figure 50: Representative Interaction case for spatial-compliance assessment at reception.

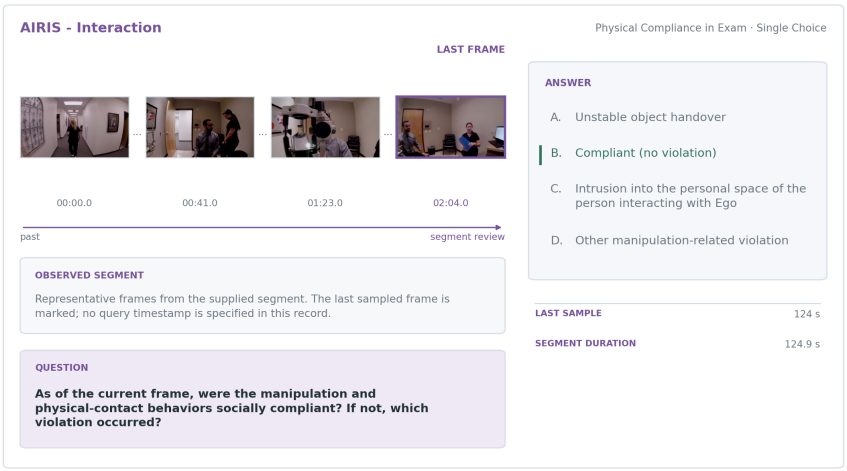


Figure 51: Representative Interaction case for physical-compliance assessment in a clinical exam.

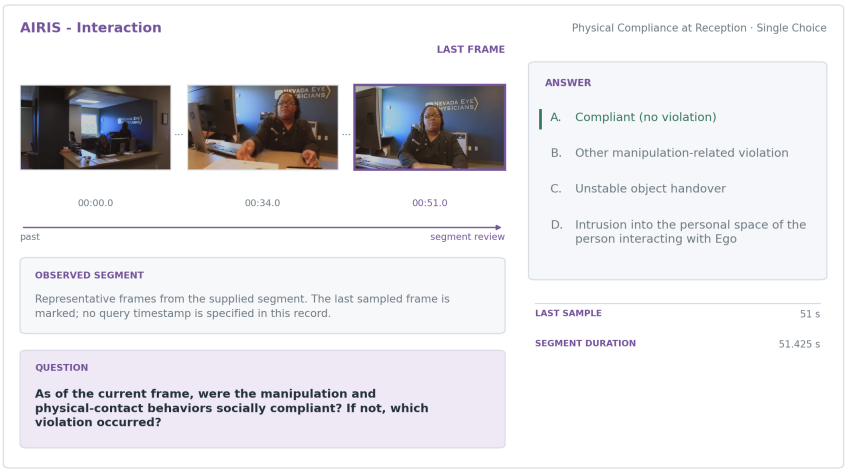


Figure 52: Representative Interaction case for physical-compliance assessment at reception.

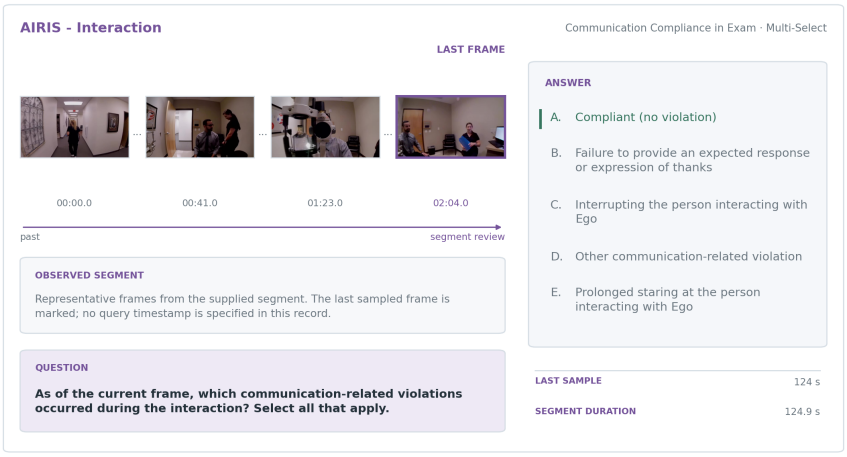


Figure 53: Representative Interaction case for communication-compliance assessment in a clinical exam.

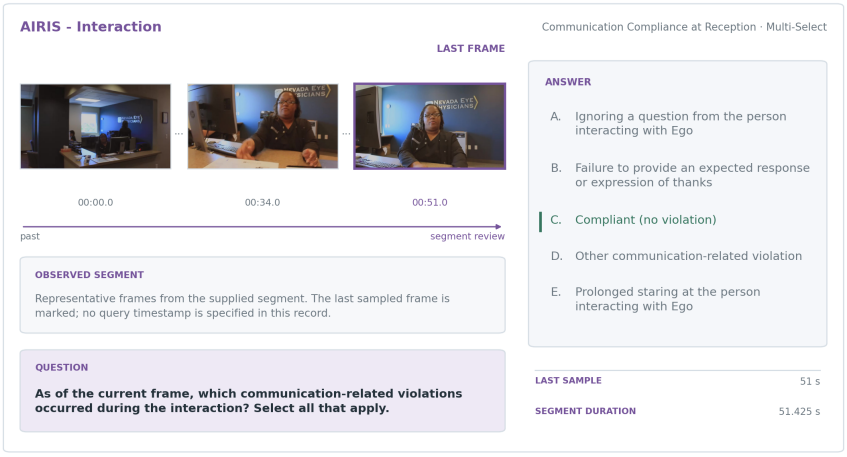


Figure 54: Representative Interaction case for communication-compliance assessment at reception.

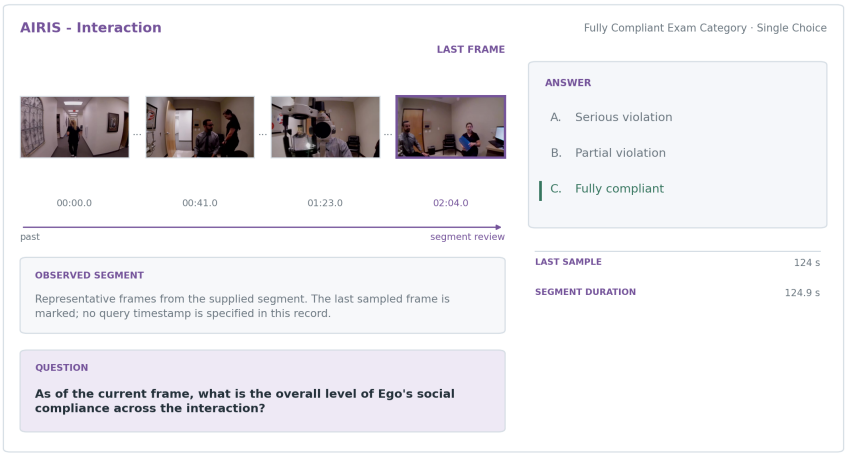


Figure 55: Representative Interaction case for overall social-compliance categorization in a clinical exam.

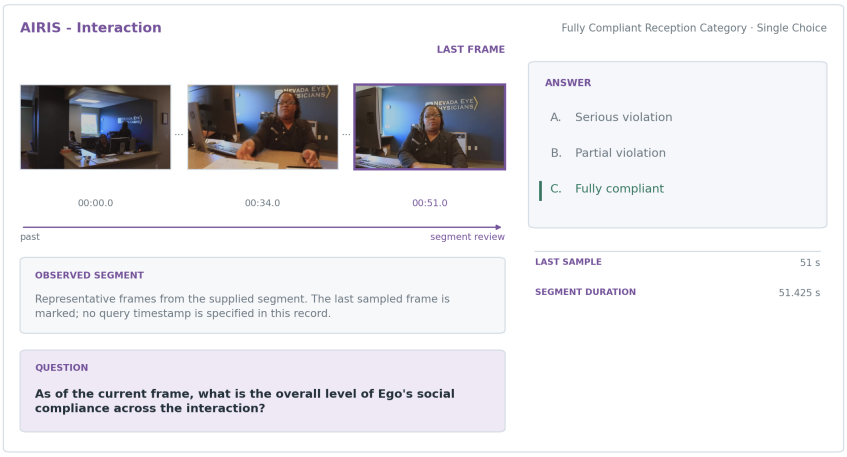


Figure 56: Representative Interaction case for overall social-compliance categorization at reception.

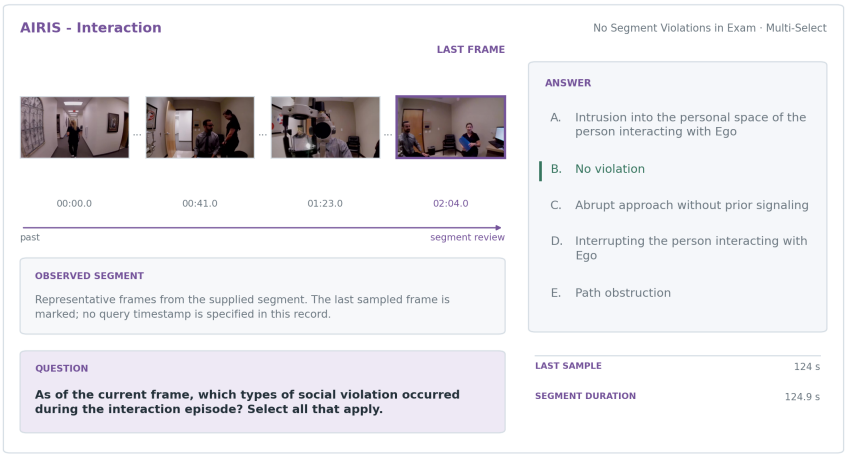


Figure 57: Representative Interaction case for segment-level violation-type assessment in a clinical exam.