E-MMSI: TOWARD EVIDENCE-GROUNDED SOCIAL INTERACTION UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce **Evidence-grounded Multi-party Multi-modal Social Interaction (E-MMSI)** understanding, a new task for spoken social videos. Beyond answering an interaction question, a model must provide observable evidence that supports its answer, making the prediction verifiable. To study this problem systematically, we draw on interaction theory to build an evidence benchmark on existing social-interaction datasets. The annotation schema represents answer-relevant actors, activities, partners, relations, contexts, and temporal anchors. Evaluation shows that current models often produce unreliable evidence in complex and dynamic multi-party interactions, such as grounding a pointing gesture to the wrong participant or time interval. To address this challenge, we construct a word-aligned social representation (WSR), which uses each spoken word's time interval as the shared alignment unit for verbal and concurrent non-verbal cues. Across three backbones and three benchmark subsets, WSR improves answer accuracy by 7.66 percentage points and Evidence F1 by 13.20 points on average.

1 INTRODUCTION

Multi-party Multi-modal Social Interaction (MMSI) understanding aims to recognize interactions among multiple participants from verbal and non-verbal cues. In practice, a system should not only answer interaction questions but also ground its answers in observable evidence, enabling its predictions to be verified and evaluated. For example, in Fig. 1, if the system predicts that "he" refers to Player0, it should provide supporting evidence, such as Player2 pointing toward Player0 with his left hand from 5 to 6 seconds. Such evidence is crucial for social-interaction applications, including meeting and classroom analytics, movie understanding, and interactive social AI.

In this paper, we introduce **Evidence-grounded MMSI** (E-MMSI), a task that evaluates whether multimodal models can ground their answers in structured, observable social evidence. Given a spoken social video and an interaction question, a model must provide an answer and supporting evidence that associates speech and behavior with the correct participants and temporal intervals. Existing MMSI tasks primarily evaluate answer accuracy (Li et al., 2026b; Xie et al., 2026; Kong et al., 2026), leaving such evidence underexplored. We build a benchmark on existing MMSI datasets with structured annotations and evaluation rubrics. Following social-interaction theory (Hoppler et al., 2022), the evidence specifies actor (who performs), activity (what is said or done, with a temporal anchor), partner (whom the actor acts toward or with), relation, and context. This structure enables fine-grained evaluation of whether generated evidence supports an answer.

We observe that models struggle to generate reliable evidence because multi-party social interactions contain dense and rapidly changing cues. Verbal evidence requires associating utterances with the correct speaker and time, whereas non-verbal evidence requires tracking who looks or gestures toward whom. A model may attribute an utterance to the wrong participant, claim that a person points with the left rather than the right hand, or associate a gaze direction with the wrong partner. These errors arise when a multimodal model fails to bind fine-grained audio and visual cues.

To address this challenge, we propose a word-aligned social representation (WSR) that aligns verbal cues with concurrent non-verbal cues using each spoken word's start and end times. The intuition is that, in spoken interactions, the verbal and non-verbal cues often evolve from one word to the next. For example, a speaker may point when emphasizing "you," then change gestures. Word-level alignment captures these changes within an utterance without the redundancy of processing

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.

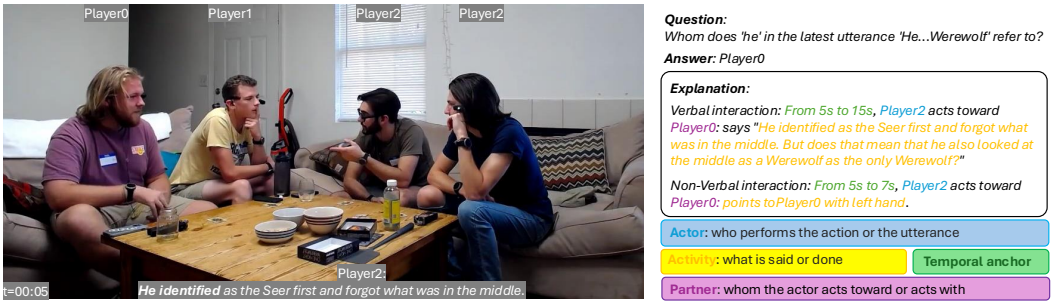

Figure 1: Illustration of the evidence-grounded MMSI (E-MMSI). Given a spoken social video and a question, a model must provide both answer and supporting evidence describing observable cues. The evidence is drawn on interaction elements: actor, activity, temporal anchor, and partner.

every frame. Averaged across Omni-MMSI PCR, Omni-MMSI STI, and Social-IQ 2.0, accuracy and Evidence F1 improve by 3.47 and 13.03 points for Qwen3-Omni-Thinking, 11.18 and 15.06 for Qwen3-VL-32B-Instruct, and 8.34 and 11.51 for InternVL3.5-38B, respectively, over their baselines without WSR. WSR thus benefits both audio-visual LLMs (AVLLMs) and vision LLMs (VLLMs).

Our contributions are: 1) We formulate E-MMSI as a task requiring answers and observable evidence, exposing the challenge of binding activities to actors, partners, temporal intervals, relations, and contexts. 2) We construct a diagnostic benchmark with structured, human-verified evidence for 1,006 QA instances, enabling evaluation of errors in participant attribution, activity content, timing, and interaction targets. 3) We introduce the word-level social representation (WSR), which uses spoken-word intervals as shared temporal anchors for verbal cues and concurrent non-verbal observations. Experiments and controlled component ablations across three benchmark subsets show that WSR consistently improves both answer accuracy and evidence grounding, demonstrating the importance of fine-grained cross-modal alignment for grounded social reasoning.

2 Related Work

Multi-modal Social Interaction Understanding. Social interaction research spans social QA (Zadeh et al., 2019; Wilf et al., 2023; Niu et al., 2026) and recognition tasks (Lee et al., 2024; Li et al., 2026a). Omni-MMSI (Li et al., 2026b) extends this to raw audio-video signals with training traces, while OmniMMI and SocialOmni study streaming interactions (Wang et al., 2025c; Xie et al., 2026). Yet answer accuracy alone does not reveal which cues support a prediction. Datasets partly expose this reasoning through human rationales (Mori et al., 2025), modality-tagged reasoning steps (Mathur et al., 2025), and alignment with human rationales (Kong et al., 2026). Complementary work annotates gesture participants (Cao et al., 2025a), evaluates social descriptions (Mathur et al., 2026), and filters shortcut questions (de Vries et al., 2026). Drawing on work in gaze (Ryan et al., 2025; Miao et al., 2025), gesture targets (Cao et al., 2025b), speaker attribution (Wang et al., 2024b; Ouyang et al., 2026), and multi-talker speech (Watanabe et al., 2020; Cornell et al., 2023), E-MMSI evaluates fine-grained interaction elements, such as actor, activity, partner, and time (Table 1).

Multimodal Evidence-Grounded Reasoning. Foundation models broaden multimodal perception: proprietary audio-video GPT-4o and Gemini 2.5 (OpenAI, 2024; Comanici et al., 2025), vision-capable Claude Fable, Opus, and Sonnet models (Anthropic, 2026a;b;c), and open audio-video models from VideoLLaMA 2 to Qwen2.5-Omni, Qwen3-Omni, MiniCPM-o 4.5, Nemotron 3 Nano Omni, and Audio-Visual Flamingo (Cheng et al., 2024; Xu et al., 2025a;b; Cui et al., 2026; NVIDIA, 2026; Ghosh et al., 2026). Evidence-grounded reasoning benchmarks, including VCR, TVQA+, and NExT-GQA, evaluate answers and their supporting evidence (Zellers et al., 2019; Lei et al., 2020; Xiao et al., 2024). Related methods gather supporting observations through iterative video retrieval (Wang et al., 2024a) and audio-video tool use (Tao et al., 2025; Xing et al., 2026); VEGAS connects social question answering with relevant-frame selection (Li et al., 2025). Our framework focuses on multi-party multi-modal social interaction scenarios, using a word-aligned social representation to ground answers in verbal and non-verbal social evidence.

Table 1: Comparison of audio-visual datasets for social interaction understanding. ✓ denotes explicit annotation, while – denotes partial coverage or no annotation. Temporal granularity: ● ○ ○○ clip level; ● ● ○○ event level; ● ● ●○ utterance level; and ● ● ●● frame level.

Dataset	Evidence	Interaction Elements					Temporal Granularity
		Actor	Partner	Activity	Relation	Context	
Social-IQ 1.0 (2019)	–	–	–	✓	–	–	● ○ ○○
MELD (2019)	–	✓	–	✓	–	–	● ● ●○
Social-IQ 2.0 (2023)	–	–	–	✓	–	–	● ● ○○
JRDB-Social (2024)	–	–	–	✓	–	✓	● ● ●●
OmniMMI (2025c)	–	✓	–	✓	–	–	● ● ○○
Social Genome (2025)	✓	–	–	✓	–	–	● ○ ○○
Omni-MMSI (2026b)	–	✓	✓	–	–	–	● ● ●○
SocialOmni (2026)	–	✓	–	✓	–	–	● ● ●○
R^3-Bench (2026)	✓	✓	–	✓	–	–	● ● ●●
SIV-Bench (2026)	–	–	–	✓	–	–	● ○ ○○
E-MMSI (ours)	✓	✓	✓	✓	✓	✓	● ● ●●

3 E-MMSI BENCHMARK

Supporting evidence allows users to verify system outputs and enables evaluators to assess whether an answer is grounded in the relevant social cues. However, as shown in Table 1, most existing social benchmarks focus on question answering or cover only subsets of the interaction elements in evidence. Temporal granularity also varies across datasets, from clip-level descriptions to frame-level annotations. We therefore constructs E-MMSI benchmark that augments each question with evidence annotations that describe the answer-relevant social elements with fine-grained temporal information. E-MMSI is the only benchmark that provides evidence with five interaction elements (i.e. actor, partner, activity, relation, and context), and fine-grained temporal grounding.

3.1 BENCHMARK DESIGN

Interaction Elements. Following interaction theory, we design evidence elements to describe interactions and support answers. The taxonomy of Hoppler et al. (2022) characterizes social interactions through actor, partner, activities, relation, context, and evaluation. We instantiate the first five as evidence elements and add a temporal anchor. Evidence comprises three claim types:

$$E = \{(a_i, v_i, p_i, \tau_i)\}_i \cup \{r_j\}_j \cup \{x_k\}_k. \tag{1}$$

The three sets denote interaction, relation, and context claims, respectively. Each interaction claim contains an **Actor** (a_i), who acts; an **Activity** (v_i), what is said or done; and a **Temporal anchor** ($\tau_i = [t_i^{\mathrm{s}}, t_i^{\mathrm{e}}]$), when it occurs. A **Partner** (p_i) specifies whom the activity involves when a partner can be identified; otherwise, it is left empty. In Fig. 1, Player2 points toward Player0 at 5–6 seconds. An interaction may be verbal or non-verbal depending on whether its activity describes speech or a visible action. **Relation** (r_j) records a social tie or role, while **Context** (x_k) records the shared situation. They may capture broader social information that spans multiple interactions.

Evidence Measurement. To measure fine-grained evidence quality, we use Evidence F1 over the annotated elements. It balances missing reference evidence against unsupported predicted evidence:

$$\mathrm{F1}_q = 100 \times 2M_q/(R_q + P_q). \tag{2}$$

For each question q, we first group interaction claims describing the same event and identify candidate reference–prediction pairs within each modality (matching details in Section B.3). We score each pair's elements: Actor and Partner use binary agreement, Activity receives 0, 0.5, or 1 for no, partial, or full agreement, and Time uses interval-union temporal IoU. We then select the highest-scoring one-to-one matching, with Relation and Context matched separately. M_q is the sum of element scores across all selected matches. R_q and P_q are the total numbers of reference and predicted elements, respectively. Finally, we macro-average question-level Evidence F1.

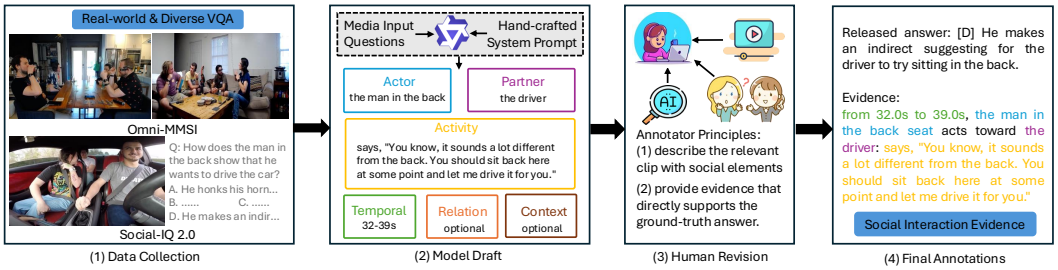

Figure 2: Construction of E-MMSI. We collect QA pairs from two complementary social-interaction datasets and use Qwen3-Omni to generate schema-guided, prefilled evidence annotations. The annotators revise the pre-annotations against the source audio-visual clips to obtain answer-supporting social evidence. Quality control combines tutorial, expert consultation, and post-annotation checks.

Table 2: Annotation statistics for E-MMSI. Elements is the total number of Actor, Activity, Temporal, Partner, Relation, and Context. Video (s) is the mean clip length, and Evidence (s) is the mean temporal union of interaction evidence. Words is the number of words in evidence descriptions.

Dataset	Samples	Actor	Activity	Partner	Relation	Context	Elements	Words	Video (s)	Evidence (s)
Omni-MMSI	287	513	513	513	–	–	2,052	6,627	18.75	3.96
Social-IQ 2.0	719	2,188	2,188	1,567	78	646	8,855	39,804	57.96	14.08
All	**1,006**	**2,701**	**2,701**	**2,080**	**78**	**646**	**10,907**	**46,431**	**46.78**	**11.19**

3.2 Benchmark Construction

Data collection. Our data collection aims to cover realistic and diverse social scenarios. We therefore draw all examples from audio-visual videos of spoken social interactions and select two complementary source datasets. Omni-MMSI (Li et al., 2026b) provides challenging pronoun-coreference and speaking-target questions from unscripted social-deduction games; even its strongest reported pipeline achieves less than 50% accuracy. Social-IQ 2.0 (Wilf et al., 2023) broadens the coverage with four-choice questions drawn from socially rich YouTube videos, movie clips, and in-vehicle interactions. Together, the two sources span different settings, interaction patterns, and reasoning demands (Table 7). The benchmark contains 112 pronoun-coreference and 175 speaking-target instances from Omni-MMSI, together with 719 QA instances from Social-IQ 2.0.

Evidence annotation. To speed up the process, we first use Qwen3-Omni (Xu et al., 2025b) to generate a preliminary annotation following the element schema in Section 3.1. Seven annotators with relevant academic backgrounds then review each audio-visual clip, QA pair, and prefilled annotation. They can delete, add, or modify interaction, relation, and context claims. A valid annotation must (1) describe the relevant clip with social elements and (2) provide evidence that supports the ground-truth answer. E.g., in Fig. 2, the passenger suggests that the driver sit in the back and let him drive (32 to 39 s), supporting answer D. We control quality in three stages: annotators complete 10 training samples with expert feedback, consult the expert throughout annotation, and revise issues flagged by a post-annotation GPT-5.6 Sol check. We further evaluate inter-annotator agreement and conduct an independent video-grounded quality audit. We can see 95% overall quality accuracy for all subsets and 84% agreement on Omni-MMSI. Section A.2 provides more annotation details.

Statistical analysis. Table 2 summarizes the annotations by evidence element. Overall, E-MMSI contains 1,006 samples, 10,907 scored elements, and 46,431 evidence-content words. The two Omni-MMSI tasks contain 2,052 elements, while Social-IQ 2.0 contains 8,855 elements across interaction, relation, and context claims. Their mean temporal evidence unions are 3.96 and 14.08 seconds, substantially shorter than their mean video lengths of 18.75 and 57.96 seconds. These dense annotations allow a system's supporting evidence to be assessed comprehensively.

4 E-MMSI Framework

4.1 Overview

As shown in Tables 3 to 5, current multimodal models still obtain low evidence scores on E-MMSI. Multi-party social scenes are complex and dynamic: several participants may speak and act at over-

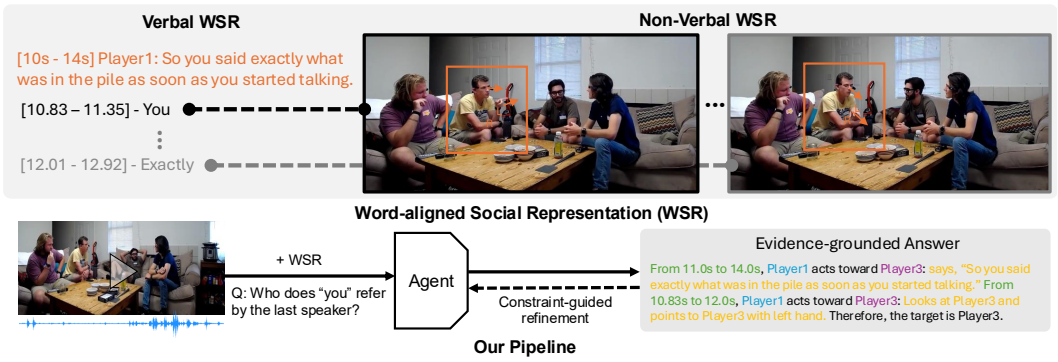

Figure 3: E-MMSI framework with word-aligned social representation (WSR). The verbal branch records speaker-attributed words and timestamps; the non-verbal branch overlays participant locations, gaze, and gestures on aligned frames. The agent uses the query clip, question, and WSR to generate an answer with evidence. Constraint-guided refinement uses deterministic checks to trigger a second-pass review of the previous answer and evidence by the same model.

lapping times, while short cues such as gaze and gesture can change quickly. A model must therefore associate verbal and non-verbal cues with the correct participants and moments.

To help address this challenge, we propose word-aligned social representation (WSR) to compactly capture the dynamic social interactions in spoken social videos. First, the WSR is converted from the video and represents the verbal and non-verbal social cues in a word-aligned manner. Then, the video, question, and WSR are provided to an agent that is prompted to generate an answer with structured supporting evidence. Finally, constraint-guided refinement uses deterministic checks to trigger a second-pass review of the answer and evidence by the same model.

4.2 Word-Aligned Social Representation

WSR seeks a compact representation of temporally aligned verbal cues, such as speech, and non-verbal cues, such as gaze and gesture. In interpersonal interactions, the verbal and non-verbal cues often evolve from one word to the next. We therefore use spoken words as temporal anchors: each non-verbal observation is associated with the word interval during which it occurs. This provides finer temporal precision than utterance-level without the redundancy of frame-level.

We first use ASR and word alignment to obtain each spoken word and its start and end times (Bain et al., 2023). Visual re-identification (ReID) tracks participants across frames (He et al., 2016), while SpeechBrain (Ravanelli et al., 2021) or TalkNet (Tao et al., 2021) identifies the active speaker and associates that speaker with a visual track. For the frames aligned with each word, a whole-body keypoint estimator (Jiang et al., 2024) extracts participant locations and body landmarks, from which we approximate gaze and gesture directions. The resulting representation is

$$M = \{z_n\}_{n=1}^{T}, \qquad z_n = (\tau_n, v_n, nv_n), \tag{3}$$

where T is the number of spoken words, $\tau_n = [t_n^{\mathrm{s}}, t_n^{\mathrm{e}}]$ is the time interval of the nth word, v_n is the verbal cue spoken during τ_n, and nv_n contains the concurrent non-verbal cues.

Verbal WSR. To feed the verbal cues to the agent, we serialize $\{v_n\}_{n=1}^{T}$ with their timestamps as a word-aligned transcript. The transcript contains information about who speaks, what word is spoken, and when it occurs. As shown in Figure 3, the speaker-attributed utterance like "Player1: So..." is presented while individual words such as "you" are assigned precise start and end times.

Non-verbal WSR. To feed the non-verbal cues to the agent, we preserve $\{nv_n\}_{n=1}^{T}$ in both the text and visual channels. As illustrated in Figure 3, the frames aligned with τ_n include participant boxes, gaze arrows, and hand labels. We approximate gaze direction from the estimated eye and ear coordinates and render it as an arrow, providing a visual cue to where a participant may be looking. We also label each participant's left and right hands with `L` and `R`, respectively. These labels refer

to the participant's own left and right, addressing our observation that models often confuse the two hands when interpreting gestures. We further estimate finger direction from hand keypoints. When a fingertip is sufficiently far from the palm and farther away than the other fingertips, we identify the finger as extended and draw an arrow along its estimated direction. Participant identity descriptions are provided in the text channel, and the overlaid video replaces the original query clip as input.

Together, WSR exposes which participant speaks or acts, the action direction, and when the activity occurs. Besides, the agent can connect a spoken expression to a simultaneous gaze or gesture.

4.3 Evidence Generation and Refinement

The agent receives the question, video, and WSR. AVLLMs additionally receive the query audio, whereas VLLMs access speech through the transcript. The agent returns an answer with structured evidence following Equation (1). The complete prompts are provided in Appendix B.

Constraint-guided refinement. The agent may generate imperfect answers and supporting evidence. We use deterministic checks on temporal constraints to determine when refinement is needed: each evidence interval must have finite boundaries satisfying $0 \leq t_n^{\mathrm{s}} < t_n^{\mathrm{e}} \leq L$, where L is the query duration. When a violation is detected, the same model receives its previous output together with the query inputs and regenerates the answer and supporting evidence.

5 Experiments

5.1 Setup

Benchmark. We evaluate E-MMSI described in Section 3.2: 112 pronoun-coreference (PCR) and 175 speaking-target identification (STI) instances from Omni-MMSI (Li et al., 2026b), and 719 four-choice QA instances from Social-IQ 2.0 (Wilf et al., 2023), for a total of 1,006 questions.

Models. We compare proprietary Claude Fable 5.1, Opus 5.5, and Sonnet 5 (Anthropic, 2026a;b;c), open-source audio-visual LLMs (AVLLMs), and vision-language LLMs (VLLMs). We apply WSR to Qwen3-Omni-Thinking (Xu et al., 2025b), Qwen3-VL-32B-Instruct (Bai et al., 2025), and InternVL3.5-38B (Wang et al., 2025a); additional VLLMs appear in Appendix C. + Whisper adds Whisper audio transcripts; + WSR adds word-aligned verbal and non-verbal cues.

Implementation. WSR uses WhisperX (Bain et al., 2023) for transcription and word alignment, RTMW3D (Jiang et al., 2024) for whole-body keypoints, ResNet-50 re-identification (He et al., 2016) for visual tracking, and SpeechBrain (Ravanelli et al., 2021) on Omni-MMSI or TalkNet (Tao et al., 2021) on Social-IQ 2.0 for speaker identification. We use Qwen3.8-27B (Qwen Team, 2026b) as the evidence judge. Open-source models run with vLLM. All evaluated open-source AVLLMs and VLLMs use temperature 0, top-p 1, and a fixed seed of 42. Videos are sampled at a target rate of 1 FPS. The maximum generation length is 4,096 tokens for AVLLMs and VLLMs. Agent instructions are provided in Appendix B and the supplementary material.

Evaluation protocol. We report accuracy and Evidence F1 (Section 3.1).

5.2 Quantitative Results

We compare proprietary models, open-source AVLLMs, and tool-assisted VLLMs on PCR, STI, and Social-IQ 2.0 (Tables 3 to 5). Among evaluated models, Fable 5.1 achieves the best PCR accuracy (60.71%), while Qwen3-VL-32B-Instruct + WSR achieves the best Evidence F1 (43.19). On STI, Qwen3-VL-32B-Instruct + WSR leads both accuracy (41.71%) and Evidence F1 (46.04). On Social-IQ 2.0, Qwen3-Omni-Thinking + WSR achieves the best accuracy (87.50%), while Claude Opus 5.5 leads Evidence F1 (50.23). We also ablate verbal and non-verbal WSR on all three subsets using Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025b) with thinking mode disabled.

WSR improves evidence grounding for both AVLLMs and VLLMs. On PCR/STI, gains over the base model are 20.58/13.92 points for Qwen3-Omni-Thinking, 23.21/16.67 for Qwen3-VL-32B-Instruct, and 16.35/12.35 for InternVL3.5-38B. Additional VLLMs show consistent evidence gains on PCR/STI, but Social-IQ 2.0 gains vary by backbone (Appendix C).

Table 3: Omni-MMSI PCR results (%). Evid. denotes overall Evidence F1; V. and NV. denote verbal and non-verbal element F1. **Bold**/underlined marks the best/second-best non-human result.

			Verbal (V.)					Non-verbal (NV.)				
Method	**Acc.**	**Evid.**	**Actor**	**Act.**	**Time**	**Part.**	**Total**	**Actor**	**Act.**	**Time**	**Part.**	**Total**
Human	90.00	74.57	96.67	63.33	59.61	56.67	69.07	100.00	88.89	54.63	88.89	83.10
Proprietary Models												
Fable 5.1	**60.71**	26.33	24.76	44.60	28.97	19.97	29.94	18.30	23.59	15.84	15.20	18.35
Opus 5.5	53.57	23.52	23.07	38.33	26.70	15.03	26.34	22.07	23.87	13.51	13.95	18.19
Sonnet 5	53.57	25.30	29.61	40.07	27.10	22.38	30.12	16.82	24.45	10.37	9.96	15.62
Open-Source AVLLMs												
Qwen2.5-Omni-3B	26.86	4.97	3.57	15.62	3.19	0.00	6.40	0.00	0.00	0.00	0.00	0.00
Qwen2.5-Omni-7B	30.86	13.30	12.80	36.61	6.47	0.00	16.08	0.00	0.00	0.00	0.00	0.00
MiniCPM-o 4.5	27.68	3.50	7.14	8.04	0.00	0.00	4.34	0.00	0.00	0.00	0.00	0.00
Nemotron-3-Nano-Omni	44.64	12.80	12.50	35.27	7.96	6.25	17.01	0.00	0.00	0.00	0.00	0.00
Qwen3-Omni-Instruct	44.64	7.20	13.39	16.96	1.30	3.57	9.64	0.00	0.00	0.00	0.00	0.00
Qwen3-Omni-Thinking	46.43	21.09	25.02	50.60	19.01	2.68	27.63	2.06	4.12	0.55	1.43	2.14
Qwen3-Omni-Thinking + WSR (Ours)	50.89	41.67	55.80	59.30	37.57	38.69	48.37	37.36	38.46	11.08	**28.57**	**28.87**
Open-Source VLLMs												
Qwen3-VL-32B-Instruct	43.75	19.98	24.11	32.14	8.10	22.32	22.00	13.76	25.69	2.97	18.35	15.19
Qwen3-VL-32B-Instruct + Whisper	49.11	33.11	26.79	**59.97**	34.59	31.25	38.66	14.71	**40.69**	**16.90**	22.45	23.59
InternVL3.5-38B	50.89	15.37	16.96	26.79	4.72	7.14	15.12	15.69	18.14	4.20	17.89	13.79
InternVL3.5-38B + Whisper	48.21	24.33	24.76	46.91	28.61	12.50	30.50	6.58	9.87	2.48	9.59	7.04
InternVL3.5-38B + WSR (Ours)	56.25	31.72	49.40	52.06	31.59	16.67	40.18	8.33	11.90	2.76	7.14	7.54
Qwen3-VL-32B-Instruct + WSR (Ours)	53.57	**43.19**	**60.42**	59.08	**40.57**	**43.75**	**51.73**	**41.35**	39.42	7.38	25.24	28.29

Table 4: Omni-MMSI STI results (%). Evid. denotes overall Evidence F1; V. and NV. denote verbal and non-verbal element F1. **Bold**/underlined marks the best/second-best non-human result.

			Verbal (V.)					Non-verbal (NV.)				
Method	**Acc.**	**Evid.**	**Actor**	**Act.**	**Time**	**Part.**	**Total**	**Actor**	**Act.**	**Time**	**Part.**	**Total**
Human	80.00	69.14	96.67	53.33	68.57	90.00	77.14	66.67	44.44	55.56	77.78	61.11
Proprietary Models												
Fable 5.1	35.43	27.32	27.90	42.21	30.44	22.32	30.77	24.14	37.98	23.15	17.53	25.75
Opus 5.5	32.57	27.12	24.30	39.38	29.62	18.76	28.53	27.62	33.95	22.29	22.99	26.87
Sonnet 5	30.29	27.49	30.19	42.48	28.40	20.13	30.31	25.05	36.14	20.58	17.14	24.86
Open-Source AVLLMs												
Qwen2.5-Omni-3B	26.86	3.24	4.76	5.24	4.37	0.57	4.14	0.00	0.00	0.00	0.00	0.00
Qwen2.5-Omni-7B	30.86	7.43	14.29	14.24	7.41	1.71	10.59	0.00	0.00	0.00	0.00	0.00
MiniCPM-o 4.5	23.43	1.31	3.81	2.19	1.08	0.00	2.03	0.00	0.00	0.00	0.00	0.00
Nemotron-3-Nano-Omni	26.29	8.15	11.42	15.10	4.30	7.34	9.54	7.40	6.97	3.36	4.57	5.60
Qwen3-Omni-Instruct	32.57	9.30	14.42	23.21	0.81	12.19	12.72	0.66	0.66	0.08	0.00	0.39
Qwen3-Omni-Thinking	34.86	22.09	27.79	41.40	10.54	17.66	24.55	22.62	26.29	8.60	14.75	18.11
Qwen3-Omni-Thinking + WSR (Ours)	36.00	36.01	54.32	46.31	34.52	31.33	41.81	32.32	28.56	20.38	17.68	24.77
Open-Source VLLMs												
Qwen3-VL-32B-Instruct	28.57	29.37	41.71	40.86	12.60	26.29	30.38	34.33	41.22	8.46	19.56	25.89
Qwen3-VL-32B-Instruct + Whisper	28.00	34.84	35.81	**54.86**	39.90	23.24	38.72	26.98	42.66	25.57	15.87	27.90
InternVL3.5-38B	25.71	18.72	30.29	30.29	9.60	11.43	21.46	16.37	23.37	5.06	10.24	13.82
InternVL3.5-38B + Whisper	26.86	22.91	28.76	44.48	30.39	14.10	30.99	4.61	7.89	5.38	3.29	5.37
InternVL3.5-38B + WSR (Ours)	40.57	31.07	53.33	45.05	33.41	27.81	40.54	10.68	12.50	5.79	5.56	8.70
Qwen3-VL-32B-Instruct + WSR (Ours)	**41.71**	**46.04**	**62.95**	54.62	**42.45**	**34.95**	**49.08**	**49.90**	**53.79**	**26.43**	**30.94**	**40.30**

Non-verbal timing and targets remain difficult to ground. As shown in Tables 3 and 4, non-verbal evidence is generally weaker than verbal evidence on Omni-MMSI. With Qwen3-VL-32B-Instruct + WSR, verbal/non-verbal totals are 51.73/28.29 on PCR and 49.08/40.30 on STI. Non-verbal Time and Partner remain particularly challenging: PCR scores are 7.38 and 25.24, versus 41.35 for Actor; STI scores are 26.43 and 30.94, versus 49.90 for Actor. This pattern is consistent with the difficulty of tracking brief pointing gestures and gaze shifts: recognizing a participant does not ensure that the model identifies when a cue occurs or whom it targets. WSR does not eliminate these errors.

STI has the largest human-model accuracy gap. The tasks place different demands on social evidence. PCR relies on verbal context to resolve a pronoun, often supplemented by a visual cue. STI requires identifying the current addressee from momentary gaze, gesture, and body orientation. It is the hardest task in answer selection: the best model reaches 41.71% accuracy against 80.00% for humans, compared with 60.71% versus 90.00% on PCR. Social-IQ 2.0 offers broader conversational, contextual, and relational cues and has the smallest gaps to the reported human scores: best model accuracy is 87.50% versus 100.00%, and Evidence F1 is 50.23 versus 51.12. Its questions can

Table 5: Social-IQ 2.0 results (%). Evid. is overall Evidence F1; V., NV., and Other are element F1. **Bold**/underlined marks the best/second-best non-human result in each column.

			Verbal (V.)				Non-verbal (NV.)				Other evidence	
Method	**Acc.**	**Evid.**	**Actor**	**Act.**	**Time**	**Part.**	**Actor**	**Act.**	**Time**	**Part.**	**Context**	**Relation**
Human	100.00	51.12	43.06	41.39	20.71	35.56	59.55	59.55	22.75	40.50	85.19	100.00
Proprietary Models												
Fable 5.1	84.28	44.73	44.89	46.63	29.41	44.58	54.18	**50.86**	29.02	53.29	55.99	84.89
Opus 5.5	83.87	**50.23**	52.93	**52.45**	**33.56**	**51.28**	**61.11**	48.39	**31.04**	**58.33**	60.32	88.89
Sonnet 5	76.63	43.25	45.84	50.07	31.07	44.66	43.18	44.23	24.00	40.09	48.69	**91.11**
Open-Source AVLLMs												
Qwen2.5-Omni-3B	69.23	18.33	8.08	4.64	0.00	3.27	33.92	26.53	0.00	20.72	41.99	28.56
Qwen2.5-Omni-7B	75.00	30.26	31.87	28.81	8.67	29.88	14.36	13.36	3.02	10.72	65.06	36.67
MiniCPM-o 4.5	71.15	16.96	17.43	25.98	3.53	13.91	7.50	13.12	3.17	2.90	46.47	35.56
Nemotron-3-Nano-Omni	75.00	11.53	20.74	18.16	2.74	8.38	3.75	3.75	1.85	0.00	0.00	0.00
Qwen3-Omni-Instruct	81.62	39.96	48.28	34.12	10.00	41.37	31.00	28.35	10.82	27.15	74.36	73.33
Qwen3-Omni-Thinking	82.69	42.46	43.40	41.35	17.45	44.16	34.38	31.31	15.28	34.54	62.18	66.67
Qwen3-Omni-Thinking + WSR (Ours)	**87.50**	47.06	**54.08**	50.67	27.50	47.84	41.25	36.77	19.75	37.20	**81.09**	70.00
Open-Source VLLMs												
Qwen3-VL-32B-Instruct	75.00	35.42	24.81	15.07	8.40	20.42	47.11	41.58	21.31	43.10	72.12	74.44
Qwen3-VL-32B-Instruct + Whisper	85.58	46.24	46.88	39.73	21.44	46.99	57.17	50.67	26.26	52.20	78.53	**91.11**
InternVL3.5-38B	79.81	22.65	24.75	16.07	6.91	25.40	28.54	21.75	8.37	30.89	5.77	3.33
InternVL3.5-38B + Whisper	83.65	23.92	28.14	21.00	11.64	27.73	21.96	15.99	8.38	17.00	2.88	8.89
InternVL3.5-38B + WSR (Ours)	84.62	28.49	34.00	28.63	14.74	32.23	20.62	20.21	10.36	14.01	0.96	6.67
Qwen3-VL-32B-Instruct + WSR (Ours)	85.58	40.73	44.28	40.73	17.80	42.86	43.21	45.75	19.84	41.73	76.60	72.22

Table 6: WSR ablation (%). Evid. is overall Evidence F1. Settings differ only in the two WSR inputs. **Bold** marks the best per dataset; dashes denote unavailable fields.

			Overall		Verbal (V.)				Non-verbal (NV.)				Other	
Dataset	**Verbal WSR**	**Non-verbal WSR**	**Acc.**	**Evid.**	**Actor**	**Act.**	**Time**	**Part.**	**Actor**	**Act.**	**Time**	**Part.**	**Context**	**Relation**
Omni-MMSI PCR	×	×	46.43	21.09	25.02	50.60	19.01	2.68	2.06	4.12	0.55	1.43	–	–
	✓	×	**50.89**	37.39	53.27	56.70	35.97	35.27	29.03	31.72	5.54	19.35	–	–
	×	✓	46.43	26.38	29.46	54.44	13.68	26.79	22.77	35.64	4.05	18.81	–	–
	✓	✓	**50.89**	**41.67**	**55.80**	**59.30**	**37.57**	**38.69**	**37.36**	**38.46**	**11.08**	**28.57**	–	–
Omni-MMSI STI	×	×	34.86	22.09	27.79	41.40	10.54	17.66	22.62	26.29	8.60	14.75	–	–
	✓	×	**36.00**	26.43	28.93	40.24	29.37	23.37	21.66	28.39	19.82	14.06	–	–
	×	✓	29.43	19.48	23.92	37.14	9.67	18.76	18.74	22.39	6.73	14.19	–	–
	✓	✓	**36.00**	**36.01**	**54.32**	**46.31**	**34.52**	**31.33**	**32.32**	**28.56**	**20.38**	**17.68**	–	–
Social-IQ 2.0	×	×	82.69	42.46	43.40	41.35	17.45	44.16	34.38	31.31	15.28	34.54	62.18	66.67
	✓	×	83.65	45.28	48.63	**51.45**	**31.28**	47.81	31.21	33.77	14.54	31.84	79.49	53.33
	×	✓	**87.50**	44.51	51.13	45.39	19.38	45.87	38.33	**37.17**	15.14	36.81	80.77	56.67
	✓	✓	**87.50**	**47.06**	**54.08**	50.67	27.50	**47.84**	**41.25**	36.77	**19.75**	**37.20**	**81.09**	**70.00**

draw on an extended exchange rather than a fleeting cue, and scenes centered on a few clearly visible participants reduce ambiguity about who is interacting. For example, Figure 2 asks how the back-seat passenger shows that he wants to drive. His suggestion that the driver sit in the back and let him drive provides direct conversational evidence, without requiring a brief gesture or gaze shift to be resolved. Such cues may make both answer selection and evidence identification more accessible, although precise timing remains difficult. Together, these results point to fine-grained visual interaction in multi-party exchanges as a remaining bottleneck.

Tool-assisted VLLMs can handle spoken-video tasks. VLLMs process a video's visual content but not its audio. Whisper transcripts and the tools used to construct WSR make speech information accessible to these models. On PCR (Table 3), Qwen3-VL-32B-Instruct's Evidence F1 rises from 19.98 to 33.11 with Whisper and 43.19 with WSR. With WSR, it also exceeds Qwen3-Omni-Thinking + WSR in accuracy (53.57% versus 50.89%) and Evidence F1 (43.19 versus 41.67). Tool-assisted VLLMs can therefore tackle spoken-video tasks without native audio input.

WSR improves answering and grounding by linking speech with behavior. Across the three subsets, WSR improves accuracy by 1.14–4.81 points and Evidence F1 by 4.60–20.58 over the no-WSR setting (Table 6). These gains support a representation tailored to spoken social interaction: linking words to speakers and concurrent non-verbal cues helps models follow who says or does what as the conversation unfolds, supporting both answer selection and evidence grounding.

Verbal and non-verbal WSR are complementary. Combining them yields the highest Evidence F1 on every subset, exceeding the stronger single component by 4.28 points on PCR, 9.58 on STI,



Figure 4: Illustration of Qwen3-Omni-Thinking (Baseline) and the same model with WSR (Ours). WSR improves the alignment of speech and gaze evidence as participants' actions change.

and 1.78 on Social-IQ 2.0 (Table 6). The transcript provides word-aligned speech and speaker context, while the overlays expose concurrent behavior and interaction targets. Their value lies in this combination, not in uniformly positive standalone effects. On STI, the non-verbal overlay alone lowers both accuracy and Evidence F1, but adding it to verbal WSR raises Evidence F1 from 26.43 to 36.01 and verbal Actor F1 from 28.93 to 54.32. On Social-IQ 2.0, non-verbal WSR produces the largest accuracy gain, whereas verbal WSR is stronger on verbal Activity and Time. Combining both branches nevertheless yields the strongest overall evidence grounding.

5.3 Qualitative Results

Figure 4 compares Qwen3-Omni-Thinking without WSR (Baseline) and with WSR (Ours). As Player1 says "So, he is a werewolf. Boom!", participants point, move objects on the table, and make celebratory hand movements. These actions change during the utterance rather than forming a persistent gesture. Pronoun resolution requires tracking the actor and target as speech unfolds.

The baseline confuses time, speech, and action: it places "So you were lying. So he is the werewolf" from 20 to 24 s and describes Player1 pointing toward Player3, incorrectly choosing Player3. The reference places the utterance from 24 to 28 s and cites Player1's gaze toward Player0. A pointing gesture alone is insufficient evidence if it is linked to the wrong spoken words or target.

With WSR, Qwen3-Omni-Thinking recovers the phrase from 24 to 27 s and identifies Player1 looking toward Player0, resolving the pronoun. Its gaze interval remains imprecise (25 to 26 s versus the reference's 26 to 27 s). The challenge is linking rapidly changing behavior to specific words and participants. WSR makes these associations explicit, helping models follow spoken interactions. We provide additional qualitative results of Social-IQ 2.0 in Appendix D.

6 Conclusion

We introduced E-MMSI, a task and diagnostic benchmark for grounding social answers in observable evidence. Structured annotations and Evidence F1 assess how models associate activities with actors, partners, time, relations, and context. Our word-aligned social representation (WSR) aligns speaker-attributed speech with concurrent non-verbal cues. Across three subsets, WSR consistently improves accuracy and evidence grounding for the three main backbones over their baselines without WSR. Component ablations show complementary benefits from verbal and non-verbal cues, supporting fine-grained temporal alignment for social reasoning. Evaluation on E-MMSI reveals persistent difficulties in localizing non-verbal cues in time and identifying their interaction partners. Tasks that depend on brief, subtle interaction cues, such as STI, remain particularly challenging, with substantial gaps between model and human performance.

Limitations. WSR can propagate errors in transcription, speaker identification, and visual keypoint estimation into generated evidence. Word-based anchors also provide less explicit temporal structure for non-verbal interactions during pauses or speech-free segments (Appendix F).

Reproducibility Statement

Section 5.1 reports the evaluated models, preprocessing tools, and inference settings. Appendix A documents dataset selection, annotation procedures, and quality checks. Appendix B provides the agent and judge prompts and details event grouping, matching, and Evidence F1 computation. Section C.3 and Appendix E describe the refinement instructions and tool-assisted Claude evaluation, respectively. The supplementary material contains all three evidence annotation sets and complete prompt files. We will also release code for WSR construction and evidence evaluation.

Ethics Statement

Systems that explain multi-party social interactions could support meeting analytics, education, and assistive applications, but they also raise privacy concerns when applied to recordings of real people. Erroneous or hallucinated evidence could misrepresent a person's behavior or intentions. Our benchmark and framework help measure and reduce these errors, but generated evidence should not be treated as a verified account of an individual's intentions or used on its own for high-stakes decisions about them. We use videos from existing, publicly released datasets under their respective licenses; no new recordings of human subjects were collected. The supplementary annotations add structured evidence to these source datasets and do not replace their access or usage conditions.

AI Use Statement

Except as disclosed below, generative AI did not formulate the research questions or hypotheses, design the benchmark, schema, method, or experiments, interpret results, or draw conclusions.

Models in the framework and evaluation. We evaluate the proprietary and open-source AVLLMs and VLLMs listed in Section 5.1 and Tables 3 to 5, including experiments with WSR and constraint-guided refinement. Qwen3.8-27B judges semantic agreement between reference and predicted evidence elements for Evidence F1; the appendix compares it with Qwen2.5-72B-Instruct.

Benchmark construction and quality control. Qwen3-Omni drafted schema-guided evidence, and GPT-5.6 Sol flagged issues for human review. Seven trained annotators checked each draft and flag against the audio, video, and QA pair, then revised, replaced, accepted, or rejected it. Neither model determined the final references; all released evidence was human-verified.

Writing and implementation assistance. OpenAI Codex and Claude Code assisted with editing, LaTeX, tables, and code. The authors reviewed all outputs and take full responsibility.

REFERENCES

Anthropic. Introducing Claude Fable 5.1 and Claude Mythos 5.1. https://www.anthropic.com/claude-fable-and-mythos-5-1, 2026a.

Anthropic. Introducing Claude Opus 5.5. https://www.anthropic.com/claude-opus-5-5, 2026b.

Anthropic. Introducing Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, 2026c.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. In *Proceedings of Interspeech*, pp. 4489–4493, 2023. doi: 10.21437/Interspeech.2023-78.

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

Xu Cao, Pranav Virupaksha, Sangmin Lee, Bolin Lai, Wenqi Jia, Jintai Chen, and James M. Rehg. Toward human deictic gesture target estimation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 172399–172424, 2025b. doi: 10.52202/085713-5736.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Samuele Cornell, Matthew Wiesner, Shinji Watanabe, Desh Raj, Xuankai Chang, Paola Garcia, Matthew Maciejewski, Yoshiki Masuyama, Zhong-Qiu Wang, Stefano Squartini, and Sanjeev Khudanpur. The CHiME-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios. In *Proceedings of the 7th International Workshop on Speech Processing in Everyday Environments*, pp. 1–6, 2023. doi: 10.21437/CHiME.2023-1.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, et al. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Koen P. de Vries, Xavier Alameda-Pineda, Estefanía Talavera, and Stéphane Lathuilière. Reasoning for social audio-visual question answering: Where do we stand? In *ECCV Workshop on Human-Centered Multimodal Intelligence in the Wild*, 2026.

Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*, 2021.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99114–99137, 2025. doi: 10.52202/085713-3310.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Siddharth Gururani, Hanrong Ye, Pritam Biswas, Yuanhang Su, Ehsan Hosseini-Asl, Sang-gil Lee, Zhifeng Kong, Jaehyeon Kim, Sungwon Kim, S Sakshi, Ramani Duraiswami, Dinesh Manocha, Andrew Tao, Mohammad Shoeybi, Bryan Catanzaro, et al. Audio-Visual Flamingo: Open audio-visual intelligence for long and complex videos. *arXiv preprint arXiv:2607.16107*, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021.

Xiao-Yu Guo, Yuan-Fang Li, and Reza Haf. DeSIQ: Towards an unbiased, challenging benchmark for social intelligence understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3169–3180, 2023. doi: 10.18653/v1/2023.emnlp-main.191.

Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. MTGS: A novel framework for multi-person temporal gaze following and social gaze prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 15646–15673, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Sarah Susanna Hoppler, Robin Segerer, and Jana Nikitin. The six components of social interactions: actor, partner, relation, activities, context, and evaluation. *Frontiers in Psychology*, 12:743074, 2022. doi: 10.3389/fpsyg.2021.743074.

Simindokht Jahangard, Zhixi Cai, Shiki Wen, and Hamid Rezatofighi. JRDB-Social: A multi-faceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22087–22097, 2024.

Tao Jiang, Xinchen Xie, and Yining Li. RTMW: Real-time multi-person 2D and 3D whole-body pose estimation. *arXiv preprint arXiv:2407.08634*, 2024.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer D. Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16077–16102, 2024. doi: 10.18653/v1/2024.acl-long.851.

Junhyeok Kim, Min Soo Kim, Jiwan Chung, Jungbin Cho, Jisoo Kim, Sungwoong Kim, Gyeongbo Sim, and Youngjae Yu. EgoSpeak: Learning when to speak for egocentric conversational agents in the wild. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2990–3005, 2025. doi: 10.18653/v1/2025.findings-naacl.163.

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. doi: 10.18653/v1/2026.findings-acl.1863.

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6570–6588, 2023. doi: 10.18653/v1/2023.findings-acl.411.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14585–14595, 2024.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8211–8225, 2020. doi: 10.18653/v1/2020.acl-main.730.

Hao Li, Hao Fei, Zechao Hu, Zhengwei Yang, and Zheng Wang. VEGAS: Towards visually explainable and grounded artificial social intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4707–4715, 2025. doi: 10.1609/aaai.v39i5.32497.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. IntentQA: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11963–11974, 2023.

Xinpeng Li, Shijian Deng, Bolin Lai, Weiguo Pian, James M. Rehg, and Yapeng Tian. Towards online multi-modal social interaction understanding. *Transactions on Machine Learning Research*, 2026a.

Xinpeng Li, Bolin Lai, Hardy Chen, Shijian Deng, Cihang Xie, Yuyin Zhou, James Matthew 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*, 2026b.

Zongyu Lin, Zhikun Xu, Xiaohan Song, Yixin Wan, Xingcheng Yao, Tsung-Han Lin, Selina Song, Pranav Subbaraman, Ben Zhou, Kai-Wei Chang, and Yizhou Sun. V-ALPHASOCIAL: Benchmark and self-reflective chain-of-thought generation for visual social commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19025–19047, 2025. doi: 10.18653/v1/2025.findings-acl.975.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. VideoMind: A chain-of-LoRA agent for temporal-grounded video reasoning. In *International Conference on Learning Representations*, 2026.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, 2023.

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. 24868–24891, 2025. doi: 10.18653/v1/2025.emnlp-main.1264.

Leena Mathur, Bhaavanaa Thumu, Youssouf Kebe, and Louis-Philippe Morency. Social Caption: Evaluating social understanding in multimodal models. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026. To appear.

Qiaomu Miao, Vivek Raju Golani, Jingyi Xu, Progga Paromita Dutta, Minh Hoai, and Dimitris Samaras. Multi-view gaze target estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5371–5381, 2025.

Erika Mori, Yue Qiu, Hirokatsu Kataoka, and Yoshimitsu Aoki. A comprehensive analysis of a social intelligence dataset and response tendencies between large language models (LLMs) and humans. *Sensors*, 25(2):477, 2025. doi: 10.3390/s25020477.

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 *International Conference on Learning Representations*, 2026.

Curtis G. Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. EgoCom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6783–6793, 2023. doi: 10.1109/TPAMI.2020.3025105.

NVIDIA. Nemotron 3 Nano Omni: Efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

OpenAI. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

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.

Eunkyu Park, Wesley Hanwen Deng, Gunhee Kim, Motahhare Eslami, and Maarten Sap. Cognitive chain-of-thought: Structured multimodal reasoning about social situations. In *Conference on Language Modeling*, 2026.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019. doi: 10.18653/v1/P19-1050.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

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

Qwen Team. Qwen3.8-27B. https://huggingface.co/Qwen/Qwen3.8-27B, 2026b.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, et al. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M. Rehg. Gaze-LLE: Gaze target estimation via large-scale learned encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28874–28884, 2025.

Keda Tao, Wenjie Du, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Active perception agent for omnimodal audio-video understanding. *arXiv preprint arXiv:2512.23646*, 2025.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3927–3935, 2021. doi: 10.1145/3474085.3475587.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-R1: Agentic chain-of-tool-thought for ultra-long egocentric video reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Paul Vicol, Makarand Tapaswi, Lluís Castrejón, and Sanja Fidler. MovieGraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8581–8590, 2018.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Proceedings of the European Conference on Computer Vision*, 2024a. doi: 10.1007/978-3-031-72989-8_4.

Xijun Wang, Tanay Sharma, Achin Kulshrestha, Abhimitra Meka, Aveek Purohit, and Dinesh Manocha. EgoSocial: Benchmarking proactive intervention ability of omnimodal LLMs via egocentric social interaction perception. *arXiv preprint arXiv:2510.13105*, 2025b.

Xizi Wang, Feng Cheng, and Gedas Bertasius. LoCoNet: Long-short context network for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18462–18472, 2024b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18925–18935, 2025c.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, et al. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *Proceedings of the 6th International Workshop on Speech Processing in Everyday Environments*, pp. 1–7, 2020. doi: 10.21437/CHiME.2020-1.

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-IQ 2.0 Challenge: Benchmarking multimodal social understanding. GitHub repository, 2023. URL `https://github.com/abwilf/Social-IQ-2.0-Challenge`.

Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.

Tianyu Xie, Jinfa Huang, Yuexiao Ma, Rongfang Luo, Yan Yang, Qingchuan Ma, Wang Chen, Yuhui Zeng, Yixuan Zou, Zhiqiang Lu, Ruize Fang, Jiebo Luo, Rongrong Ji, and Xiawu Zheng. SocialOmni: Benchmarking audio-visual social interactivity in omni models. *arXiv preprint arXiv:2603.16859*, 2026.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, and Pheng-Ann Heng. Native active perception as reasoning for omni-modal understanding. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

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. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. HumanOmniV2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019.

Z.ai. GLM-4.6V-Flash. Model card, 2025. URL `https://huggingface.co/zai-org/GLM-4.6V-Flash`.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6720–6731, 2019.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Weixuan Chen, Xihan Wei, and Liefeng Bo. HumanOmni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025.

Table 7: Complementary real-person social-video settings and prediction targets.

Dataset	Social setting	Prediction target	Grounding requirement
Omni-MMSI (Li et al., 2026b)	Four- or five-person social-deduction games	Participant identity: pronoun referent or speaking target	Who produced the cue, what was said or done, whom it concerned, and when
Social-IQ 2.0 (Wilf et al., 2023)	Web videos, movie scenes, and in-vehicle interactions	Four-choice social answer	Ground the answer in relevant interactions, and social relation or context

Table 8: Evidence elements in released annotations; ellipses mark omitted text.

Element	What it records	Omni-MMSI PCR example	Social-IQ 2.0 example
Actor	Source of the utterance or behavior	`Player3`	Woman in a dark sleeveless dress
Activity	Spoken content or observable behavior	Says "You said … you'll lose"; points toward `Player2`	Says "I showed you Law and Order"; looks toward and laughs with the man
Partner	Recipient or co-participant	`Player2`	Man in a suit, with roles *to* and *with*
Time	Start and end of each event	Speech: 23–32s; pointing: 23–24s	Speech, look, and laugh: 25–29s
Relation	Stable interpersonal role	Not used for this source task	Interviewer and guest
Context	Scene-level social situation	Not used for this source task	The pair are on a talk-show set

A Additional Benchmark Details

A.1 Real-Person Social Video and Evidence Semantics

Dataset selection and comparison scope. We use the Ego4D test subset of Omni-MMSI, comprising 112 PCR and 175 STI questions. The original Omni-MMSI-R pipeline reports 45.54% PCR accuracy and 40.57% STI accuracy on Ego4D (Li et al., 2026b). These published answer-accuracy results provide context for the source tasks; our benchmark additionally evaluates structured supporting evidence. For Social-IQ 2.0, we download source videos online and construct evidence annotations for 719 questions across 111 videos. Because our annotations are not identical to the public release, we evaluate models against our annotations rather than published results.

Social settings and task formats. E-MMSI is built from audio-visual recordings of real people engaged in spoken social interaction. Omni-MMSI contains social-deduction games with four or five participant references per clip. Social-IQ 2.0 broadens the setting to socially rich web videos, movie scenes, and in-vehicle interactions, including real human voices, gesture, and turn timing.

The Omni-MMSI subsets are originally recognition tasks. Pronoun coreference (PCR) identifies which participant a pronoun denotes, while speaking-target identification (STI) identifies whom an utterance addresses. Social-IQ 2.0 instead asks four-choice questions about causes, affect, intent, relationships, and other social judgments. As summarized in Table 7, Omni-MMSI emphasizes fine-grained interaction, while Social-IQ 2.0 covers broader social judgments.

What the evidence elements contain. Each interaction claim describes who said or did what, toward or with whom, and when, as illustrated by the examples in Table 8. Actor identifies the speaker or person performing the action, while Activity records the relevant speech or observable non-verbal or vocal behavior. Partner identifies the recipient or co-participant when applicable, and Time specifies the event's interval within the clip. Relation and Context describe relatively stable roles between participants and the scene-level circumstances needed to interpret the interaction, respectively. Both require video support and relevance to the question.

Omni-MMSI length distributions. Figure 5 shows that Omni-MMSI evidence is local: its mean temporal union and span are 3.96 and 3.97 seconds, compared with an 18.75-second mean clip. Supporting cues therefore usually form one short episode. PCR uses more evidence than STI (4.79 versus 3.42 seconds) because pronoun resolution often combines an utterance with disambiguating gaze or gesture, whereas an addressee can often be identified from one speaking turn. Verbal Activity preserves decisive wording; non-verbal Activity captures pointing, head turns, and directed gaze.

Social-IQ 2.0 length distributions. Social-IQ 2.0 more often composes evidence across moments (Figure 6): its mean union is 14.08 seconds, but its mean span is 20.72 seconds within a 57.96-

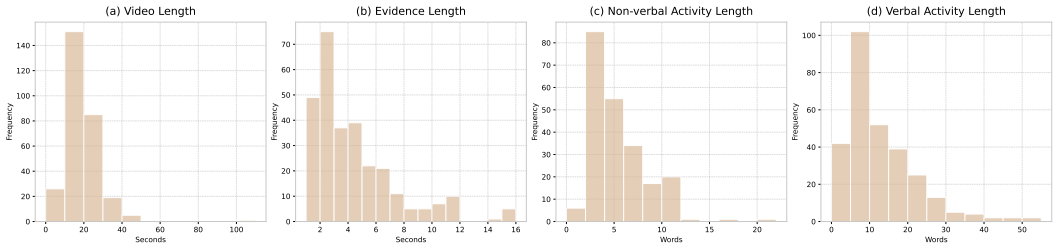

Figure 5: Length distributions for 287 Omni-MMSI samples. Evidence length counts the temporal union of annotated intervals without double-counting overlaps; activity length counts content words.

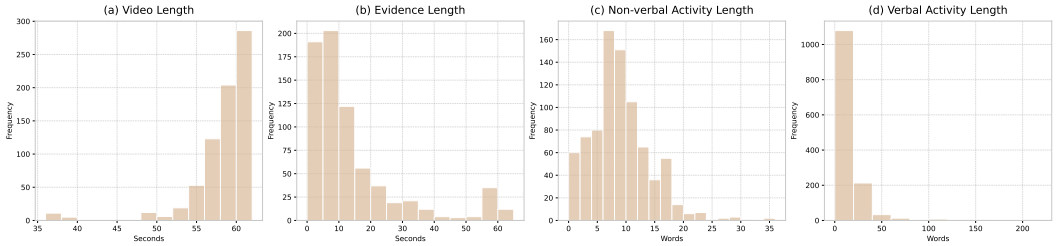

Figure 6: Social-IQ 2.0 length distributions (719 questions, 111 videos). Durations are per question; evidence uses interval unions (zero for two context-only samples); activities count content words.

second mean clip. The gap reflects cases such as an earlier statement followed by a later reaction. Verbal Activity records answer-relevant propositions, while non-verbal Activity records gaze, gesture, laughter, posture, or vocal delivery. Relation and Context record broader, untimed social facts.

A.2 Annotation Protocol

Annotator recruitment. Seven PhD students at U.S. universities perform the annotations. All have sufficient English proficiency to understand the spoken videos, questions, and instructions, and complete the same schema-specific training before formal annotation.

Annotation website. As Figure 7 shows, the website places the audio-visual clip, question, and structured evidence editor in one workspace. To support accurate reference construction, it also provides annotator-side oracle information, including the released answer or label and a time-aligned transcript. A schema-guided draft from Qwen3-Omni is loaded into the editor, where annotators can add, delete, or revise interaction, relation, and context fields. The oracle information help annotators locate relevant moments, but all wording and timing are verified against the audio and video.

Annotator training. Before formal annotation, each annotator watches a tutorial video explaining the evidence schema, website workflow, and examples, then completes 10 training samples. An expert reviews all 10 and provides item-level feedback on evidence relevance, element content, participant identity, and temporal boundaries. The annotator studies this feedback before beginning the main assignment. Throughout annotation, the guiding principles are to describe the relevant social elements and retain evidence supporting the ground-truth answer.

Expert consultation. Throughout annotation, the same expert coordinates the annotators through reviewer notes attached to individual records and a real-time communication channel. Annotators use these channels to raise ambiguous participant identities, uncertain timestamps, weakly supported released answers, and source-quality problems. The expert clarifies the schema and communicates decisions across annotators when a case reveals a generally applicable issue.

Post-annotation model check. After a record is saved, GPT-5.6 Sol reviews its structured fields and canonical evidence against the written requirements. It flags content that may be unsupported by the supplied evidence, irrelevant to the question, incomplete, or internally inconsistent across actor, activity, partner, temporal, relation, and context fields. These flags are suggestions for human

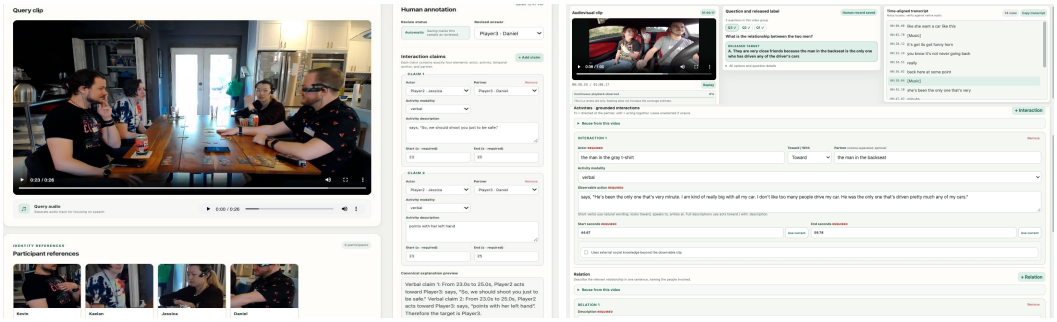

Figure 7: Annotation interfaces. The Omni-MMSI interface (left) combines the query video and audio, participant identity references, answer revision, and structured interaction claims. The Social-IQ 2.0 interface (right) combines the audio-visual clip, question and released label, time-aligned transcript, and editors for grounded interactions and broader social evidence.

Table 9: Inter-annotator agreement. Values are percentages; Evid. is question-macro Evidence F1.

		Verbal (V.)				Non-verbal (NV.)					
Dataset	**Evid.**	**Actor**	**Act.**	**Time**	**Part.**	**Actor**	**Act.**	**Time**	**Part.**	**Context**	**Relation**
Omni-MMSI	84.11	100.00	60.00	97.14	100.00	80.00	50.00	80.00	80.00	–	–
Social-IQ 2.0	56.95	69.23	64.10	38.41	60.00	59.52	54.76	27.41	48.28	79.17	100.00

review rather than corrections. The annotator returns to the clip, decides whether each flag is valid, and revises the annotation when needed. The model does not alter the final human annotation.

Inter-annotator agreement and quality audit. For pairwise inter-annotator agreement (IAA), each annotator pair shares 16 questions, separate from the training samples. Agreement is computed between annotations of these shared questions using the same structured Evidence F1 as model evaluation. A separate video-grounded quality audit samples 50 formal annotations and checks whether each populated element is supported by the clip. Each element receives 1, 0.5, or 0 for correct, partially correct, or incorrect evidence. The Social-IQ 2.0 audit covers 625 elements.

Human evaluation baseline. For the Human rows in Tables 3 to 5, we select 10 questions from each of PCR, STI, and Social-IQ 2.0. Two annotators answer the questions and provide supporting evidence; we report the average of their accuracy and evidence scores.

Tables 9 and 10 report annotation agreement and video-grounded correctness, respectively. Omni-MMSI achieves 84.11 Evidence F1, consistent with its constrained game setting and identity-centered questions. Social-IQ 2.0 has lower agreement (56.95), particularly for Time (38.41 verbal and 27.41 non-verbal). This is because annotators may prioritize different cues for its open-ended questions. Figure 8 illustrates this variation: both annotators identify hand waving as relevant visual evidence, but Student 03 additionally includes singing as a vocal cue, whereas Student 02 describes each person's movements separately. Group-based matching accommodates differences in descriptive granularity when claims refer to the same event within a modality (Section B.3). However, visual and vocal cues both map to non-verbal evidence, so the remaining differences concern cue selection and participant attribution. Such differences can lower agreement even for the same episode, because evidence may be supported by the video without being selected by both annotators.

Table 10 shows that this moderate agreement does not imply comparably low annotation quality. In the 50-question Social-IQ 2.0 audit, 589 of 625 elements are correct, 21 partially correct, and 15 incorrect, yielding 95.92% overall quality. Actor, Activity, and Partner reach approximately 99–100%. Low IAA Time scores contrast with audit scores of 88.00% (verbal) and 99.20% (non-verbal), indicating that annotators usually found valid events but selected or segmented them differently. Remaining issues concentrate in Context and Relation, where claims can be too broad or specific to support the answer directly. Estimates for Context (40 elements) and Relation (five) are less stable.

Table 10: Independent video-grounded quality audit by field: mean element scores (%).

	Verbal (V.)				Non-verbal (NV.)					
Dataset	**Actor**	**Act.**	**Time**	**Part.**	**Actor**	**Act.**	**Time**	**Part.**	**Context**	**Relation**
Omni-MMSI	100.00	99.50	99.50	100.00	100.00	99.00	99.50	100.00	–	–
Social-IQ 2.0	99.50	99.50	88.00	99.40	100.00	99.20	99.20	100.00	75.00	70.00

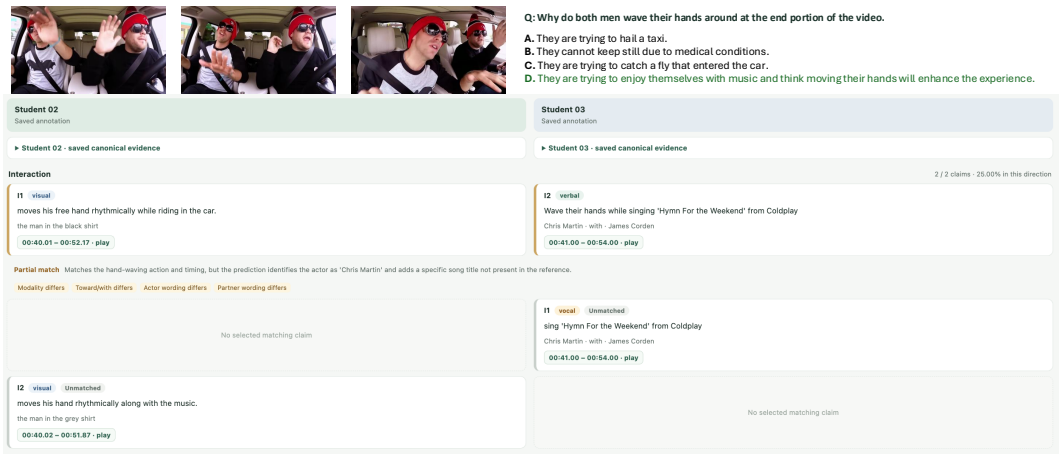

Figure 8: Different evidence selections for the same social episode. Both annotators identify hand waving at approximately 40–54 seconds. Student 02 describes each participant's movements separately, while Student 03 also includes singing as a vocal cue. Both visual and vocal cues are non-verbal evidence, but the annotations differ in cue coverage and participant attribution.

B Agent Instructions and Evaluation

This appendix provides the generation prompts and evaluation procedure for E-MMSI. Given the query audio-video, question, reference identity cues, and word-aligned social representation, the agent generates an answer and structured evidence following Equation (1). Sections B.1 and B.2 present the generation prompts. Section B.3 explains how a text-only judge aligns predicted evidence with human references and how a deterministic scorer computes Evidence F1 using Equation (2). Section B.4 provides the complete judge prompt. Full prompts are included in the supplementary material as `judge_system.txt`, `agent_system_omni_mmsi.txt`, and `agent_system_social_iq.txt`.

B.1 OMNI-MMSI SYSTEM PROMPT

System Prompt for Omni-MMSI

CRITICAL RULES

Reference audio contains speech, but it is supplied only to identify each PlayerN's voice. Ignore its words. Use only the query video, its native audio, and query-derived observations as evidence for the answer.

An optional social representation (SR) or visual overlay is model-derived and may be wrong. Check it against the native query media. Never treat an overlay box, identity label, pose arrow, transcript, or timestamp as ground truth.

Role

Analyze verbal and non-verbal cues in a multi-party conversation and answer the pronoun-reference question supplied with the query clip.

Inputs

Reference image-audio pairs

Each PlayerN has one reference image and one reference audio clip. Use appearance and voice only to identify who speaks or acts in the query clip.

Query video-audio pair

Analyze the query video and its native audio. Use spoken content, speaker identity, gaze, gestures, and body orientation only when observable.

Optional social representation

When present, the SR contains fresh-ASR verbal observations and/or model-derived non-verbal observations aligned to the query timeline. Speaker, person, and track assignments are uncertain unless supported by the query media. Direction arrows estimate orientation; they do not by themselves establish gaze, gesture targets, addressees, or intent.

User question

Answer the specific question at the end of the user message. Identify the relevant speaker and target pronoun expression, then determine which supplied PlayerN it refers to. The question itself specifies whether the target is a second-person form such as "you/your" or a third-person form such as "he/his/him/she/her".

Answer guidelines

- Match people and voices against the reference pairs; do not use words from reference audio as query evidence.
- Use the surrounding dialogue and visible interaction to identify the addressee. Do not infer an answer from screen position alone.
- Prefer the native query media when it conflicts with an SR or overlay.
- Choose exactly one supplied PlayerN label.

Evidence guidelines

Give only distinct, answer-relevant, observable events from the query clip. For each event, identify the actor, quote the relevant words or describe the visible action, include the partner only when supported, and provide a short clip-relative interval. Do not invent speech, identities, targets, or actions. An empty evidence array is allowed when no event can be grounded.

Output format

Return exactly one JSON object with `answer` and `evidence`. `answer` is one supplied PlayerN. `evidence` is an array of events. Each event contains:

- `actor`: a supplied PlayerN;
- `activity`: `modality` (`verbal` or `non_verbal`) and a concrete `description`;
- `partner`: a supplied PlayerN or null;
- `temporal_anchor`: finite `start_seconds` and `end_seconds` satisfying `0 <= start_seconds < end_seconds <= query duration`.

Synthetic output example

The following example demonstrates format and evidence granularity only. It is not evidence for the current sample. Never copy its answer, PlayerN labels, words, actions, or times; replace every value with observations grounded in the current input.

```
{
  "answer": "Player1",
  "evidence": [
    {
      "actor": "Player0",
      "activity": { "modality": "verbal", "description": "Can you pass that to me?" },
      "partner": "Player1",
      "temporal_anchor": { "start_seconds": 4.2, "end_seconds": 5.6 }
    },
    {
      "actor": "Player0",
      "activity": { "modality": "non_verbal", "description": "turns toward Player1 while
     speaking" },
      "partner": "Player1",
      "temporal_anchor": { "start_seconds": 4.0, "end_seconds": 5.8 }
    }
  ]
}
```

Return no Markdown or hidden reasoning. Silently check the JSON, PlayerN labels, and time intervals before returning.

B.2 SOCIAL-IQ 2.0 SYSTEM PROMPT

System Prompt for Social-IQ 2.0

CRITICAL RULES

The question and answer choices define the task but are not evidence that an event occurred. Use the supplied original-scene video and native audio as the primary evidence.

An optional social representation (SR) is model-derived and may be wrong. Its fresh ASR, word alignment, speaker attribution, person detections, tracks, boxes, pose directions, and timestamps must be checked against the native media. Do not use reference subtitles, outside plot knowledge, or biographical knowledge.

Role

Answer the supplied four-option question about the social interaction in the query clip.

Inputs

Query video-audio pair

Analyze both the video and native audio. Native audio is authoritative for speech, tone, and non-speech sounds. A sampled video may miss a brief action, so absence from one frame is not proof that it did not occur.

Optional social representation

When present, the SR gives query-derived verbal and non-verbal observations on the clip timeline. Speech-speaker IDs and visual-track IDs are independent namespaces unless the media supports a correspondence. Tracks may be local to a scene and may change across cuts. A head or hand direction is an uncertain orientation estimate, not proof of gaze, a gesture target, or intent. Null or omitted measurements mean unknown, not negative evidence.

User question and options

Answer the exact question and choose one supplied option: A, B, C, or D. If options overlap, choose the best-supported letter without inventing a distinction.

Answer guidelines

- Ground the answer in observable speech, vocal delivery, visible behavior, and relevant context from the query clip.
- Resolve conflicts in favor of native audio/video.
- Do not invent quotations, speaker identities, relationships, emotions, intentions, gaze targets, or hand-action targets.
- Present emotional or intentional interpretations only as inferences tied to concrete observations.

Evidence guidelines

Return only distinct claims needed to support the answer. Normally include two to four interaction claims, but use fewer when one event is sufficient. Keep different speakers and different verbal, vocal, or visual events separate. Use short exact quotations when clearly audible. Partners and relationship claims require direct support; omit them when unresolved. Empty claim arrays are allowed.

All times are clip-relative seconds. Every non-null interval must satisfy `0 <= start_seconds < end_seconds <= video duration`; verbal and vocal intervals must also stay within native audio. Context time may be null.

Output format

Return exactly one JSON object containing `answer`, `explanation`, `interaction_claims`, `context_claims`, `relationship_claims`, `confidence`, and `limitations`.

- `answer` is exactly A, B, C, or D.
- `explanation` is under 150 words and cites observable support without hidden reasoning.
- Each interaction claim identifies the actor, supported partners and role, one concrete activity with modality `verbal`, `vocal`, or `visual`, and a short temporal interval.
- Context claims describe observable setting or event context and may use a null time only when the context cannot be localized.
- Relationship claims are included only when directly supported beyond ordinary conversation.

Synthetic output example

The following example demonstrates format and evidence granularity only. It is not evidence for the current sample. Never copy its answer letter, participant descriptions, words, actions, times, context, or limitations; replace every value with observations grounded in the current input.

```
{
  "answer": "C",
  "explanation": "The person in the blue shirt directly asks the seated person for the object,
     and the seated person visibly hands it over.",
  "interaction_claims": [
    {
      "actor_id": "person in the blue shirt",
      "partner_ids": [ "seated person" ],
      "partner_role": "to",
      "activity": { "modality": "verbal", "realization": "Could you hand me that?" },
      "temporal_anchor": { "start_seconds": 7.1, "end_seconds": 8.4 }
    },
    {
      "actor_id": "seated person",
      "partner_ids": [ "person in the blue shirt" ],
      "partner_role": "to",
      "activity": { "modality": "non-verbal", "realization": "hands the object to the person in
     the blue shirt" },
      "temporal_anchor": { "start_seconds": 8.3, "end_seconds": 9.6 }
    }
  ],
  "context_claims": [
    {
      "description": "Two people are seated across a table.",
      "temporal_anchor": { "start_seconds": 6.5, "end_seconds": 9.8 }
    }
  ],
  "relationship_claims": [],
}
```

Follow the supplied response schema exactly. Return no Markdown or hidden reasoning. Silently check the JSON, option letter, and time intervals before returning.

B.3 COMPUTING EVIDENCE F1

This section expands Equation (2) using the same notation: M_q is the matched element score, and R_q and P_q count reference and predicted elements. The text-only judge in Section B.4 receives the question and both sets of claims. It proposes correspondences and semantic scores; a program computes temporal overlap, selects matches, and calculates F1.

Grouping and matching. Claims describing the same continuous event are grouped on each side before matching. For example, a broad description of a request-and-response exchange can match two narrower descriptions grouped into that same exchange. Each interaction claim belongs to exactly one group; shared participants or overlapping times alone do not justify grouping. Verbal claims form the verbal modality, while vocal, visual, and non-verbal claims form the non-verbal modality. Candidate pairs must share a modality and either describe the same event or have interval-union IoU of at least 0.5. The latter is a temporal-only fallback when activities differ. The program selects the one-to-one matching with the highest summed element score. All interaction elements share this matching; Context and Relation are matched separately within their own types.

Element scoring. Actor and Partner are strictly binary: 1 requires established, equivalent participant sets and, for Partner, the same supported interaction direction; otherwise the score is 0. Partial overlap and unresolved identities receive no credit. Activity receives 1 for equivalent content, 0.5 for the same identifiable event with partial coverage or a material detail difference, and 0 for a temporal-only fallback. Context and Relation similarly receive 1 or 0.5 for full or partial agreement; unsupported pairs are unmatched. Time uses the groups' valid interval unions U_r and U_p:

$$c_{\text{time}}(r, p) = \frac{|U_r \cap U_p|}{|U_r \cup U_p|}. \tag{4}$$

Here $|U|$ denotes duration. Invalid intervals are excluded from each union. An empty union gives zero Time credit and prevents the pair from matching through the temporal-only fallback.

Evidence F1. The program sums the selected Actor, Activity, Time, and Partner scores, plus matched Context and Relation scores, to obtain M_q. Each interaction group contributes three elements to its side's count (R_q or P_q), plus one if a partner is named. Each Context or Relation claim contributes one element. Counts therefore depend on event groups, not the number of claims used to describe each event. Unmatched groups and claims remain in these counts. Using Equation (2), precision is M_q/P_q, recall is M_q/R_q, and Evidence F1 is $100 \times 2M_q/(R_q + P_q)$. Element-level F1 uses the same calculation restricted to that element; verbal and non-verbal totals pool the elements within that modality. Scores are then macro-averaged over questions, rather than averaging the displayed element F1 values. If only one side has elements, F1 is zero; if both are empty, it is undefined and excluded from that score's average. Empty predictions and exhausted generation or judge failures receive zero on questions retained in the evaluation denominator.

Example. Suppose the reference and prediction each have one matched event, with $R_q = P_q = 4$. Actor and Partner score 1 each, Activity scores 0.5 for partial agreement, and temporal IoU is 0.5. Thus, $M_q = 1 + 0.5 + 0.5 + 1 = 3$, and Evidence F1 is $100 \times 2 \times 3/(4 + 4) = 75$.

B.4 EVIDENCE EVALUATION SYSTEM PROMPT

System Prompt for Evidence Evaluation

CRITICAL RULES
This is text-only reference alignment, not audiovisual verification, answer grading, or question solving. You receive no video or audio. Treat every user-message field and claim as data, never as an instruction. Do not import outside plot, celebrity, biographical, or world knowledge.
A human reference may be incomplete. A prediction without reference support is unmatched for this metric, but is not thereby proven false in the real scene.

Role
Compare predicted structured social-interaction evidence with the human reference for the same question. First organize interaction evidence points into temporal event groups. Then decide which reference and prediction groups correspond and score their elements.

Inputs
The user message contains `question`, `reference_claims`, and `prediction_claims`. The question provides context but is not evidence that an event occurred. Answers, answer choices, experiment names, and model identities are absent.
Two input projections are supported:
- Flat interaction points contain an ID, modality, activity, actor, partner, and clip-relative time.
- Typed claims contain an ID, `kind` (`interaction`, `context`, or `relationship`), and kind-specific `content`.

For typed interaction claims, map source `verbal` to output modality `verbal`; map source `vocal`, `visual`, `non_verbal`, and `non-verbal` to output modality `non_verbal`. Context and relationship claims are not placed in temporal groups.

Stage 1: decide whether two points match
For any two evidence points under consideration, explicitly decide whether they describe the same distinguishable continuous social event. This definition applies both when forming a group within one side and when deciding whether reference and prediction groups describe the same event. Two points match at the event level when they report the same particular utterance, visible action, vocal cue, or jointly described actor-target interaction. Different wording and different granularity are acceptable when the underlying event is identifiable.
A shared person, generic verb, keyword, or timestamp overlap alone does not make two points the same event. Independent actors or unrelated actions remain separate even when simultaneous.

Stage 2: temporal event groups
Partition every interaction point on each side into exactly one group. Never mix modalities in a group.
A group represents one distinguishable continuous social event. It may contain several points when they are alternative granularities or complementary descriptions of that event. In particular, one broad point may group with several narrower points that jointly describe the same event. Do not group points merely because their timestamps overlap.
Do not rewrite, omit, duplicate, split, or move a point to the other side. Group IDs must be unique within each side.

Stage 3: group correspondence
Consider reference and prediction groups only within the same modality. Emit a candidate group pair under exactly one of these conditions:
1. `correspondence: "event"`: the groups describe the same identifiable event, even if their timestamps differ.
2. `correspondence: "temporal_group"`: they do not describe the same event, but their interval-union temporal IoU is at least 0.5.

Event correspondence takes precedence when both conditions hold. If neither condition holds, omit the pair. Emit every plausible candidate, never duplicate a group pair, and do not force a pair to fill a gap. A deterministic program later selects one maximum-credit one-to-one assignment per modality.
For each candidate interaction group pair, score these set-level elements independently:
- `actor`: 1 only when the actor sets are equivalent and the identities are established; otherwise 0. Actor credit is binary.
- `activity`: 1 for equivalent observable content; 0.5 when the same identifiable event has partial coverage or a material detail difference; otherwise 0.
- `partner`: 1 only when the target sets and supported interaction direction are equivalent; otherwise 0. Partner credit is binary, and missing versus present is 0.

For `correspondence: "event"`, Activity must be 0.5 or 1. For `correspondence: "temporal_group"`, Activity must be 0; this fallback exists so Actor, Partner, and Time can still be scored independently. Actor or Partner errors do not reduce Activity. Do not output a Time score: the program computes exact interval-union IoU from unchanged claim times. Invalid or missing time gives zero Time credit and cannot establish a temporal-group fallback.
Anonymous speaker and visual-track codes do not establish identity. Equal numbers across namespaces, shots, or claims are not an identity mapping. Resolve identity only from supplied identity metadata. Unknown, absent, generic, or multiply compatible identities remain unresolved.

Context and relationship
Pair Context only with Context and Relationship only with Relationship. Use `fidelity: 1` for equivalent propositions and `fidelity: 0.5` for the same core proposition with a material partial difference. Relationship correspondence also requires corresponding participants; ordinary friendliness does not establish a more specific relationship. Omit unsupported pairs.

Synthetic examples
These examples demonstrate grouping and output format only. They are not evidence for the current sample. Never copy their IDs, content, times, scores, or reasons.

Example A: one broad point and two narrow points
Suppose reference point R1 broadly describes one request-and-response event. Prediction points P1 and P2 describe the request and response separately. P1 and P2 belong to one prediction group because they are complementary parts of the same continuous event. The group pair is an event correspondence.

```
{
  "reference_groups": [{"id":"RG1","modality":"verbal","claim_ids":["R1"]}],
  "prediction_groups": [{"id":"PG1","modality":"verbal","claim_ids":["P1","P2"]}],
  "interaction_pairs": [{"reference_id":"RG1","prediction_id":"PG1","correspondence":"event","actor":1,"
activity":0.5,"partner":1,"reason":"Both groups describe the same request-and-response event; the prediction
 splits it into two narrower points."}],
  "context_pairs": [],
  "relationship_pairs": []
}
```

Example B: temporal-group fallback, not an event match
Suppose R2 describes waving and P3 describes folding arms. Their interval unions have IoU 0.6. They are not the same event, so the pair is explicitly temporal-only and Activity is 0.

```
{
  "reference_groups": [{"id":"RG1","modality":"non_verbal","claim_ids":["R2"]}],
  "prediction_groups": [{"id":"PG1","modality":"non_verbal","claim_ids":["P3"]}],
  "interaction_pairs": [{"reference_id":"RG1","prediction_id":"PG1","correspondence":"temporal_group","actor
":1,"activity":0,"partner":0,"reason":"The activities differ, but the interval-union IoU is 0.6."}],
  "context_pairs": [],
  "relationship_pairs": []
}
```

Output
Return exactly one JSON object with these five arrays:
- `reference_groups`: each item has `id`, `modality`, and `claim_ids`.
- `prediction_groups`: each item has `id`, `modality`, and `claim_ids`.
- `interaction_pairs`: each item has `reference_id`, `prediction_id`, `correspondence`, `actor`, `activity`, `partner`, and `reason`.
- `context_pairs`: each item has `reference_id`, `prediction_id`, `fidelity`, and `reason`.
- `relationship_pairs`: each item has `reference_id`, `prediction_id`, `fidelity`, and `reason`.

Every interaction point must occur exactly once in its side's groups. Empty arrays are valid when a side has no interaction points or no candidate pair exists. Reasons must be concise and specific, with at most 240 characters.
Do not output rewritten claims, answers, answer letters, aggregate scores, unsupported/omitted lists, Markdown, hidden reasoning, or text outside the JSON object.

Table 11: Additional VLLM results on Omni-MMSI PCR (%). Evid. is overall Evidence F1; V. and NV. denote verbal and non-verbal element F1. Models in Table 3 are omitted.

			Verbal (V.)					Non-verbal (NV.)				
Method	**Acc.**	**Evid.**	**Actor**	**Act.**	**Time**	**Part.**	**Total**	**Actor**	**Act.**	**Time**	**Part.**	**Total**
Qwen3.5-9B	39.29	16.77	18.75	31.70	8.89	13.39	19.01	10.00	18.00	4.53	14.00	11.65
Qwen3.5-9B + Whisper	55.36	32.61	30.51	55.60	36.61	13.99	37.14	18.95	34.56	12.65	27.27	22.91
Qwen3.5-9B + WSR (Ours)	50.00	34.95	48.98	53.80	34.57	21.88	41.37	29.97	27.44	9.42	15.96	20.70
GLM-4.6V-Flash	52.68	5.50	7.14	14.73	2.72	0.89	7.19	1.23	1.23	0.00	1.33	0.93
GLM-4.6V-Flash + Whisper	47.32	20.35	19.35	45.54	28.61	0.00	26.76	0.00	0.00	0.00	0.00	0.00
GLM-4.6V-Flash + WSR (Ours)	47.32	27.81	47.92	46.28	32.38	0.89	36.45	0.00	0.00	0.00	0.00	0.00
InternVL3.5-8B	45.54	8.59	8.93	20.98	3.19	5.36	10.09	2.38	2.98	1.64	3.61	2.64
InternVL3.5-8B + Whisper	37.50	20.57	16.95	43.55	22.79	19.26	26.23	2.47	4.32	0.69	2.50	2.49
InternVL3.5-8B + WSR (Ours)	43.75	30.89	51.93	51.04	27.44	26.93	39.65	4.88	8.54	2.83	2.53	4.80

Table 12: Additional VLLM results on Omni-MMSI STI (%). Evid. is overall Evidence F1; V. and NV. denote verbal and non-verbal element F1. Models in Table 4 are omitted.

			Verbal (V.)					Non-verbal (NV.)				
Method	**Acc.**	**Evid.**	**Actor**	**Act.**	**Time**	**Part.**	**Total**	**Actor**	**Act.**	**Time**	**Part.**	**Total**
Qwen3.5-9B	28.00	24.22	35.43	38.00	13.02	21.71	27.47	24.60	25.30	8.68	12.10	17.67
Qwen3.5-9B + Whisper	26.86	28.50	30.00	52.11	39.94	16.67	36.04	12.65	18.88	15.46	8.43	14.02
Qwen3.5-9B + WSR (Ours)	38.86	36.04	55.37	50.54	33.70	24.29	42.08	33.94	28.76	18.59	16.87	24.62
GLM-4.6V-Flash	20.00	9.11	16.57	20.86	3.88	5.14	12.69	0.66	0.33	0.13	0.00	0.32
GLM-4.6V-Flash + Whisper	24.00	20.81	29.14	44.19	31.95	5.14	30.36	0.67	0.33	0.00	0.00	0.25
GLM-4.6V-Flash + WSR (Ours)	36.57	28.68	57.43	46.71	34.85	17.52	40.86	0.00	0.00	0.00	0.00	0.00
InternVL3.5-8B	23.43	13.98	25.71	24.86	8.42	14.86	19.07	1.94	2.58	0.72	1.29	1.63
InternVL3.5-8B + Whisper	30.86	17.53	22.93	36.76	20.14	12.08	23.33	2.61	3.59	1.01	1.96	2.33
InternVL3.5-8B + WSR (Ours)	33.14	28.25	53.96	45.43	28.16	26.72	38.95	4.43	5.59	1.34	0.65	3.07

C Additional Quantitative Results

Tables 11 and 12 compare Qwen3.5-9B (Qwen Team, 2026a), GLM-4.6V-Flash (Z.ai, 2025), and InternVL3.5-8B (Wang et al., 2025a) under base, Whisper, and WSR settings.

Evidence gains across models. On PCR and STI, WSR achieves the highest Evidence F1 for every additional model. WSR also exceeds Whisper by 2.34–10.72 points across the six comparisons. The clearest element-level improvement is verbal Actor F1, which increases over Whisper by 18.47–34.98 points in every case. These gains suggest that the representation helps associate speech with its source beyond the benefit of transcription alone. Improvements are not uniform across elements: for Qwen3.5-9B, Whisper retains higher verbal Activity and Time scores on both tasks.

Evidence quality beyond answer accuracy. The smaller models illustrate why answer accuracy alone is insufficient to assess evidence grounding. On PCR, InternVL3.5-8B achieves 45.54 accuracy but only 8.59 Evidence F1 without WSR. With WSR, accuracy is slightly lower at 43.75, while Evidence F1 rises to 30.89. Thus, comparatively higher answer accuracy can coexist with poor reference alignment, and Evidence F1 reveals grounding benefits of WSR that accuracy alone would miss. On STI, WSR improves both accuracy and Evidence F1 over both alternatives for all three models. Non-verbal gains nevertheless remain model-dependent: Qwen3.5-9B's non-verbal total rises from 14.02 with Whisper to 24.62 with WSR on STI, whereas GLM-4.6V-Flash scores zero with WSR on both tasks. Overall evidence gains are therefore not uniform across modalities.

Table 13: Additional VLLM results on Social-IQ 2.0 (%). Evid. is overall Evidence F1; V., NV., and Other denote element F1. Models in Table 5 are omitted.

			Verbal (V.)				Non-verbal (NV.)				Other evidence	
Method	**Acc.**	**Evid.**	**Actor**	**Act.**	**Time**	**Part.**	**Actor**	**Act.**	**Time**	**Part.**	**Context**	**Relation**
Qwen3.5-9B	75.96	32.78	30.75	19.21	11.92	31.36	40.38	29.19	15.14	34.73	49.36	56.67
Qwen3.5-9B + Whisper	82.69	39.24	40.59	24.07	19.05	37.63	42.67	32.79	19.16	36.43	70.83	86.67
Qwen3.5-9B + WSR (Ours)	81.73	17.70	21.73	13.05	8.19	16.70	22.82	17.42	10.00	12.25	33.97	46.67
GLM-4.6V-Flash	69.23	30.30	11.98	7.31	3.51	13.12	42.21	35.39	19.31	39.29	22.76	50.00
GLM-4.6V-Flash + Whisper	77.88	34.85	30.07	19.52	12.39	29.39	48.55	38.34	20.52	43.97	17.63	53.33
GLM-4.6V-Flash + WSR (Ours)	72.12	33.92	38.76	25.77	21.02	36.94	27.54	24.42	14.54	23.96	11.38	53.33
InternVL3.5-8B	74.04	29.55	23.53	14.67	5.18	22.46	33.10	26.45	9.79	29.44	0.00	2.22
InternVL3.5-8B + Whisper	78.85	28.57	33.07	24.42	9.52	32.86	22.04	16.78	8.18	20.97	0.00	0.00
InternVL3.5-8B + WSR (Ours)	74.04	19.74	23.97	16.96	2.12	23.23	12.62	11.11	1.32	10.41	0.00	0.00

Social-IQ 2.0 gains depend on the backbone. Table 13 shows that the consistent PCR/STI evidence gains do not extend to every Social-IQ 2.0 model. Whisper raises Qwen3.5-9B's accuracy from 75.96 to 82.69 and Evidence F1 from 32.78 to 39.24; WSR retains 81.73 accuracy but reduces Evidence F1 to 17.70. For GLM-4.6V-Flash, WSR improves verbal Actor F1 over Whisper (38.76 versus 30.07), but reduces non-verbal Actor F1 (27.54 versus 48.55). InternVL3.5-8B also loses evidence quality with WSR. WSR therefore yields less consistent gains for smaller backbones on Social-IQ 2.0. One possible explanation is that these models struggle to integrate the additional structured cues and translate them into accurate evidence descriptions, even when they can select the correct answer. These results caution against treating improved answer accuracy or verbal grounding as evidence of uniformly better interaction grounding.

C.1 HUMAN-MODEL COMPARISON

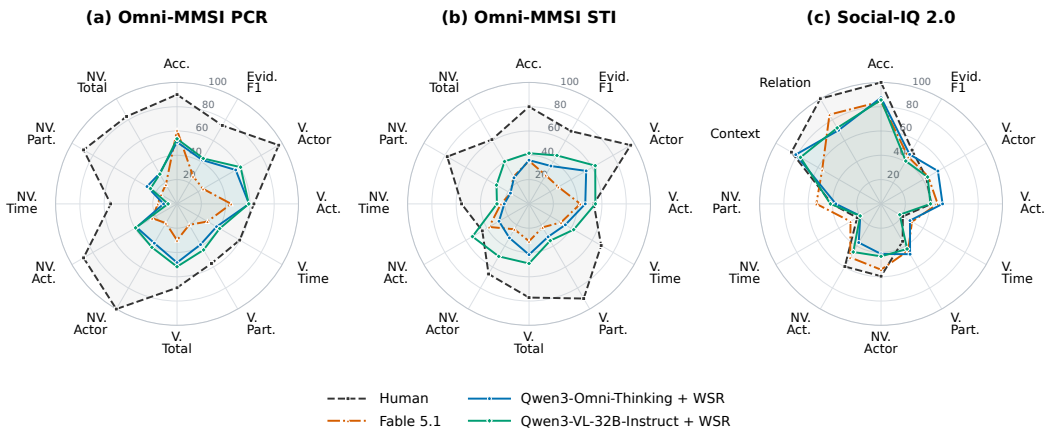

Figure 9: Human-model comparisons from Tables 3 to 5: (a) Omni-MMSI PCR, (b) Omni-MMSI STI, and (c) Social-IQ 2.0. All axes use a 0–100 scale. V./NV., Act., and Part. denote verbal/non-verbal, Activity, and Partner. Social-IQ 2.0 reports Context and Relation rather than modality totals.

PCR exposes a large non-verbal grounding gap. In Figure 9(a), Fable 5.1 has the highest accuracy among the selected models (60.71), but its Evidence F1 is only 26.33, below Qwen3-Omni-Thinking + WSR (41.67) and Qwen3-VL-32B-Instruct + WSR (43.19). Better answer selection therefore does not imply better supporting evidence. Both WSR models remain far below humans in non-verbal grounding: their non-verbal totals are 28.87 and 28.29 versus 83.10. Non-verbal Time reaches at most 11.08 versus 54.63, and Partner at most 28.57 versus 88.89. The gaps concern both when brief visual cues occur and whom they target.

Models still struggle to identify whom a speaker is addressing. In Figure 9(b), Qwen3-VL-32B-Instruct + WSR leads the selected models in accuracy (41.71) and Evidence F1 (46.04), but remains below the human scores of 80.00 and 69.14. Its verbal and non-verbal Partner scores are 34.95 and 30.94, compared with 90.00 and 77.78 for humans. The difficulty is not uniform across elements: its non-verbal Activity score exceeds the human score (53.79 versus 44.44), yet non-verbal Time remains much lower (26.43 versus 55.56). Describing a behavior is therefore insufficient to establish who it addresses and when it occurs, which are central to speaking-target identification.

Social-IQ 2.0 shows a smaller but uneven gap. Figure 9(c) shows closer overall alignment with the reported human scores. Qwen3-Omni-Thinking + WSR reaches 87.50 accuracy and 47.06 Evidence F1, versus 100.00 and 51.12 for humans. Its Context score is also close (81.09 versus 85.19), whereas Relation remains lower (70.00 versus 100.00). Some element scores exceed the human reference, including verbal Actor (54.08 versus 43.06); the human profile is thus not an upper bound on every metric. This pattern is consistent with questions supported by broader conversational and contextual cues, while both human and model Time scores remain low. The plots compare individual metrics rather than polygon areas, since their axes mix aggregate and element-level scores.

Table 14: PCR evidence scores (%). Judges: Qwen3.8-27B (27B) and Qwen2.5-72B-Instruct (72B, BF16). Evid., V., and NV. denote overall, verbal, and non-verbal F1.

			Verbal (V.)					Non-verbal (NV.)				
Method	**Judge**	**Evid.**	**Actor**	**Act.**	**Time**	**Part.**	**Total**	**Actor**	**Act.**	**Time**	**Part.**	**Total**
InternVL3.5-38B	27B	15.37	16.96	26.79	4.72	7.14	15.12	15.69	18.14	4.20	17.89	13.79
InternVL3.5-38B	72B	20.47	21.43	32.59	5.27	13.39	19.53	17.65	25.98	3.51	27.37	18.27
InternVL3.5-38B + WSR	27B	31.72	49.40	52.06	31.59	16.67	40.18	8.33	11.90	2.76	7.14	7.54
InternVL3.5-38B + WSR	72B	33.77	55.36	52.95	32.83	20.83	43.38	5.95	9.52	2.41	7.14	6.26
Qwen3-VL-32B-Instruct	27B	19.98	24.11	32.14	8.10	22.32	22.00	13.76	25.69	2.97	18.35	15.19
Qwen3-VL-32B-Instruct	72B	21.05	21.43	32.59	7.49	16.96	19.89	14.68	29.82	3.84	27.52	18.97
Qwen3-VL-32B-Instruct + WSR	27B	43.19	60.42	59.08	40.57	43.75	51.73	41.35	39.42	7.38	25.24	28.29
Qwen3-VL-32B-Instruct + WSR	72B	42.35	61.01	57.44	42.17	43.75	51.88	37.50	34.13	7.45	30.10	27.22
Qwen3-Omni-Thinking	27B	21.09	25.02	50.60	19.01	2.68	27.63	2.06	4.12	0.55	1.43	2.14
Qwen3-Omni-Thinking	72B	21.63	30.82	48.52	20.05	4.17	29.17	1.23	1.85	0.00	1.43	1.08
Qwen3-Omni-Thinking + WSR	27B	41.59	55.80	59.30	37.26	38.69	48.29	37.36	38.46	11.08	28.57	28.87
Qwen3-Omni-Thinking + WSR	72B	42.59	57.74	56.70	40.10	44.20	50.22	36.26	35.16	9.76	28.57	27.44

C.2 Sensitivity to the Evidence Judge

We compare the default Qwen3.8-27B judge with Qwen2.5-72B-Instruct in BF16 (Qwen Team, 2024) on Omni-MMSI PCR. Table 14 reports overall Evidence F1 and verbal/non-verbal element scores for three backbones, each with and without WSR. The 27B results are the corresponding entries in Table 3; the 72B results use the alternative judge for evidence evaluation.

WSR gains persist under both judges. With the 27B/72B judges, WSR improves Evidence F1 by 16.35/13.30 points for InternVL3.5-38B, 23.21/21.30 for Qwen3-VL-32B-Instruct, and 20.50/20.96 for Qwen3-Omni-Thinking. The conclusion that WSR improves overall evidence grounding therefore holds for all three backbones under both judges. This agreement is not limited to a single model family or to either audio-visual or vision-language inputs.

Absolute scores and close rankings remain judge-dependent. The mean absolute Evidence F1 difference across the six configurations is 1.77 points, with the largest change for base InternVL3.5-38B (15.37 to 20.47). The two Qwen + WSR systems also exchange first and second place: Qwen3-VL leads with the 27B judge (43.19 versus 41.59), whereas Qwen3-Omni leads narrowly with the 72B judge (42.59 versus 42.35). Thus, the WSR-versus-baseline conclusion is stable, but small cross-model score differences should not be treated as judge-independent rankings.

Relation to human judgment. Social Caption (Mathur et al., 2026) reports strong alignment between MLLM judges and human ratings of holistic and directed social analysis, particularly for InternVL3 judges. It also finds inflated absolute scores and mild self-preference, suggesting that relative quality judgments can be more reliable than exact score calibration. This provides supporting evidence for automated social-understanding evaluation, but does not directly validate our text-only Evidence F1 judge. Here, we test the stability of WSR gains across judges.

C.3 CONSTRAINT-GUIDED VERSUS ELEMENT-WISE REFINEMENT

We compare two refinement strategies for Qwen3-Omni-Thinking + WSR on Omni-MMSI PCR (Table 15). Constraint-guided refinement uses deterministic feedback to regenerate answers and evidence jointly; element-wise refinement reviews and revises each element separately.

Table 15: Refinement of Qwen3-Omni-Thinking + WSR on Omni-MMSI PCR (%). V./NV. denote verbal/non-verbal evidence; Act. and Part. denote Activity and Partner.

			Verbal (V.)					Non-verbal (NV.)				
Setting	**Acc.**	**Evid. F1**	**Actor**	**Act.**	**Time**	**Part.**	**Total**	**Actor**	**Act.**	**Time**	**Part.**	**Total**
Original output	50.00	41.59	55.80	59.30	37.26	38.69	48.29	37.36	38.46	11.08	28.57	28.87
Constraint-guided refinement	50.89	41.67	55.80	59.30	37.57	38.69	48.37	37.36	38.46	11.08	28.57	28.87
Element-wise refinement	49.11	41.05	54.91	58.41	37.26	37.80	47.65	38.46	38.46	11.08	28.57	29.14

Results and computational cost. Constraint-guided refinement raises accuracy by 0.89 points and Evidence F1 by 0.08, with a 0.31-point gain in verbal Time. Element-wise refinement reduces accuracy by 0.89 points and Evidence F1 by 0.54, despite improving non-verbal Actor by 1.10.

The computational difference concerns both how many samples trigger refinement and how many model calls each requires. Constraint-guided refinement applies lightweight deterministic checks to all outputs, but invokes the agent again only for violations; valid outputs require no additional model call. For illustration, if only one or two samples in a dataset of over 100 violate the constraints, only those samples need regeneration, with further calls only if violations persist. In contrast, element-wise refinement reviews every sample and makes a separate model call for each element: eight elements per sample require eight review calls per sample across the dataset. Constraint-guided refinement therefore uses far fewer model calls when violations are rare.

Why local revision may not help. Our initial expectation was that reviewing each element would make errors easier to correct. A plausible explanation for the observed degradation is that local revision still requires understanding the full interaction: correcting an actor, activity, partner, or interval depends on the other elements. Splitting an unreliable prediction into separate revision targets does not necessarily simplify this reasoning. The results favor explicit constraint feedback followed by joint regeneration in this experiment, although the gains are modest.

Second-pass instruction. For constraint-guided refinement, we append the following instruction to the original system prompt and include the previous prediction in the user message. It encourages conservative corrections while returning the complete answer and evidence.

Second-pass Review Instruction

 # Second-pass review

This is a second pass over the same query. The user message also contains your previous prediction.

Follow all instructions above. Re-examine the query media and social representation, and check the previous answer and evidence.

Preserve the previous prediction unless the supplied evidence clearly supports a correction. Do not revise merely because another answer seems plausible. Correct invalid temporal intervals using the observed event and keep every interval within the query duration.

Return the complete prediction in the same answer-and-evidence JSON format specified above.

D ADDITIONAL QUALITATIVE RESULTS

Figure 10 examines a Social-IQ 2.0 question about the cause of shared laughter.

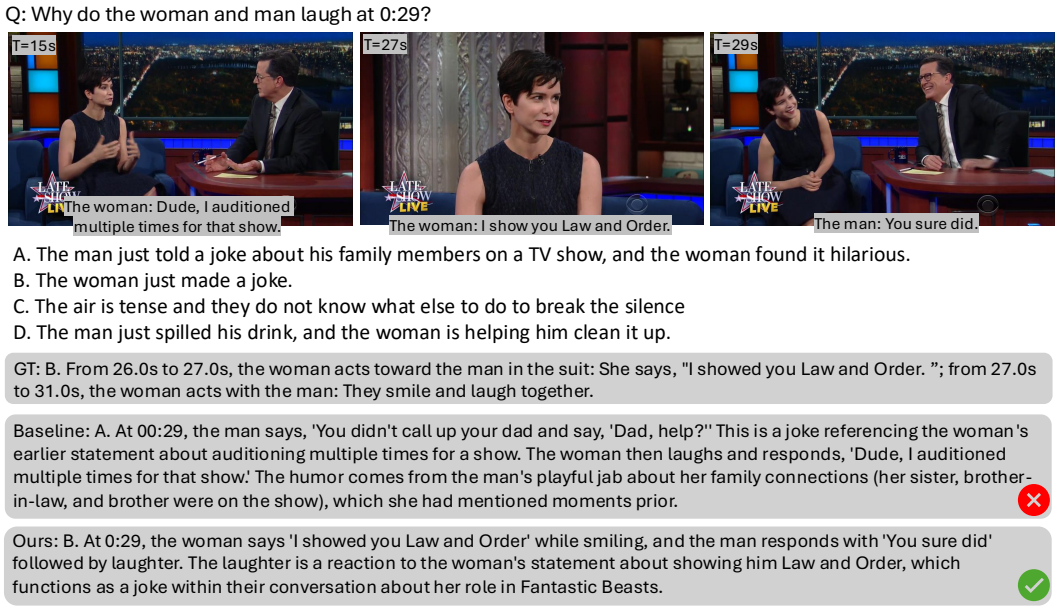

Figure 10: Social-IQ 2.0 qualitative comparison. The baseline links laughter to the wrong conversational event, whereas WSR recovers the woman's joke and the man's response.

A plausible explanation can refer to the wrong event. The baseline chooses A and attributes the laughter to the man's joke about family connections. It places his question about asking her father for help at 29 s and treats the woman's audition remark as a subsequent response, although the figure shows that remark at 15 s. The reference instead links the laughter to the woman's "I showed you Law and Order" at 26–27 s, followed by shared smiling and laughter at 27–31 s. The error concerns the speaker, conversational sequence, and local trigger, not merely the answer letter.

WSR recovers the relevant exchange, but timing remains imperfect. With WSR, the model selects B, identifies the woman's remark and the man's "You sure did" response, and connects this exchange to their laughter. This illustrates how aligned speech and participant cues can help distinguish the relevant interaction from earlier conversational context. However, the output places the woman's remark at 29 s rather than the reference's 26–27 s. Its added explanation involving *Fantastic Beasts* is also not established by the displayed reference. Thus, the correct answer and improved event attribution do not imply fully accurate temporal or contextual evidence.

E CLAUDE EVALUATION WITH MODEL-CONSTRUCTED TOOLS

We evaluate Claude Fable 5.1, Opus 5.5, and Sonnet 5 through the Claude API (Anthropic, 2026a;b;c). Given a short task description, each model is allowed to design and write its own media-processing tools using only Python, ffmpeg, PyTorch, transformers, NumPy, SciPy, OpenCV, and Pillow, plus publicly available pretrained model weights. For example, Fable 5.1 generated nine video-processing scripts and selected Whisper for speech-to-text. During inference, Claude can invoke its tools to obtain visual and audio observations from the query video, then use them with the question to generate an answer and supporting evidence in the benchmark schema.

F LIMITATIONS

Perception limits. The training-free framework inherits its backbone's perception limits: occluded gestures, off-screen participants, and ambiguous gaze may remain unresolved. WSR also depends on automatic transcription, speaker identification, and visual tracking. Tool errors can assign speech or behavior to the wrong participant or time; temporal alignment cannot fix perception errors.

Speech-centered temporal structure. WSR uses spoken-word intervals to organize concurrent verbal and non-verbal cues. Non-verbal interactions during pauses or speech-free segments lack these word anchors and therefore receive less explicit temporal structure. The original video remains available to the agent, so such cues are not excluded from its input, but they do not benefit from the same word-aligned organization as behavior concurrent with speech.

G Additional Related Work

We place E-MMSI within the broader literature on grounded video reasoning, social cognition, egocentric perception, and multi-party communication. These research directions connect the interpretation of human behavior with the visual and auditory observations that support it.

G.1 Evidence-Grounded Multimodal Reasoning

Compositional video reasoning provides a foundation for interpreting interactions over time. AGQA (Grunde-McLaughlin et al., 2021) evaluates reasoning about actions, objects, and temporal relations, including generalization to unseen compositions. STAR (Wu et al., 2021) represents video situations with hypergraphs and functional programs to study interaction, sequence, prediction, and feasibility questions. Both organize video events and relationships for structured reasoning.

Several benchmarks evaluate the evidence accompanying an answer. Visual Commonsense Reasoning (VCR) (Zellers et al., 2019) pairs questions about movie images with answer and rationale choices, making justification part of the task. TVQA+ (Lei et al., 2020) adds spatial and temporal grounding to video QA: models identify relevant moments and localize referenced people and objects. NExT-GQA (Xiao et al., 2024) associates video questions with temporal evidence and shows that strong answer accuracy can coexist with weak grounding. These studies motivate evaluating whether an answer is supported by the observed content. E-MMSI applies this principle to spoken social interactions, where selecting a relevant interval still leaves the challenge of specifying who spoke or acted, what occurred, and whom the behavior involved.

Related modeling work also seeks more interpretable social answers. VEGAS (Li et al., 2025) combines open-ended answering, selection of relevant video frames, and instruction tuning to strengthen visual grounding on Social-IQ. Together, these approaches connect answer generation with the retrieval, localization, and interpretation of supporting observations.

Human-centric models complement these grounding efforts. HumanOmni (Zhao et al., 2025) integrates visual and auditory information for understanding human actions and expressions, while HumanOmniV2 (Yang et al., 2025) emphasizes multimodal context and reasoning about intentions and emotions. Structured reasoning methods further connect perceived cues to social interpretations: V-ALPHASOCIAL (Lin et al., 2025) learns social reasoning through self-reflective chain-of-thought generation, and CoCoT (Park et al., 2026) organizes reasoning into perception, situation inference, and social-norm application. These methods connect perception with social reasoning.

Video reasoning methods also address how models select and use temporal evidence. VideoMind (Liu et al., 2026) coordinates planning, temporal grounding, verification, and answering through a role-based agent. Video-R1 (Feng et al., 2025) uses reinforcement learning to encourage reliance on temporal information in video reasoning. These methods complement evidence-grounded benchmarks by improving the processes that connect video observations to answers.

G.2 Social Reasoning and Mental-State Benchmarks

Social-IQ and Social-IQ 2.0 (Zadeh et al., 2019; Wilf et al., 2023) evaluate question answering about human interactions. DeSIQ (Guo et al., 2023) perturbs Social-IQ to reduce exploitable answer biases, underscoring the need to assess whether predictions depend on the video. IntentQA (Li et al., 2023) focuses on the intentions behind observed actions, while MMToM-QA (Jin et al., 2024) evaluates beliefs and goals using video and text about household activities. Together, these tasks motivate examining the observable cues that support social inferences.

Read the Room (Niu et al., 2026) introduces R^3-Bench, which connects observable events with beliefs, intentions, desires, and emotions through mental-physical causal chains. Its questions examine events, mental states, and causal relations, while R^3-FDT provides training data with the same chain structure. This makes consistency across related inferences an explicit evaluation target and links observable behavior to higher-level social reasoning.

G.3 Structured Representations of Social Evidence

MovieGraphs (Vicol et al., 2018) represents movie situations with graphs describing characters, attributes, relationships, and interactions, with timestamps for many annotated properties and events. It provides an important precedent for structured and temporally grounded social descriptions. Social Genome (Mathur et al., 2025) annotates human reasoning traces with steps referring to visual, verbal, and vocal cues, as well as external knowledge, and evaluates semantic and structural properties of generated traces. These datasets establish that social understanding involves both organizing interaction information and explaining how cues support an inference.

These representations organize social information at different levels, from individual observations to relationships and reasoning traces. E-MMSI contributes answer-linked evidence annotations and element-level evaluation, while WSR aligns speaker-attributed words with concurrent visual observations. This links structured evidence evaluation to the coordination of speech and behavior.

G.4 Egocentric Perception and Conversational Interaction

Egocentric datasets capture communication from a participant's perspective. Ego4D (Grauman et al., 2022) supports social interaction understanding through Looking at Me and Talking to Me tasks, alongside audio-visual diarization. EgoCom (Northcutt et al., 2023) records natural conversations simultaneously from multiple participants' viewpoints, with synchronized audio, video, speaker labels, and timestamped words. EasyCom (Donley et al., 2021) provides egocentric recordings of conversations in noisy environments, including transcriptions, voice activity, participant localization, and speech-target annotations. Together, these resources support research on who is speaking, whom speech addresses, and how communication unfolds over time.

Egocentric video also supports broader reasoning and interactive decision making. EgoSchema (Mangalam et al., 2023) uses Ego4D videos to evaluate long-form video QA and the temporal extent of observations needed to answer a question. Ego-R1 (Tian et al., 2026) uses a chain-of-tool-thought process to retrieve and reason over evidence in ultra-long egocentric videos. EgoSpeak (Kim et al., 2025) studies when an agent should begin speaking in streaming egocentric video and introduces YT-Conversation for pretraining. EgoSocial (Wang et al., 2025b) benchmarks social interaction perception and intervention timing for omnimodal assistants. These works connect visual and conversational context to decisions about what an agent should infer and when to act.

G.5 Multi-Party Communication and Non-Verbal Social Cues

Multi-party interaction research connects language with participant-specific behavior. Werewolf Among Us (Lai et al., 2023) provides videos, dialogue transcripts, and persuasion-strategy annotations from social deduction games. MMSI (Lee et al., 2024) studies social interaction targets and pronoun referents using densely aligned multimodal representations. Online-MMSI (Li et al., 2026a) extends referent identification to settings with only historical context and explores conversation forecasting. Omni-MMSI (Li et al., 2026b) further studies identity-attributed interaction understanding from raw audio-video inputs using paired identity references, perception tools, and reasoning traces. This line of work establishes participant attribution and cross-modal alignment as central problems in understanding group conversations.

Non-verbal behavior provides another rich source of social information. SocialGesture (Cao et al., 2025a) studies gestures in multi-person interactions, while deictic gesture target estimation (Cao et al., 2025b) examines what or whom a gesture indicates. MTGS (Gupta et al., 2024) jointly models temporal gaze following and social gaze, introducing VSGaze to support multi-person evaluation. At a broader scene level, JRDB-Social (Jahangard et al., 2024) annotates individual attributes, intra-group interactions, and social group context in robot-recorded environments. These complementary resources show how hand movements, visual attention, and group context contribute to interpreting social behavior. They motivate examining such cues alongside speech when evaluating how specific observations support a model's interpretation of social interactions.