Omni-MMSI: Toward Identity-attributed Social Interaction Understanding

Supplementary Material

Xinpeng Li[1] Bolin Lai[2] Hardy Chen[3] Shijian Deng[1]
Cihang Xie[3] Yuyin Zhou[3] James M. Rehg[4] Yapeng Tian[1]
[1]University of Texas at Dallas [2]Georgia Institute of Technology
[3]University of California, Santa Cruz [4]University of Illinois Urbana-Champaign
{xinpeng.li, shijian.deng, yapeng.tian}@utdallas.edu
bolin.lai@gatech.edu {hchen403, cixie, yzhou284}@ucsc.edu jrehg@illinois.edu

A. Implementations

A.1. Human Study

We randomly selected 30 samples for each participant and first presented the raw video input (Omni-MMSI setting), followed by the version with provided social cues used in the previous setting, as shown in Fig. 1. This ordering prevents participants from being biased by the provided cues.

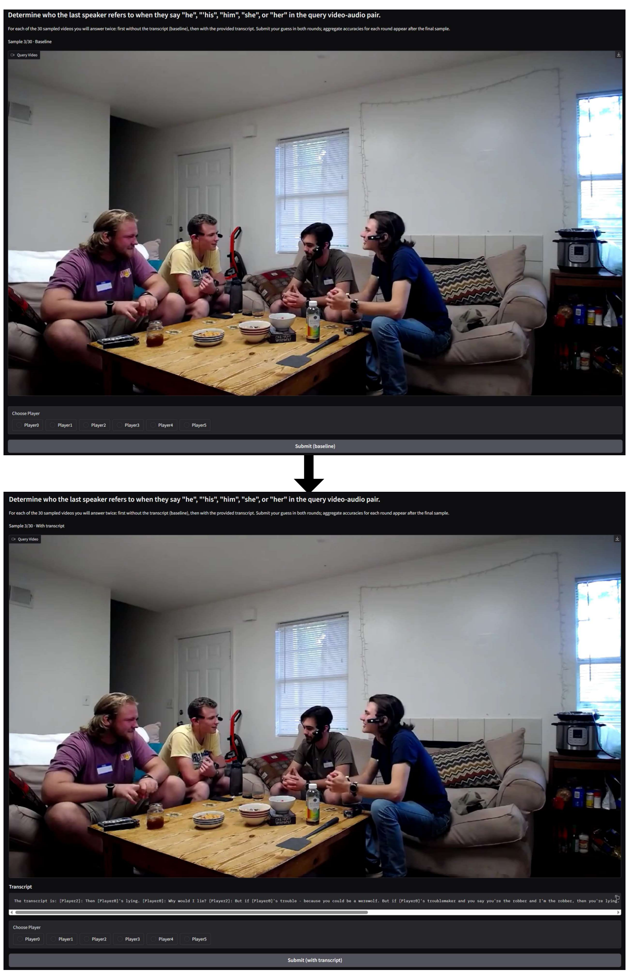

Figure 1. Illustration of human study.

A.2. Human Filtering of CoT Reasoning

As shown in Fig. 2, we examine all reasoning traces that pass the automatic answer-matching step. If a trace contains pervasive errors that fundamentally contradict the audio-visual evidence, we discard it entirely. When only a small number of inaccuracies appear, we manually correct them rather than removing the whole trace. Typical corrections include: a) removing incorrect non-verbal cues, for example, deleting statements such as Player3 looking at Player1 when such gaze does not occur; b) supplementing missing salient evidence, such as adding pointing gestures from the speaker when they serve as a clearer cue than gaze; and c) adding additional non-verbal cues from other participants, for instance, when multiple players are pointing toward the referent but the generated reasoning mentions only the speaker. This process ensures that the final reasoning traces are factually accurate, complete, and faithful.

A.3. System Prompts

To generate CoT annotations from the reference-based input, we use the system prompt shown in Figure 5. This prompt is carefully designed to explicitly instruct the model to identify verbal and non-verbal cues, perform last speaker confirmation and infer the correct referent in a structured step-by-step manner. Its detailed formulation helps the model focus on extracting evidence grounded in the audio-visual input and prevents it from hallucinating unsupported cues. Derived from the CoT-generation prompt, we adopt the system prompt in Figure 6 for model fine-tuning, the prompt in Figure 7 for evaluating Omni-LLMs without references, and the prompt in Figure 8 for evaluating Omni-LLMs with references. Overall, these system prompts are not generic instructions; they are deliberately designed and empirically refined to guide the model toward faithful evidence-based reasoning and maximize the effectiveness of reference-based social interaction understanding.

A.4. System Latency and Parameters

We report the latency and parameter size of each component in the Omni-MMSI-R pipeline for completeness. All measurements are obtained on an NVIDIA RTX A6000 GPU.

Table 1. Latency and parameter size of the components in the Omni-MMSI-R pipeline, measured on an NVIDIA RTX A6000 GPU.

	YOLO + OSNet	Whisper + SpeechBrain	Qwen2.5 Omni (Answer)	Qwen2.5 Omni (CoT)
Latency	0.16s	0.21 real-time factor	1.05s	12.69s
Parameters	43.69M + 2.17M	1541.57M + 22.15M	8.93B	8.93B

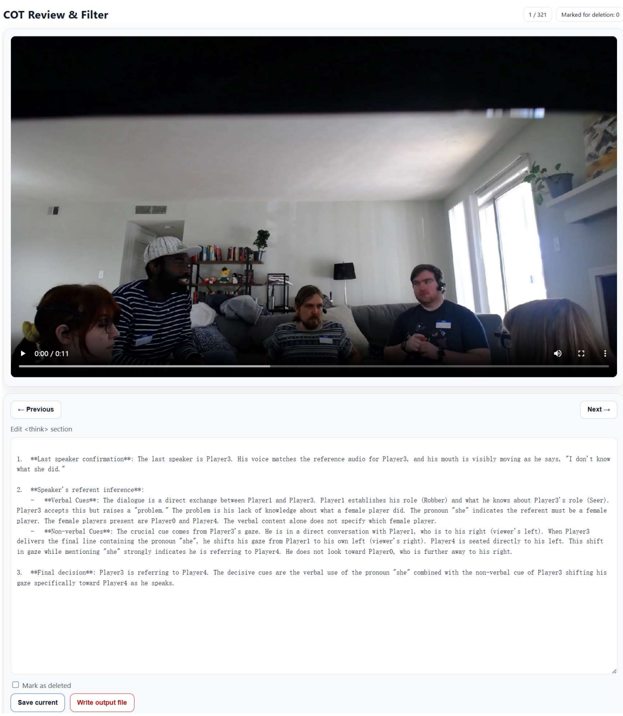

Figure 2. Illustration of human filtering.

For identity-attributed non-verbal cue extraction, YOLO and OSNet together require 0.16s per clip, with 43.69M and 2.17M parameters, respectively. For identity-attributed verbal cue extraction, Whisper and SpeechBrain jointly operate at a 0.21 real-time factor and contain 1541.57M and 22.15M parameters. For the reasoning module, Qwen2.5-Omni (8.93B parameters) produces a direct answer for Omni-MMSI in 1.05s, while enabling chain-of-thought reasoning increases the latency to 12.69s. These numbers characterize the computational profile of the current implementation and serve as a reference for future optimization.

A.5. Identity Attribution Accuracy Computation

To compute verbal identity attribution, we first perform sentence-level matching between the predicted utterances $\hat{u}_i$ and the ground-truth utterances u_i using a semantic similarity score. A predicted utterance is considered matched when its similarity exceeds a threshold τ_{sem}=0.9, forming the matched index set $\mathcal{M}_{\text{verb}} = \{\, i \mid \mathrm{sim}(\hat{u}_i, u_i) > \tau_{\text{sem}} \,\}$, where $\mathrm{sim}(\cdot)$ denotes the cosine similarity between sentence embeddings. The accuracy is then computed:

$$\mathrm{Acc}_{\text{verb}} = \frac{1}{|\mathcal{M}_{\text{verb}}|} \sum_{i \in \mathcal{M}_{\text{verb}}} \mathbb{1}\left[\hat{s}_i = s_i\right], \tag{1}$$

where $\hat{s}_i$ and s_i represent the predicted and ground-truth speaker identities in the matched pairs, respectively.

For non-verbal identity attribution, we first perform IoU-based matching between the predicted person boxes $\hat{b}_i$ and the ground-truth boxes b_i on the last frame. A predicted box is considered matched when its intersection-over-union (IoU) exceeds a threshold τ_{IoU}=0.9, forming the matched index set $\mathcal{M}_{\text{non-verb}} = \{\, i \mid \mathrm{IoU}(\hat{b}_i, b_i) > \tau_{\text{IoU}} \,\}$. The non-verbal attribution accuracy is then computed:

$$\mathrm{Acc}_{\text{non-verb}} = \frac{1}{|\mathcal{M}_{\text{non-verb}}|} \sum_{i \in \mathcal{M}_{\text{non-verb}}} \mathbb{1}\left[\hat{y}_i = y_i\right], \tag{2}$$

where $\hat{y}_i$ and y_i denote the predicted and ground-truth visual identities of the participants, respectively.

A.6. Task Selection

We omit Mentioned Player Prediction (MPP) used in prior MMSI [1, 2]. In the original MMSI formulation, MPP aims to predict the identity referred to by an explicitly mentioned name in a dialogue. The task is constructed by masking a player name (e.g., replacing it with a [MASK] token) and requiring the model to recover the mentioned identity. However, this task is less realistic in practice: AI assistants can typically retrieve explicit names directly from ASR, requiring little social reasoning. Instead, other tasks, STI and PCR, require deeper multimodal cue grounding and social interaction inference. For this reason, we omit MPP in Omni-MMSI and focus on STI and PCR. Since the models are trained and evaluated independently for each task, this omission does not affect comparability with prior works.

A.7. Task Novelty

Omni-MMSI is fundamentally different from prior MMSI formulations [1, 2]. (1) The task assumptions differ. Prior MMSI assumes that identity-attributed social cues are perfectly available, typically via manual annotation or oracle preprocessing. In contrast, Omni-MMSI requires models to automatically extract identity-attributed cues directly from raw inputs. (2) The input modality differs. Previous formulations primarily take visual and textual social cues as input, whereas Omni-MMSI operates on raw multi-modal inputs, including visual, text, and audio signals from videos. Notably, audio is essential for modeling social dynamics such as speaker turns, interruptions, and overlapping speech, which are not supported in prior problems.

A.8. Reference Reliance

When reference information is not pre-stored, the reference bank can be updated automatically. For example, when a person is encountered, the system extracts visual or vocal identity cues, matches them against existing references, and registers a new identity if similarity falls below a threshold. This can be achieved, for example, through a brief greeting-based enrollment step in social scenarios. When references are difficult to obtain (*e.g.*, missing visual), the system degrades to a non-reference mode using raw inputs.

B. More Results

B.1. Robustness of Reference Pairs

This experiment aims to evaluate the robustness of our reference-based pipeline under audio and visual degradation conditions (on Ego4D), focusing on how noise and occlusion affect verbal and non-verbal attribution accuracy and downstream social interaction understanding tasks.

For audio degradation, we inject additive white Gaussian noise into the reference audio at signal-to-noise ratio (SNR) levels of $\{$Clean, 20, 10, 5$\}$ dB, where the Clean setting corresponds to no noise injection. For visual degradation, random occlusion masks are applied to the reference images with occlusion ratios of $\{0.0, 0.1, 0.3, 0.4\}$. The degraded references are used during both the attribution and reasoning stages to assess their overall influence.

As shown in Figure 3, decreasing the SNR from 20 to 5 only slightly decreases verbal attribution accuracy from 71.0% to 70.2%, with negligible changes in STI and PCR performance (less than 0.5%). This indicates that the audio branch of our reference-based framework is highly robust to moderate background noise. In contrast, visual degradation results in a moderate performance drop: as the occlusion ratio increases from 0.0 to 0.4 (severe occlusion), non-verbal attribution accuracy decreases from 86.4% to 72.2%. Nevertheless, the model maintains stable performance on downstream STI and PCR tasks, showing only marginal variations (around 1%), demonstrating that high-level social interaction understanding remains robust even under severe visual occlusion. The results indicate that our pipeline remains stable under audio-vision degradation.

B.2. Cross-Architecture Generalization

To further assess robustness of the proposal across scales, we add experiments on Qwen2.5 Omni 3B [3] across two datasets and two tasks. From Tab. 2, we observe consistent performance improvements after SFT, both with and without reference inputs. More importantly, incorporating expert tools and CoT reasoning further improves performance.

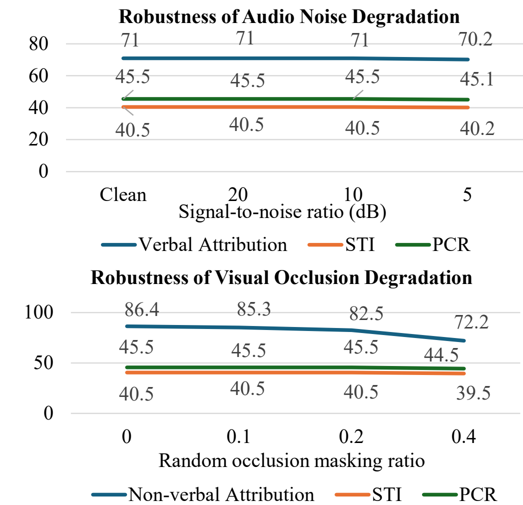

Figure 3. Robustness of the reference-based pipeline under audio and visual degradation. (a) Audio noise degradation evaluates the impact of Gaussian noise on verbal attribution and MMSI tasks. (b) Visual occlusion degradation tests the effect of partial masking on non-verbal attribution and MMSI tasks. The results indicate that our pipeline is highly resilient under audio-visual degradation.

Table 2. Comparison of different settings on Ego4D and YouTube (%). The Omni-LLM backbone is Qwen2.5 Omni 3B. *ZS*, *SFT*, *ref*, *tool* and *CoT* denote zero-shot inference, supervised fine-tuning, the use of raw reference pairs, tools for extracting identity-attributed social cues and chain-of-thought reasoning supervision.

Setting (Ego4D)	STI	PCR	Avg. Acc.
ZS (w/o ref)	21.14	13.61	17.38
SFT (w/o ref)	26.29	32.14	29.22
ZS (w/ ref)	21.23	10.71	15.97
SFT (w/ ref)	28.57	32.85	30.71
SFT (+tool+CoT)	**33.14**	**36.56**	**34.85**
Setting (YouTube)	**STI**	**PCR**	**Avg. Acc.**
ZS (w/o ref)	17.30	23.84	20.57
SFT (w/o ref)	23.39	25.53	24.46
ZS (w/ ref)	13.61	14.97	14.29
SFT (w/ ref)	24.52	28.53	26.52
SFT (+tool+CoT)	**34.88**	**30.21**	**32.54**

B.3. Additional Comparison on Referential Pipeline

We tested the zero-shot performance of Omni-LLMs on Ego4D using tool-extracted social cues. The results show that they remain substantially worse than our SFT model. This confirms that the observed improvements are not solely

Table 3. Comparison on Ego4D using tool-extracted social cues. Our model significantly outperforms zero-shot Omni-LLMs, indicating that the gains are not solely from access to extracted cues but also from effective task formulation and CoT supervision.

Model	STI	PCR	Avg. Acc.
OmniVinci	31.43	12.50	21.96
Qwen3 Omni 30B	30.28	21.43	25.86
Gemini 2.5 Pro	36.00	33.92	34.96
Omni-MMSI-R	**40.57**	**45.54**	**43.06**

due to access to extracted cues, which might contain speech errors. The performance gain can also arise from effective task formulation and CoT reasoning supervision.

B.4. Example of CoT Reasoning Trace

We present an example of a curated CoT for pronoun coreference recognition. As illustrated in Fig. 4, the CoT performs two key steps-last speaker confirmation and referent inference-to reach the final decision, leading to reliable prediction. The model leverages the identity-attributed transcript and reference audio from the reference-guided input in the last speaker confirmation step. In addition, non-verbal cues such as gaze and gesture are incorporated to complement the verbal evidence in speaker's referent inference. Through supervision from such CoT annotations, the model learns not only structured step-by-step reasoning but also more effective integration of all available social cues.

1. **Last speaker confirmation**: The last speaker is Player3. His voice matches the reference audio for Player3, and his mouth is visibly moving as he says, "I don't know what she did."

2. **Speaker's referent inference**:
   - **Verbal Cues**: The dialogue is a direct exchange between Player1 and Player3. Player1 establishes his role (Robber) and what he knows about Player3's role (Seer). Player3 accepts this but raises a "problem." The problem is his lack of knowledge about what a female player did. The pronoun "she" indicates the referent must be a female player. The female players present are Player0 and Player4. The verbal content alone does not specify which female player.
   - **Non-verbal Cues**: The crucial cue comes from Player3's gaze and gesture. He is in a direct conversation with Player1, who is to his right (viewer's left). When Player3 delivers the final line containing the pronoun "she", he shifts his gaze from Player1 to his own left (viewer's right). Player4 is seated directly to his left. This shift in gaze while mentioning "she" strongly indicates he is referring to Player4. Besides, Both Player3 and Player1 are pointing to Player4.

3. **Final decision**: Player3 is referring to Player4. The decisive cues are the verbal use of the pronoun "she" combined with the non-verbal cue of Player3 shifting his gaze and pointing toward Player4 as he speaks.

Figure 4. Example of CoT. The CoT performs two key steps-last speaker confirmation and referent inference-to reach the final decision based on reference-based input, leading to reliable prediction.

C. Future Works

Omni-MMSI and Omni-MMSI-R demonstrate promising progress toward identity-attributed social interaction understanding, which better supports future exploration of richer social scenes and social tasks. A current limitation is that the datasets used in this work represent controlled scenarios where all participants remain visible under a fixed game setting. Although such setups make manually identity attribution easy, they capture only a narrow portion of real-world social dynamics. In natural environments, people enter or exit the scene and camera viewpoints often change abruptly. Multi-person interactions in movies, television content, and outdoor gatherings also involve frequent camera cuts and heterogeneous visual contexts. Under our reference-based design, these multi-shot and camera-switching scenes, which were previously difficult to study because prior methods could not maintain consistent attribution across shots, become substantially more feasible to curate. Stable reference identities allow reliable cross-shot identity grounding, enabling richer and more realistic social scenarios to be included in future datasets. Extending Omni-MMSI to larger and more diverse environments is an important direction for improving real-world applicability.

D. Societal Impacts and Concerns

While Omni-MMSI aims to advance socially-intelligent AI assistants by enabling real-world perception and reasoning over individual-level verbal and non-verbal cues, these same capabilities introduce potential societal risks. In particular, the ability to align speech, gaze, and gestures with specific individuals, central to our reference-based cue attribution framework, could be misused for intrusive monitoring of social interactions, workplace surveillance, or targeted behavioral manipulation if deployed without consent or appropriate safeguards. Moreover, because Omni-MMSI operates on imperfectly extracted audio-visual cues, systematic errors in speech recognition, tracking, or gaze estimation may disproportionately affect certain demographic groups, potentially amplifying existing biases in downstream decisions. These concerns highlight that the contributions in this work are intended strictly for research on accurate multi-modal social understanding rather than for surveillance applications. Responsible deployment of such systems requires strong privacy protections, transparent usage policies, and governance mechanisms that prevent misuse, especially in real-world settings where individual-level attribution carries heightened ethical implications.

System Prompt for CoT Generation

Inputs
Reference image-audio pairs
You are provided with a reference image-audio pair for each player (referred to as PlayerN). These references are provided to help the model extract attributed verbal and non-verbal social cues by aligning and comparing them with the reference image-audio pairs, enabling the alignment of social cues.

Bounding box coordinates
Each player is associated with a bounding box in the format: `[PlayerN]`: $[x_{min}, y_{min}, x_{max}, y_{max}]$. These coordinates are preliminary and may contain errors or drift. Use them to locate players in the query video to help reason about who is speaking and who the speaker is referring to.

Transcription
You are provided with a transcription with several utterance segments for the query video-audio pair in the format: `[PlayerN]`: [utterance content]. These segments are preliminary and may contain incorrect speaker tags or wording. Use them to help reason about who is speaking and who the speaker is referring to. You can transcribe the target audio and use commonsense to correct preliminary transcript.

Query video-audio pair
This is the video-audio pair in which you must extract attributed verbal and non-verbal social cues and determine who the last speaker is referring to.

Task
Determine **who the last speaker refers to when they say "he", "his", "him", "she", or "her"** in the query video-audio pair.
Treat the task as two mandatory stages:
1. Extract attributed verbal and non-verbal social cues, including speaker-attributed transcript and speaker-attributed visual behaviors like gaze and gesture.
2. Analyze these social cues to infer the last speaker and their referent, especially speech content, dialog turn-taking, and visual engagement.

Decision rules
1. Denote speakers as Player0, Player1, etc., based on their position from left to right in the video.
2. Referent must be a PlayerN, including off-screen or occluded Players that are included in conversation.
3. Every conclusion must cite both **verbal** signals (e.g., speaker content matching with previous dialog content, previous speaker) and **non-verbal** cues (e.g., speaker and who are making mutual eye contact, speaker is pointing at whom). Generic statements like "they are near each other" without interaction detail are insufficient.

Required social cues structure
Inside <cue>... </cue> explicitly cover:
- Speaker-attributed transcript. Each utterance represents one utterance segment in chronological order.
- Speaker-attributed visual behaviors. You must include the bounding box coordinates [x_min, y_min, x_max, y_max] for each player when describing their visual behaviors.

Required reasoning structure
Inside <think>... </think> explicitly cover:
1. Last speaker confirmation with provided audio, vision, and speech evidence.
2. Speaker's referent inference from the verbal and non-verbal interaction cues.
3. Final decision that names the referent and states the decisive cues.

Output format
1. A social cues trace wrapped in <cue>... </cue> that follows the structure above.
2. A reasoning trace wrapped in <think>... </think> that follows the structure above.
3. The final answer wrapped in <answer>PlayerN </answer>.

Examples
<cue>The verbal cues of all players are: `[Player1]`: And then this- `[Player2]`: That one was you three. `[Player1]`: Yeah. `[Player3]`: So it's between us two? `[Player0]`: Yeah. `[Player3]`: You said, "If you keep going on the rejection strategy you'll lose," meaning he's not on the good team, because I am." The non-verbal cues of all players are: Player0 ([0.001, 0.719, 0.165, 0.992]) is looking at Player4's hands. Player1 ([0.125, 0.597, 0.333, 0.992]) glances at the cards in their hand, then looks up. Player2 ([0.379, 0.728, 0.519, 0.992]) has a look at his cards on the table and then looks at Player4. Player3 ([0.546, 0.664, 0.703, 0.989]) is looking at his watch. Player4 ([0.679, 0.728, 0.872, 0.989]) is only visible from the left rear side, but it can be inferred that she looks at the cards on the table and then faces Player2 directly. </cue><think>1. Last speaker: Player4 delivers the final utterance "Okay. Do you need the script?" and their voice matches the Player4 reference while the Player4 bounding box shows their mouth moving. 2. Speaker referents: Player2 (responded just before), Player3 (standing nearby but disengaged). 3. Decision: Player4 addresses Player2 based on directed gesture and mutual gaze. </think><answer>Player2 </answer>

Figure 5. System prompt for CoT generation.

System Prompt for Model Training

Inputs
Reference image-audio pairs
You are provided with a reference image-audio pair for each player (referred to as PlayerN). These references are provided to help the model extract attributed verbal and non-verbal social cues by aligning and comparing them with the reference image-audio pairs, enabling the alignment of social cues.

Bounding box coordinates
Each player is associated with a bounding box in the format: `[PlayerN]`: $[x_{min}, y_{min}, x_{max}, y_{max}]$. These coordinates are preliminary and may contain errors or drift. Use them to locate players in the query video to help reason about who is speaking and who the speaker is referring to.

Transcription
You are provided with a transcription with several utterance segments for the query video-audio pair in the format: `[PlayerN]`: [utterance content]. These segments are preliminary and may contain incorrect speaker tags or wording. Use them to help reason about who is speaking and who the speaker is referring to. You can transcribe the target audio and use commonsense to correct preliminary transcript.

Query video-audio pair
This is the video-audio pair in which you must extract attributed verbal and non-verbal social cues and determine who the last speaker is referring to.

Task
Determine **who the last speaker refers to when they say "he", "his", "him", "she", or "her"** in the query video-audio pair.
Treat the task as two mandatory stages:
1. Extract attributed verbal and non-verbal social cues, including speaker-attributed transcript and speaker-attributed visual behaviors like gaze and gesture.
2. Analyze these social cues to infer the last speaker and their referent, especially speech content, dialog turn-taking, and visual engagement.

Decision rules
1. Denote speakers as Player0, Player1, etc., based on their position from left to right in the video.
2. Referent must be a PlayerN, including off-screen or occluded Players that are included in conversation.
3. Every conclusion must cite both **verbal** signals (e.g., speaker content matching with previous dialog content, previous speaker) and **non-verbal** cues (e.g., speaker and who are making mutual eye contact, speaker is pointing at whom). Generic statements like "they are near each other" without interaction detail are insufficient.

Required reasoning structure
Inside <think>... </think> explicitly cover:
1. Last speaker confirmation with provided audio, vision, and speech evidence.
2. Speaker's referent inference from the verbal and non-verbal interaction cues.
3. Final decision that names the referent and states the decisive cues.

Output format
1. A reasoning trace wrapped in <think>... </think> that follows the structure above.
2. The final answer wrapped in <answer>PlayerN </answer>.

Examples
<think>1. Last speaker: Player4 delivers the final utterance "Okay. Do you need the script?" and their voice matches the Player4 reference while the Player4 bounding box shows their mouth moving. 2. Speaker referents: Player2 (responded just before), Player3 (standing nearby but disengaged). 3. Decision: Player4 addresses Player2 based on directed gesture and mutual gaze. </think><answer>Player2 </answer>

Figure 6. System prompt for model training.

References

[1] 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 *CVPR*, pages 14585–14595, 2024. 2

[2] Xinpeng Li, Shijian Deng, Bolin Lai, Weiguo Pian, James M Rehg, and Yapeng Tian. Towards online multi-modal social interaction understanding. *arXiv preprint arXiv:2503.19851*, 2025. 2

[3] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 3

System Prompt for Omni-LLMs Evaluation w/o Reference

Inputs
Query video-audio pair
This is the video-audio pair in which you must extract attributed verbal and non-verbal social cues and determine who the last speaker is referring to.

Task
Determine **who the last speaker refers to when they say "he", "his", "him", "she", or "her"** in the query video-audio pair.
Treat the task as two mandatory stages:
1. Extract attributed verbal and non-verbal social cues, including speaker-attributed transcript and speaker-attributed visual behaviors like gaze and gesture.
2. Analyze these social cues to infer the last speaker and their referent, especially speech content, dialog turn-taking, and visual engagement.

Decision rules
1. Denote speakers as Player0, Player1, etc., based on their position from left to right in the video.
2. Referent must be a PlayerN, including off-screen or occluded Players that are included in conversation.
3. Every conclusion must cite both **verbal** signals (e.g., speaker content matching with previous dialog content, previous speaker) and **non-verbal** cues (e.g., speaker and who are making mutual eye contact, speaker is pointing at whom). Generic statements like "they are near each other" without interaction detail are insufficient.

Required social cues structure
Inside <cue>... </cue> explicitly cover:
- Speaker-attributed transcript. Each utterance represents one utterance segment in chronological order.
- Speaker-attributed visual behaviors. You must include the bounding box coordinates [x_min, y_min, x_max, y_max] for each player when describing their visual behaviors.

Required reasoning structure
Inside <think>... </think> explicitly cover:
1. Last speaker confirmation with provided audio, vision, and speech evidence.
2. Speaker's referent inference from the verbal and non-verbal interaction cues.
3. Final decision that names the referent and states the decisive cues.

Output format
1. A social cues trace wrapped in <cue>... </cue> that follows the structure above.
2. A reasoning trace wrapped in <think>... </think> that follows the structure above.
3. The final answer wrapped in <answer>PlayerN </answer>.

Examples
<cue>The verbal cues of all players are: `[Player1]`: And then this- `[Player2]`: That one was you three. `[Player1]`: Yeah. `[Player3]`: So it's between us two? `[Player0]`: Yeah. `[Player3]`: You said, "If you keep going on the rejection strategy you'll lose," meaning he's not on the good team, because I am." The non-verbal cues of all players are: Player0 ([0.001, 0.719, 0.165, 0.992]) is looking at Player4's hands. Player1 ([0.125, 0.597, 0.333, 0.992]) glances at the cards in their hand, then looks up. Player2 ([0.379, 0.728, 0.519, 0.992]) has a look at his cards on the table and then looks at Player4. Player3 ([0.546, 0.664, 0.703, 0.989]) is looking at his watch. Player4 ([0.679, 0.728, 0.872, 0.989]) is only visible from the left rear side, but it can be inferred that she looks at the cards on the table and then faces Player2 directly. </cue><think>1. Last speaker: Player4 delivers the final utterance "Okay. Do you need the script?" and their voice matches the Player4 reference while the Player4 bounding box shows their mouth moving. 2. Speaker referents: Player2 (responded just before), Player3 (standing nearby but disengaged). 3. Decision: Player4 addresses Player2 based on directed gesture and mutual gaze. </think><answer>Player2 </answer>

Figure 7. System prompt for Omni-LLMs evaluation without reference.

System Prompt for Omni-LLMs Evaluation w/ Reference

Inputs
Reference image-audio pairs
You are provided with a reference image-audio pair for each player (referred to as PlayerN). These references are provided to help the model extract attributed verbal and non-verbal social cues by aligning and comparing them with the reference image-audio pairs, enabling the alignment of social cues.

Query video-audio pair
This is the video-audio pair in which you must extract attributed verbal and non-verbal social cues and determine who the last speaker is referring to.

Task
Determine **who the last speaker refers to when they say “he”, “his”, “him”, “she”, or “her”** in the query video-audio pair.
Treat the task as two mandatory stages:
1. Extract attributed verbal and non-verbal social cues, including speaker-attributed transcript and speaker-attributed visual behaviors like gaze and gesture.
2. Analyze these social cues to infer the last speaker and their referent, especially speech content, dialog turn-taking, and visual engagement.

Decision rules
1. Denote speakers as Player0, Player1, etc., based on their position from left to right in the video.
2. Referent must be a PlayerN, including off-screen or occluded Players that are included in conversation.
3. Every conclusion must cite both **verbal** signals (e.g., speaker content matching with previous dialog content, previous speaker) and **non-verbal** cues (e.g., speaker and who are making mutual eye contact, speaker is pointing at whom). Generic statements like “they are near each other” without interaction detail are insufficient.

Required social cues structure
Inside <cue>... </cue> explicitly cover:
- Speaker-attributed transcript. Each utterance represents one utterance segment in chronological order.
- Speaker-attributed visual behaviors. You must include the bounding box coordinates [x_min, y_min, x_max, y_max] for each player when describing their visual behaviors.

Required reasoning structure
Inside <think>... </think> explicitly cover:
1. Last speaker confirmation with provided audio, vision, and speech evidence.
2. Speaker’s referent inference from the verbal and non-verbal interaction cues.
3. Final decision that names the referent and states the decisive cues.

Output format
1. A social cues trace wrapped in <cue>... </cue> that follows the structure above.
2. A reasoning trace wrapped in <think>... </think> that follows the structure above.
3. The final answer wrapped in <answer>PlayerN </answer>.

Examples
<cue>The verbal cues of all players are: `[Player1]`: And then this- `[Player2]`: That one was you three. `[Player1]`: Yeah. `[Player3]`: So it’s between us two? `[Player0]`: Yeah. `[Player3]`: You said, “If you keep going on the rejection strategy you’ll lose,” meaning he’s not on the good team, because I am.” The non-verbal cues of all players are: Player0 ([0.001, 0.719, 0.165, 0.992]) is looking at Player4’s hands. Player1 ([0.125, 0.597, 0.333, 0.992]) glances at the cards in their hand, then looks up. Player2 ([0.379, 0.728, 0.519, 0.992]) has a look at his cards on the table and then looks at Player4. Player3 ([0.546, 0.664, 0.703, 0.989]) is looking at his watch. Player4 ([0.679, 0.728, 0.872, 0.989]) is only visible from the left rear side, but it can be inferred that she looks at the cards on the table and then faces Player2 directly. </cue><think>1. Last speaker: Player4 delivers the final utterance “Okay. Do you need the script?” and their voice matches the Player4 reference while the Player4 bounding box shows their mouth moving. 2. Speaker referents: Player2 (responded just before), Player3 (standing nearby but disengaged). 3. Decision: Player4 addresses Player2 based on directed gesture and mutual gaze. </think><answer>Player2 </answer>

Figure 8. System prompt for Omni-LLMs evaluation with reference