This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Omni-MMSI: Toward Identity-attributed Social Interaction Understanding

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

Abstract

*We introduce **Omni-MMSI**, a new task that requires comprehensive social interaction understanding from raw audio, vision, and speech input. The task involves perceiving identity-attributed social cues (e.g., who is speaking what) and reasoning about the social interaction (e.g., whom the speaker refers to). This task is essential for developing AI assistants that can perceive and respond to human interactions. Unlike prior studies that operate on oracle-preprocessed social cues, Omni-MMSI reflects realistic scenarios where AI assistants must perceive and reason from raw data. However, existing pipelines and multi-modal LLMs perform poorly on Omni-MMSI because they lack reliable identity attribution capabilities, which leads to inaccurate social interaction understanding. To address this challenge, we propose **Omni-MMSI-R**, a reference-guided pipeline that produces identity-attributed social cues with tools and conducts chain-of-thought social reasoning. To facilitate this pipeline, we construct participant-level reference pairs and curate reasoning annotations on top of the existing datasets. Experiments demonstrate that Omni-MMSI-R outperforms advanced LLMs and counterparts on Omni-MMSI. Project page: https://sampson-lee.github.io/omni-mmsi-project-page.*

1. Introduction

Multi-modal Multi-party Social Interaction Understanding (MMSI), aiming to interpret human behaviors in social situations, is fundamental for advancing socially-intelligent AI systems [24, 44–47]. As shown in Figure 1, given audio-video input, the system is required to extract identity-attributed verbal and non-verbal social cues. For instance, the chronological utterances, *[Player2]: All right. [Player4]: Okay. Do you need the script?*, and their corresponding bounding boxes, *[0.018, 0.736, 0.186, 0.992] and [0.668, 0.742, 0.875, 0.989]*, constitute essential identity-attributed social cues. Then, the system should analyze these multi-modal social cues to infer the social interaction, *i.e. determine whom the last speaker refers to in the query audio-video*. These capabilities are essential for enabling AI assistants that can perceive, reason over, and respond to human interactions in social scenarios [4, 20, 32].

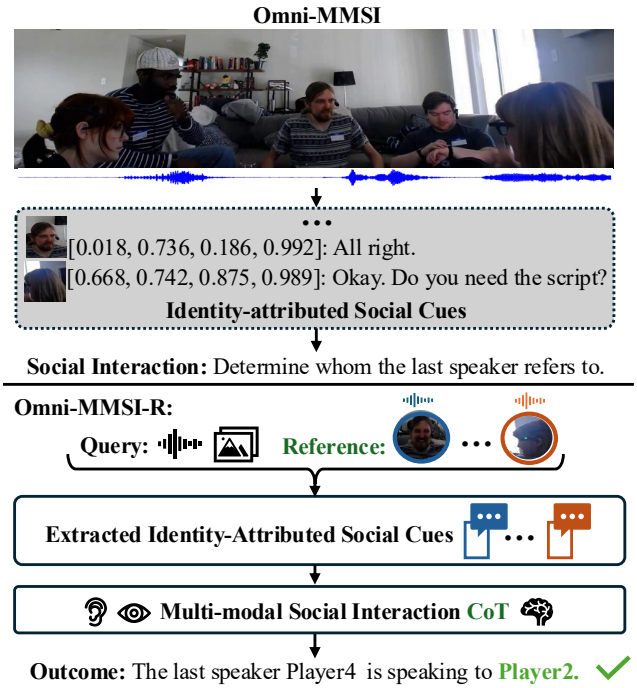

Figure 1. Overview of the Omni-MMSI task and Omni-MMSI-R pipeline. The Omni-MMSI explores social interaction understanding in a multi-party social scene only using raw audio and video, unlike prior studies that assume identity-attributed social cues are perfectly provided. To address the challenge of attribution, our Omni-MMSI-R is explicitly guided by individual references to generate identity-attributed multi-modal cues and performs CoT reasoning for accurate social interaction understanding.

Recent computer vision studies have explored social interaction understanding and advanced it with representation alignment [45] and conversation forecasting [50]. Despite the rapid progress, they remain limited in scope: they assume the individual-attributed social cues are perfectly provided, typically via oracle-preprocessing. However, in real-world deployment, AI assistants must understand social interactions from raw data input. To better align with realistic applications, we introduce a new task, named ***Omni-MMSI***, which requires social interaction understanding on raw audio-video input. The system needs to extract identity-attributed social cues, including who speaks what and where they are, and then infer the social interaction.

However, identity attribution is challenging in multi-party scenes, where people show subtle movements, and their voices also sound alike, with a lot of overlap. First, the off-the-shelf extractors [42, 68] that can be used in earlier studies were designed for single-person scenarios and fail to handle the crucial attribution step required in Omni-MMSI. Second, while Omni-modal Large Language Models (Omni-LLMs) demonstrate strong cue extraction, they still struggle to correctly associate these cues with individuals across modalities. Therefore, prior pipelines and Omni-LLMs degrade significantly when transitioning from oracle identity-attributed cues to raw inputs. As shown in Fig. 2, the accuracy of prior pipelines [45, 50] drops by an average of 28.1%, and even human annotators and advanced Omni-LLMs [18, 90] exhibit an average decline of 9.52%.

To tackle this challenge, we propose **Omni-MMSI-R**, a LLM-based pipeline that utilizes references to guide identity attribution. Our key insight is that humans remember the appearance and voice of familiar people, and readily associate their gestures or speech with these memories when interpreting social interactions. In practical use, these references are usually easy to collect on devices through the enrollment or verification processes [16, 40]. As shown in Fig. 1, to generate accurate identity-attributed social cues, task-specific tools associate cues with references. Then, to further enhance MMSI ability, the model performs chain-of-thought (CoT) reasoning. To facilitate such a pipeline, we manually construct paired image-audio references for each sample and curate a CoT reasoning dataset.

We evaluate Omni-MMSI-R on two social interaction tasks across two social datasets, Ego4D and YouTube [45]. Our method outperforms previous studies by 12% on Ego4D and 15.1% on YouTube in social interaction understanding and exceeds advanced LLMs by 23.7% on Ego4D and 18.9% on YouTube in identity attribution, demonstrating that Omni-MMSI-R benefits from reference guidance.

In summary, our contributions are four-fold:

- We present Omni-MMSI, a new task for realistic scenarios that requires multi-party multi-modal social interaction understanding only using raw audio-vision input.
- We propose Omni-MMSI-R, a reference-guided pipeline that generates identity-attributed social cues with tools and performs CoT reasoning for accurate MMSI.
- We curate paired audio-vision references and CoT reasoning annotations for two current datasets for future study.
- Experiments on two social interaction tasks across two datasets demonstrate that the proposal benefits from reference guidance and achieves state-of-the-art performance.

2. Related Works

2.1. Multi-modal Social Interaction Understanding

MMSI aims to interpret complex interactions among multiple participants by using verbal and non-verbal cues [5, 31, 44–47, 50, 65, 93]. The non-verbal social cues include visual behaviors such as body gestures, gaze patterns, and facial expressions [3, 6, 7, 15, 28, 41, 43, 48, 49, 58, 61, 63, 72, 73, 76, 83, 87, 100, 102]. The verbal social cues include linguistic signals such as conversational dynamics, speaker intent, speaker diarization, and dialogue sentiment [11, 13, 23, 24, 30, 33, 37, 52–54, 60, 62, 67, 69, 71, 88].

Despite these advances, these works all assume perfectly provided individual-attributed cues as model input, overlooking the gap between raw audio-visual input and attributed social cues in realistic deployment. In contrast, Omni-MMSI focuses on social interaction understanding only using streaming audio and video, where the system must first extract identity-attributed verbal and non-verbal social cues and then reason about the social interaction.

2.2. Multi-modal Foundation and Reasoning Model

Multi-modal foundation models pave the way toward better intelligent systems. While proprietary models [17, 36, 78] often showcase strong performance, open-weight models [1, 2, 9, 12, 14, 55, 56, 79–82, 90, 92, 94, 96, 97, 101, 103, 104, 108] provide more opportunities for specialized downstream tasks, making them useful for multi-modal social interaction understanding. Reasoning [86] as an emergent ability of LLMs [85] has recently attracted attention recently for its effectiveness under text-only settings [29, 39, 64, 66]. Multi-modal reasoning models extend this success to general image understanding [8, 10, 19, 59, 74, 89], video understanding [22, 51, 84, 95, 99] and some vertical domains like medical image understanding [35, 75]. Tooling further extends LLMs' ability to perform a broad spectrum of tasks through the use of tools [21, 25–27, 57, 98, 106].

However, CoT reasoning and tooling paradigms remain unexplored in MMSI. To advance computer vision and social AI community, we curate paired audio-vision references and CoT reasoning traces on top of existing datasets, and demonstrate the effectiveness of CoT and tooling.

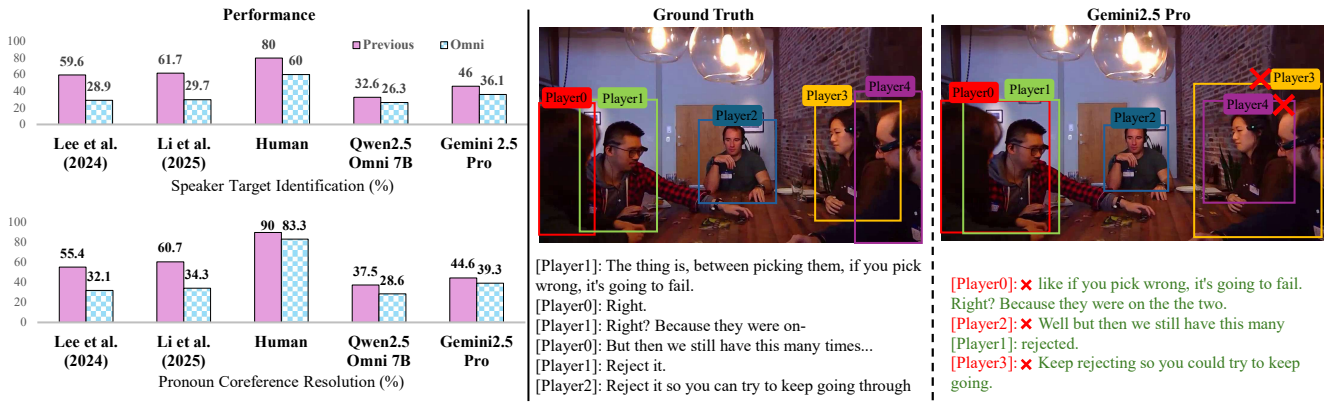

Figure 2. Illustration of the challenge in Omni-MMSI. The quantitative results (left) show prior pipelines, humans, and advanced Omni-LLMs show substantial accuracy drops when transitioning from oracle cues to raw audio-video input. Typical attribution failures (right), where speech and bounding boxes are mismatched to identities, reveal the weak multi-modal identity attribution of advanced Omni-LLMs.

3. Problem Formulation and Challenges

Omni-MMSI pursues the MMSI abilities on raw audio-visual input instead of relying on oracle cues. Specifically, we study two typical MMSI tasks [45, 50]: Speaking Target Identification (STI) and Pronoun Coreference Resolution (PCR). STI aims to identify who the speaker is talking to when the utterance contains a second-person reference, e.g., "you" and "your"; PCR focuses on resolving which participant a third-person pronoun refers to, e.g., "he", "she", "him", "her" and "his". The inputs are a raw audio-video segment I_{AV} and system prompt P that configures a specific task. The output X_{answer} is the predicted referent identity. The Omni-MMSI is to build a system f:

$$f : (P, I_{AV}) \rightarrow X_{answer}. \tag{1}$$

Unlike previous studies [45, 50] that assume oracle-preprocessing social cues as input, Omni-MMSI operates on the raw audio-video segment, requiring models to automatically extract social cues and infer social interaction. To assess the challenge, we evaluate performance on the social Ego4D dataset across two social tasks when transferring from oracle input to raw-data one. As shown in Fig. 2, previous pipelines [45, 50] and advanced Omni-LLMs such as Qwen2.5 Omni 7B [90] and Gemini 2.5 Pro [18] exhibit significant performance drops, confirming that the Omni-MMSI poses a significant challenge. It also underscores that current LLMs still fall short of human-level understanding in multi-modal and multi-party social reasoning [38, 77].

The major bottleneck is identity attribution ability on raw audio-visual input. On the one hand, off-the-shelf extractors in prior pipelines [45, 50] are designed for single-person scenarios, failing to attribute cues to individuals in a multi-party setting. On the other hand, although recent Omni-LLMs [18, 90] have shown promising performance in extracting cues, they still struggle to associate detected cues with the corresponding subjects. As illustrated in Fig. 2, Gemini 2.5 Pro often assigns speech content or bounding boxes to the wrong identity. Specifically, for visual attribution, Gemini 2.5 Pro attributes participants based on their left-to-right spatial order, but this assumption leads to identity swaps when detection fails under occlusion or overlapping. For speech attribution, Gemini 2.5 Pro often mismatches the recognized utterance with wrong identity. Such weak multi-modal association results in inaccurate social cues, ultimately degrading social interaction reasoning.

4. Methodology

4.1. Overview of Omni-MMSI-R

To tackle the difficulty of social cues attribution, we introduce Omni-MMSI-R that leverages references $\mathcal{R}$ to generate identity-attributed social cues and perform CoT social reasoning. The system target can be formulated as:

$$f : (P, I_{AV}, \mathcal{R}) \rightarrow X_{answer}. \tag{2}$$

As shown in Fig. 3, given a query audio-video segment, Omni-MMSI-R loads a set of reference audio-image pairs that store representative visual and acoustic profiles for each individual. Based on these references, task-specific tools generate identity-attributed multi-modal social cues, such as conversation transcripts and individual locations. Then, an Omni-LLM performs CoT reasoning on the audio-video segment and reference audio-image pairs, along with generated attributed cues, and produces an accurate answer.

4.2. Reference Guidance

To address the difficulty of identity attribution on raw audio-video input, we propose to associate social cues with guided references. The insight is that humans rely on the appearance and voice of memorized people to guide identity as-

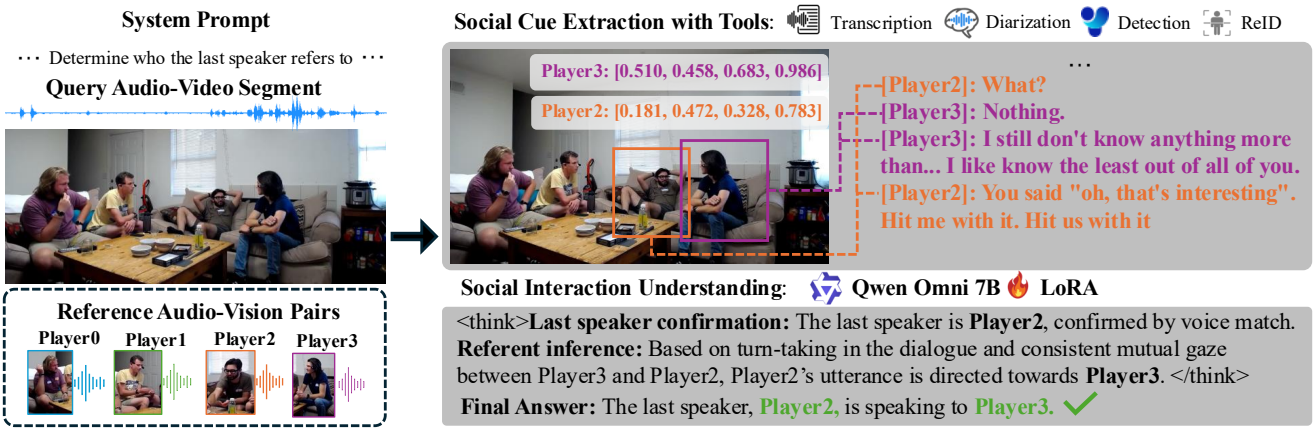

Figure 3. Overview of the Omni-MMSI-R pipeline. Given a query audio-video segment with multiple participants, the system first retrieves reference audio-vision pairs that represent each individual. Task-specific tools, for transcription, diarization, detection and ReID, generate identity-attributed verbal and non-verbal social cues, specifying who speaks what and where they are. These cues, together with the references and the raw audio-video stream, form the reference-guided input. The Omni-LLM (Qwen2.5 Omni 7B fine-tuned with LoRA) then performs chain-of-thought reasoning over this input to produce an accurate response for social interaction understanding.

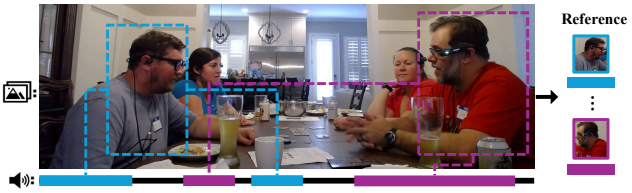

Figure 4. Illustration of preparation of reference audio-vision pairs for each participant, which serve as anchors for identity attribution.

sociation in multi-party situations. In practical use, the references are usually easy to collect on devices through the enrollment or verification processes [16, 40].

For research purposes, we manually crop each participant's upper body image and extract several corresponding voice clips to build the reference pairs, as shown in Figure 4. In total, we curate 69 audio-visual reference profiles covering different participants across the experimental datasets.

Omni-MMSI-R can access the reference audio-visual set $\mathcal{R} = \{(a_i, v_i)\}_{i=1}^{N}$ for all N participants in the scene, where a_i and v_i denote the representative voice and appearance of participant i. These references anchor identities across modalities and time, reducing common failures, like identity swaps under occlusion and cross-modal mismatches, and yielding accurate identity-attributed social cues.

4.3. Social Cue Extraction with Tools

To generate accurate cues, we leverage tools to help detect social cues and associate them with reference identities.

Audio Tools. We first apply Whisper [68] to transcribe the query audio into a sequence of utterances with timestamps. For each utterance, SpeechBrain [70] performs speaker verification by encoding both the utterance audio and each reference voice into embeddings and computing their cosine similarity. The reference with the highest similarity is selected as the predicted speaker identity. This process yields identity-attributed verbal social cues that contain transcribed speech and the corresponding speaker identity.

Visual Tools. We first leverage YOLO [42] to detect all visible participants in the last frame of the query video. For every detected bounding box, we then employ OSNet [107] for person re-identification. Specifically, both the detected image crop and each reference image are encoded into visual embeddings, and the similarity between them is computed. The reference with the highest similarity is selected as the predicted visual identity. This produces identity-attributed non-verbal social cues that specify both the spatial position and identity of each participant in the scene.

After extraction, the identity-attributed social cues $\mathcal{S}$, along with the query audio-video segment I_{AV} and the reference audio-image pairs $\mathcal{R}$, are fed into an Omni-LLM.

4.4. Social Interaction Understanding with CoT

Omni-MMSI naturally involves multi-step fine-grained understanding: (1) confirming the last speaker from audio, visual, and speech evidence, and (2) inferring the speaker's referent by integrating verbal cues, such as matching the utterance with prior dialog and speaker context, and non-verbal interaction signals, including mutual eye contact or pointing. Training models to directly output answers often fails to capture the fine-grained evidence, resulting in less reliable responses. Therefore, we propose to supervise the model with structured CoT reasoning traces.

CoT Data Curation. To facilitate the training, we curate CoT annotations by a generate-and-filter pipeline. As illustrated in Fig. 5, (i) we upload query segment, reference input, and social cues to Gemini 2.5 Pro and request it to

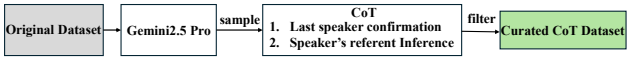

Figure 5. Illustration of the construction of CoT datasets.

generate both social reasoning traces and a final answer, including last speaker confirmation and referent inference with verbal and non-verbal evidence. (ii) Based on the rejection sampling principle, a generated sample is retained only if the reasoning trace leads to a final answer that is consistent with the ground truth. Otherwise, we repeatedly query Gemini 2.5 Pro until a correct answer is obtained, or stop after 10 attempts. (iii) To further ensure the CoT quality, we perform a lightweight human review to discard or minimally revise reasoning traces that are implausible or inconsistent with the audio-visual evidence. Through this process, we obtain a set of samples with reliable and interpretable social reasoning traces. Fig. 3 shows a CoT example: *<think>Last speaker confirmation: The last speaker is Player2, confirmed by voice match. Speaker's referent inference: Based on turn-taking in the dialogue and consistent mutual gaze between Player3 and Player2, Player2's utterance is directed towards Player3. </think>*

Model Training. After obtaining the CoT reasoning traces X_{think}, we train the model for MMSI, formulated as:

$$X_{answer}, X_{think} = f_\theta^{\text{Omni-LLM}}\big(P, I_{AV}, \mathcal{R}, \mathcal{S}\big), \quad (3)$$

where $f_\theta^{\text{Omni-LLM}}$ denotes the Omni-LLM. By learning reasoning over raw data, augmented with references and tool-extracted cues, the model can address Omni-MMSI.

5. Experiments

5.1. Implementation Details

We select **Qwen2.5-Omni-7B** [90] as our omni-modal large language model and LLaMA-Factory [105] framework for supervised fine-tuning (SFT). We apply LoRA [34] fine-tuning with a rank of 8 while other LoRA hyperparameters follow LLaMA-Factory defaults. Training uses cross-entropy loss, a cosine learning-rate scheduler with 10% warm-up and a context length of 16,384 tokens. We train for 3 epochs with per-device batch size 1 and gradient accumulation 1. The learning rate is set to 1×10^{-4} empirically for the speaking target identification and the pronoun coreference resolution task. The query segment is standardized to contain 5 dialogue turns, with average duration of 14 seconds. The reference audio clips are trimmed to 5 seconds, whereas the reference images vary in size. Additional implementation details are provided in the supplementary.

5.2. Dataset and Metrics

Experiments are conducted on the Werewolf Among Us dataset, which comprises two subsets (YouTube and Ego4D) of social deduction games [44]. We follow Lee et al. [45] and Li et al. [50] to set up STI and PCR tasks. In Omni-MMSI, we remove oracle cues, transcript and keypoints, but provide references and CoT reasoning traces.

Table 1. Performance comparison of different pipelines on STI and PCR tasks. The upper block reports results (%) on Ego4D, and the lower on YouTube. The results highlight the effectiveness of our Omni-MMSI-R pipeline design for Omni-MMSI.

Pipeline (Ego4D)	**STI**	**PCR**	**Avg. Acc.**
Qwen2.5 Omni 7B [90]	26.29	28.57	27.43
Phi-4-Multimodal [1]	14.86	8.93	11.90
HumanOmni [104]	21.71	14.29	18.00
R1-Omni [103]	0.57	0.00	0.29
OmniVinci [92]	24.57	37.50	31.04
Qwen3 Omni 30B [91]	30.86	28.57	29.72
Gemini 2.5 Pro [18]	36.12	39.28	<u>37.70</u>
Lee et al.* [45]	28.98	32.14	30.56
Li et al.* [50]	29.73	32.27	31.00
Omni-MMSI-R (ours)	**40.57**	**45.54**	**43.06**
Pipeline (YouTube)	**STI**	**PCR**	**Avg. Acc.**
Qwen2.5 Omni 7B [90]	14.00	26.18	20.09
Phi-4-Multimodal [1]	16.21	21.11	18.66
HumanOmni [104]	20.80	28.02	24.41
R1-Omni [103]	0.15	0.19	0.17
OmniVinci [92]	19.57	29.17	24.37
Qwen3 Omni 30B [91]	18.81	36.85	27.83
Gemini 2.5 Pro [18]	36.13	53.47	<u>44.80</u>
Lee et al.* [45]	29.01	34.80	31.91
Li et al.* [50]	26.30	30.14	28.22
Omni-MMSI-R (ours)	**37.46**	**56.62**	**47.04**

YouTube contains 3,255 samples for STI and 2,679 samples for PCR, with an average of 5 individuals per sample. For each sample, we manually construct reference audio-image pairs. For the training split, we generate CoT reasoning traces and filter out 2,124 samples for STI (average 202 words) and 1,935 samples for PCR (average 220 words).

Ego4D contains 832 STI samples and 503 PCR samples, with each sample involving an average of 5 individuals. For each sample, we manually construct reference audio-image pairs. For the training split, we generate CoT reasoning traces and filter out 521 samples for STI (average 206 words) and 321 samples for PCR (average 226 words).

Evaluation. To evaluate social interaction understanding, we follow previous studies [45, 50] to report the overall accuracy of the predicted referent. To further evaluate identity attribution ability, we compute the accuracy of attributed identity of each utterance (verbal attribution) and detected location on the last frame (non-verbal attribution).

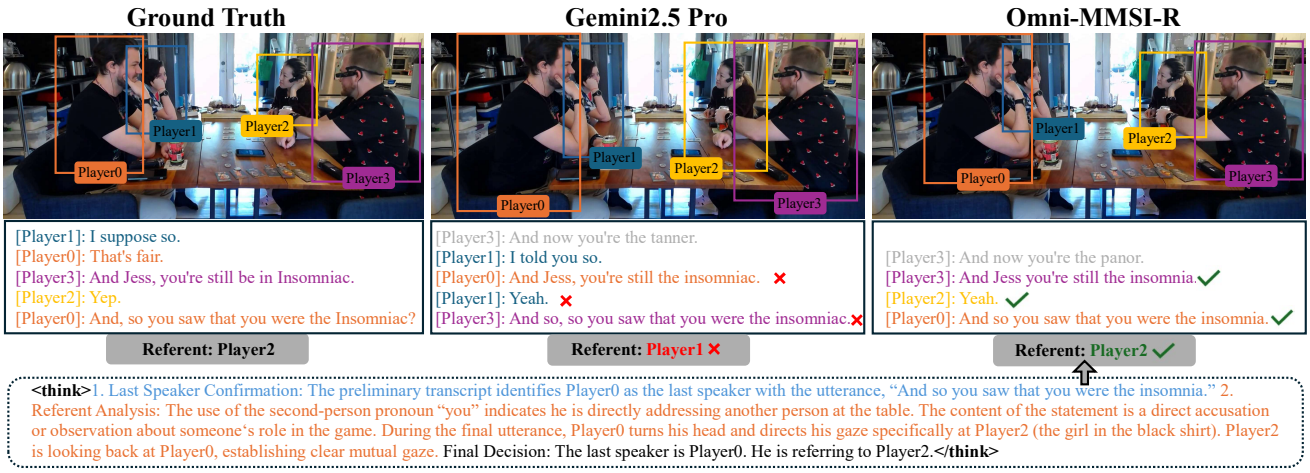

Figure 6. Qualitative comparison between Gemini2.5 Pro and our proposed Omni-MMSI-R on multi-party situations. Gemini2.5 Pro often misattributes utterances to incorrect visual identities, leading to wrong referent predictions, while Omni-MMSI-R accurately aligns verbal and non-verbal cues with individual references, yielding reliable identity-attributed social cues for social interaction reasoning.

5.3. Pipeline Performance Comparison

To evaluate the effectiveness of different pipelines for (1), we conduct evaluation on Ego4D and YouTube. For recent advanced Omni-LLMs [1, 18, 90–92, 103, 104], we directly feed them with the query audio-video pairs and prompt them to generate attributed social cues and social interaction answers. Note that participant identities are deterministically defined by spatial ordering in the system prompt. For the previous MMSI counterparts [45, 50], which overlook the attribution process, we first generate unattributed social cues using extractors and then feed them to the model.

Tab. 1 quantitatively compares the pipelines on social interaction understanding, including STI and PCR. Omni-MMSI-R achieves state-of-the-art performance, reaching 43.06% on Ego4D and 47.04% on YouTube. Relative to existing Omni-LLMs, Omni-MMSI-R improves the average accuracy by 5.36% on Ego4D and 2.24% on YouTube. Compared to previous MMSI pipelines, the improvement reaches 12.06% on Ego4D and 15.13% on YouTube. These results confirm that explicit reference guidance greatly strengthens multi-modal social interaction reasoning.

Beyond interaction reasoning, Tab. 2 reports evaluation results on social cues attribution, including verbal and non-verbal attribution accuracy. Omni-MMSI-R substantially outperforms strong Omni-LLMs, improving the average attribution accuracy by 23.68% on Ego4D and 18.91% on YouTube. Note that since some weak Omni-LLMs [1, 90, 103, 104] fail to generate valid identity-attributed social cues during inference, their attribution accuracy is not reported. These improvements indicate that references significantly enhance the pipeline's ability of identity attribution.

Fig. 6 illustrates qualitative comparisons between Gemini 2.5 Pro and our proposed Omni-MMSI-R. We can see

Table 2. Performance comparison of different pipelines on social cues attribution, including Verbal, Non-Verbal, and Average Attribution accuracy (%). The upper block reports results on Ego4D, and the lower block reports results on YouTube. The results show our Omni-MMSI-R pipeline can achieve better identity attribution.

Pipeline (Ego4D)	**Verbal**	**Non-Verbal**	**Avg.**
OmniVinci [92]	54.04	27.42	40.73
Qwen3 Omni 30B [91]	52.61	57.61	<u>55.11</u>
Gemini 2.5 Pro [18]	44.75	26.52	35.64
Omni-MMSI-R (ours)	**71.09**	**86.48**	**78.79**
Pipeline (YouTube)	**Verbal**	**Non-Verbal**	**Avg.**
OmniVinci [92]	58.10	32.35	45.23
Qwen3 Omni 30B [91]	57.73	52.54	55.14
Gemini 2.5 Pro [18]	50.57	65.51	<u>58.04</u>
Omni-MMSI-R (ours)	**67.57**	**86.33**	**76.95**

Gemini 2.5 Pro fails to attribute utterances to the right visual identities, leading to inaccurate referent prediction. It reflects its limited ability in cross-modal attribution for multi-modal social interaction understanding. Instead, Omni-MMSI-R correctly aligns verbal and non-verbal cues to individual references, producing more reliable identity-attributed social cues. Based on these cues, the model performs CoT reasoning with last speaker confirmation and referent analysis to obtain an accurate social interaction answer. Overall, these quantitative and qualitative results validate the effectiveness of our reference-guided pipeline.

5.4. Referential Pipeline Comparison

To evaluate the effectiveness of different pipelines for (2), we compare our proposal with Omni-LLMs [18, 90–92] on Ego4D and YouTube. We provide the references along

Table 3. Performance comparison of different referential pipelines on STI and PCR tasks. The upper block reports results (%) on Ego4D, and the lower on YouTube. The results highlight the effectiveness of our Omni-MMSI-R pipeline design for Omni-MMSI.

Pipeline (Ego4D)	STI	PCR	Avg. Acc.
Qwen2.5 Omni 7B [90]	21.23	10.71	15.97
OmniVinci [92]	24.00	36.61	30.31
Qwen3 Omni 30B [91]	28.57	39.28	33.94
Gemini 2.5 Pro [18]	40.57	44.64	42.61
Omni-MMSI-R (ours)	**40.57**	**45.54**	**43.06**
Pipeline (YouTube)	**STI**	**PCR**	**Avg. Acc.**
Qwen2.5 Omni 7B [90]	12.08	15.16	13.62
OmniVinci [92]	20.95	28.29	24.62
Qwen3 Omni 30B [91]	34.78	36.19	35.48
Gemini 2.5 Pro [18]	39.87	57.58	**48.72**
Omni-MMSI-R (ours)	37.46	56.62	47.04

Table 4. Comparison of referential pipelines on social cues attribution, including Verbal, Non-Verbal, and Average Attribution accuracy (%). The upper block reports results on Ego4D, and the lower block on YouTube. The results show Omni-MMSI-R, using tools, provides the strongest attribution performance.

Pipeline (Ego4D)	Verbal	N-Verbal	Avg.
Qwen3 Omni 30B [91]	21.92	71.26	46.59
Gemini 2.5 Pro [18]	59.09	35.24	47.17
Omni-MMSI-R (ours)	**71.09**	**86.48**	**78.79**
Pipeline (YouTube)	**Verbal**	**N-Verbal**	**Avg.**
Qwen3 Omni 30B [91]	23.37	50.39	36.88
Gemini 2.5 Pro [18]	60.61	65.45	63.03
Omni-MMSI-R (ours)	**67.57**	**86.33**	**76.95**

with query videos to Omni-LLMs for social cue attribution and social interaction understanding. For social interaction understanding, as shown in Tab. 3, Omni-MMSI-R achieves comparable accuracy to Gemini 2.5 Pro and exceeds open-source Omni-LLMs by 9.12% on Ego4D and 11.56% on YouTube. Compared to non-reference setting, large Omni-LLMs like Qwen3 Omni 30B [91] and Gemini 2.5 Pro [18] obtain performance gains, demonstrating the benefits of reference guidance. However, small Omni-LLMs like Qwen2.5 Omni 7B [90] and OmniVinci [92] degrade in performance. Small models might not be able to utilize the reference, showing the necessity of using tools.

For identity attribution, Tab. 4 shows reference guidance generally improves attribution for Gemini 2.5 Pro, which achieves gains of 11.53% on Ego4D and 4.99% on YouTube. However, not all Omni-LLMs can reliably incorporate references. OmniVinci [92] cannot produce valid social cues when receiving both query and reference audio-vision pairs, so its attribution accuracy cannot be reported. Qwen3 Omni 30B [91] shows lower attribution accuracy after including reference pairs. This indicates LLMs alone struggle to use references effectively for identity attribution. In addition, generating identity-attributed verbal and non-verbal cues through LLMs introduces considerable inference latency. In comparison, the lightweight tools in Omni-MMSI-R provide fast and reliable social cues.

5.5. Effects of Different Reference-guided Input

To further investigate how different reference-guided input in (3) contribute, we conduct ablation studies on Ego4D for social interaction understanding. The baseline means finetuning with only the query audio-video segment. The complete reference-guided input further includes (i) the reference voice-image pairs anchoring individual identities, and (ii) the tool-extracted social cues, consisting of attributed verbal and non-verbal cues. Note that modality is paired: audio references enable verbal cues, while visual references enable non-verbal cues. When one modality is removed, the corresponding attributed cues are also excluded.

As shown in Table 5, the baseline model that finetuned with the query audio-video segment achieves an average accuracy of 33.97%. Adding audio-vision references without attributed cues improves the performance to 35.98%, indicating that raw references alone already help ground more reliable social cues implicitly. Adding attributed cues without audio-image pairs boosts the performance to 39.44%, showing that explicit extracted cues help model understanding social interaction in raw data. By jointly using raw reference data and extracted cues, the model obtains the highest performance 43.05%. It demonstrates that our LLM is not restricted to blindly trusting extracted cues: (1) The LLM is prompted to jointly use extracted identity cues and direct audio-visual evidence from the video, allowing inaccurate cues to be complemented by raw evidence or corrected. (2) The LLM performs explicit reflection on the last speaker identity in its CoT reasoning. As illustrated in **??**, the first CoT step verifies the last speaker by jointly examining voice similarity and visible mouth movement.

Compared to the baseline, incorporating the audio modality together with its attributed verbal cues further increases the accuracy to 39.84%, while adding the vision modality and its attributed non-verbal cues yields 38.56%. When all modalities and their attributed cues are jointly used, the model achieves the highest accuracy of 43.06%, demonstrating complementary contribution of different modalities. These results show multi-modal reference and extracted identity-attributed cue together provide the strongest social cues for social interaction reasoning.

Table 5. Effect of different reference-guided input configurations on social interaction understanding (%). *RA*: Reference Audio, *RV*: Reference Visual Image, *VC*: Verbal Cues, *NC*: Non-Verbal Cues. The results show leveraging audio-visual reference and tool-extracted cue together brings the highest performance.

RA	RV	VC	NC	STI	PCR	Avg. Acc.
✗	✗	✗	✗	29.19	38.68	33.97
✓	✓	✗	✗	32.78	39.18	35.98
✗	✗	✓	✓	37.14	41.75	39.44
✓	✗	✓	✗	37.89	41.78	39.84
✗	✓	✗	✓	34.78	42.33	38.56
✓	✓	✓	✓	**40.57**	**45.54**	**43.06**

5.6. Effectiveness of CoT Reasoning

To analyze the effectiveness of CoT reasoning, we conduct ablation studies on Ego4D dataset. First, we remove the reference pairs and extracted social cues from the input to examine whether reasoning helps social understanding from raw query input. Since last-speaker confirmation depends on the references, we remove that part and keep only referent inference in the reasoning traces to supervise the model. Tab. 6 shows adding only the CoT reasoning enhances performance over the baseline by 1.5%, indicating that reasoning benefits complex social interaction understanding. This may be because the model is trained with fine-grained evidence grounding contained in the reasoning traces. Therefore, the model can better exploit the multi-modal cues, such as pointing and spoken utterances. When jointly using reference-guided input and CoT reasoning, our model achieves the best performance with an average accuracy of 43.06%, a significant improvement of 9.1% compared to baseline, demonstrating their complementary roles: reference-guided input provides more reliable cues, while CoT supervision brings more accurate social interaction understanding. For instance, as you can see in Fig. 6, with accurate reference guidance, the model performs more accurate reasoning to confirm the last speaker as Player0; with CoT, the model exploits fine-grained multi-modal cues.

Second, we investigate the effect of reasoning granularity in CoT supervision, which determines how many intermediate reasoning steps are included during model training. We define four levels of reasoning granularity. The *None* setting provides no intermediate reasoning. The *1-step* setting performs referent inference, where the model explicitly reasons about the target referent of the last speaker. The *2-step* setting further adds last speaker confirmation before referent inference, and the *3-step* setting additionally introduces social cues extraction on top of *2-step*, where the model itself is required to recognize more fine-grained verbal and non-verbal social cues prior to reasoning.

As shown in Tab. 6, introducing moderate reasoning granularity substantially improves performance. Adding referent inference slightly increases the average accuracy compared with no reasoning, and further including last speaker confirmation yields the highest average accuracy. However, adding one more reasoning stage, explicit social cues extraction, leads to a noticeable decrease in performance. This may be attributed to three factors: first, overly long reasoning sequences could distract the model from focusing on the key reasoning path; second, the model's limited ability to accurately perceive and utilize social cues makes such explicit cue extraction overly demanding; and third, training data size may be insufficient to support effective learning of such multi-step reasoning processes. Overall, these results demonstrate that a *2-step* CoT supervision strategy, which includes first confirming the speaker and then inferring the referent, achieves the best performance.

Table 6. Effect of CoT reasoning and different granularity on Ego4D. *Reference* indicates whether reference pairs and attributed social cues are provided. *Reasoning* controls the level of CoT supervision: *None* denotes no reasoning; *CoT* denotes generic reasoning without structured decomposition; *1-step*, *2-step*, and *3-step* represent increasingly fine-grained reasoning strategies. The results show that CoT improves performance even without reference, while combining reference with structured reasoning yields the best results, with *2-step* achieving the optimal balance.

Reference	Reasoning	STI	PCR	Avg. Acc.
✗	✗	29.19	38.75	33.97
✗	✓	30.71	40.18	35.45
✓	None	36.57	42.25	39.41
✓	1-step	36.65	42.75	39.70
✓	2-step	**40.57**	**45.54**	**43.06**
✓	3-step	29.71	39.14	34.43

6. Conclusion

We introduced Omni-MMSI, a new task that requires understanding multi-party social interactions from raw audio-visual input without access to oracle-provided identity-attributed social cues. This setting reflects realistic deployment scenarios where AI systems must operate on automatically extracted cues. To address the resulting challenge of identity attribution, we proposed Omni-MMSI-R, a reference-guided pipeline that aligns multi-modal cues with individual references and performs chain-of-thought (CoT) social reasoning. Through extensive experiments on two social interaction tasks and two social datasets, Omni-MMSI-R demonstrates clear advantages over previous pipelines and advanced Omni-LLMs, achieving state-of-the-art performance. We hope this work establishes a step toward socially intelligent AI that can perceive, reason about, and interact with humans in natural environments.

Acknowledgements

We thank Teng Wang for early-stage inspiration that shaped this line of work. We also thank our colleagues and peers for their valuable feedback and suggestions on this paper.

References

[1] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. 2, 5, 6

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[3] Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji Yanai. Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *2020 25th international conference on pattern recognition (ICPR)*, pages 4340–4347. IEEE, 2021. 2

[4] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. Social robotics. *Springer handbook of robotics*, pages 1935–1972, 2016. 1

[5] 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 Computer Vision and Pattern Recognition Conference*, pages 19509–19519, 2025. 2

[6] Xu Cao, Pranav Virupaksha, Sangmin Lee, Bolin Lai, Wenqi Jia, Jintai Chen, and James Matthew Rehg. Toward human deictic gesture target estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[7] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346–1366, 2023. 2

[8] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models, 2025. 2

[9] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[10] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02. 2

[11] Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817, 2023. 2

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2

[13] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024. 2

[14] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2

[15] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5396–5406, 2020. 2

[16] Jason Clarke, Yoshihiko Gotoh, and Stefan Goetze. Speaker embedding informed audiovisual active speaker detection for egocentric recordings. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2, 4

[17] ClaudeAI. 2

[18] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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. 2, 3, 5, 6, 7

[19] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles, 2025. 2

[20] MM Elsherbini, Ola Mohammed Aly, Donia Alhussien, Ohamed Amr, Moataz Fahmy, Mahmoud Ahmed, Mohamed Adel, Mohamed Fetian, Mahmoud Hatem, Mayar Khaled, et al. Towards a novel prototype for superpower glass for autistic kids. *International Journal of Industry and Sustainable Development*, 4(1):10–24, 2023. 1

[21] Sunqi Fan, Jiashuo Cui, Meng-Hao Guo, and Shuojin Yang. Tool-augmented spatiotemporal reasoning for streamlining video question answering task. *arXiv preprint arXiv:2512.10359*, 2025. 2

[22] 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, 2025. 2

[23] Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gašić. Emowoz: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. *arXiv preprint arXiv:2109.04919*, 2021. 2

[24] Xiachong Feng, Longxu Dou, Minzhi Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. A Survey on Large Language Model-Based Social Agents in Game-Theoretic Scenarios. *Transactions on Machine Learning Research (TMLR)*, 2025. 1, 2

[25] Tianhong Gao, Yannian Fu, Weiqun Wu, Haixiao Yue, Shanshan Liu, and Gang Zhang. Mmat-1m: A large reasoning dataset for multimodal agent tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1484–1494, 2025. 2

[26] Zhi Gao, Yuntao Du, Xintong Zhang, Xiaojian Ma, Wenjuan Han, Song-Chun Zhu, and Qing Li. Clova: A closed-loop visual assistant with tool usage and update. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13258–13268, 2024.

[27] Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. *arXiv preprint arXiv:2412.15606*, 2024. 2

[28] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, 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*, pages 18995–19012, 2022. 2

[29] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[30] Hongcheng Guo, Shaosheng Cao, Boyang Wang, Lei Li, Liang Chen, Xinze Lyu, Zhe Xu, Yao Hu, Zhoujun Li, et al. Sns-bench: Defining, building, and assessing capabilities of large language models in social networking services. In *Forty-second International Conference on Machine Learning*, 2025. 2

[31] 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. *Advances in Neural Information Processing Systems*, 37:15646–15673, 2024. 2

[32] Nick Haber, Catalin Voss, and Dennis Wall. Making emotions transparent: Google glass helps autistic kids understand facial expressions through augmented-reaiity therapy. *IEEE Spectrum*, 57(4):46–52, 2020. 1

[33] Yuqi Hou, Zhongqun Zhang, Nora Horanyi, Jaewon Moon, Yihua Cheng, and Hyung Jin Chang. Multi-modal gaze following in conversational scenarios. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1186–1195, 2024. 2

[34] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[35] Xiaoke Huang, Ningsen Wang, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Medvlsynther: Synthesizing high-quality visual question answering from medical documents with generator-verifier lmms, 2025. 2

[36] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[37] Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. Smile: Multimodal dataset for understanding laughter in video with language models. *arXiv preprint arXiv:2312.09818*, 2023. 2

[38] Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, and Tatsuya Kawahara. An llm benchmark for addressee recognition in multi-modal multi-party dialogue. *arXiv preprint arXiv:2501.16643*, 2025. 3

[39] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 2

[40] Yidi Jiang, Ruijie Tao, Zexu Pan, and Haizhou Li. Target active speaker detection with audio-visual cues. *arXiv preprint arXiv:2305.12831*, 2023. 2, 4

[41] Swati Jindal and Roberto Manduchi. Contrastive representation learning for gaze estimation. In *Gaze Meets Machine Learning Workshop*, pages 37–49. PMLR, 2023. 2

[42] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*, 2022. 2, 4

[43] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid–hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4572–4581, 2024. 2

[44] 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. *Association for Computational Linguistics: ACL 2023*, 2023. 1, 2, 5

[45] 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*, pages 14585–14595, 2024. 2, 3, 5, 6

[46] Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, et al. Towards social ai: A survey on understand-

ing social interactions. *arXiv preprint arXiv:2409.15316*, 2024.

[47] Wanhua Li, Zibin Meng, Jiawei Zhou, Donglai Wei, Chuang Gan, and Hanspeter Pfister. Socialgpt: Prompting llms for social relation reasoning via greedy segment optimization. *Advances in Neural Information Processing Systems*, 37:2267–2291, 2024. 1, 2

[48] Xinpeng Li, Xiaojiang Peng, and Changxing Ding. Sequential interactive biased network for context-aware emotion recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2021. 2

[49] Xinpeng Li, Teng Wang, Jian Zhao, Shuyi Mao, Jinbao Wang, Feng Zheng, Xiaojiang Peng, and Xuelong Li. Two in one go: Single-stage emotion recognition with decoupled subject-context transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9340–9349, 2024. 2

[50] Xinpeng Li, Shijian Deng, Bolin Lai, Weiguo Pian, James M Rehg, and Yapeng Tian. Towards online multimodal social interaction understanding. *arXiv preprint arXiv:2503.19851*, 2025. 2, 3, 5, 6

[51] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning, 2025. 2

[52] Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, et al. Ov-mer: Towards open-vocabulary multimodal emotion recognition. *ICML 2025*, 2024. 2

[53] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *ICML 2025*, 2025.

[54] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22932–22941, 2023. 2

[55] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2

[56] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[57] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In *European conference on computer vision*, pages 126–142. Springer, 2024. 2

[58] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10631–10642, 2021. 2

[59] Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation, 2025. 2

[60] Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 735–745, 2022. 2

[61] Shuyi Mao, Xinpeng Li, Fan Zhang, Xiaojiang Peng, and Yang Yang. Facial action units as a joint dataset training bridge for facial expression recognition. *IEEE Transactions on Multimedia*, 27:3331–3342, 2025. 2

[62] Victoria Mingote, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. Audio-visual speaker diarization: Current databases, approaches and challenges. *arXiv preprint arXiv:2409.05659*, 2024. 2

[63] Shu Nakamura, Yasutomo Kawanishi, Shohei Nobuhara, and Ko Nishino. Deepoint: Visual pointing recognition and direction estimation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 20577–20587, 2023. 2

[64] OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. 2

[65] Liangyang Ouyang, Yifei Huang, Mingfang Zhang, Caixin Kang, Ryosuke Furuta, and Yoichi Sato. Multi-speaker attention alignment for multimodal social interaction. *arXiv preprint arXiv:2511.17952*, 2025. 2

[66] Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025. Accessed: 2025-01-24. 2

[67] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317, 2022. 2

[68] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 2, 4

[69] Chirag Raman, Jose Vargas Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. Conflab: A data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild. *Advances in Neural Information Processing Systems*, 35:23701–23715, 2022. 2

[70] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021. 4

[71] Fiona Ryan, Hao Jiang, Abhinav Shukla, James M Rehg, and Vamsi Krishna Ithapu. Egocentric auditory attention localization in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14663–14674, 2023. 2

[72] 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 Computer Vision and Pattern Recognition Conference, pages 28874–28884, 2025. 2

[73] Andrey Savchenko. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *International Conference on Machine Learning*, pages 30119–30129. PMLR, 2023. 2

[74] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model, 2025. 2

[75] Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, et al. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *arXiv preprint arXiv:2504.01886*, 2025. 2

[76] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20935–20946, 2023. 2

[77] Chao-Hong Tan, Jia-Chen Gu, and Zhen-Hua Ling. Is chatgpt a good multi-party conversation solver? *arXiv preprint arXiv:2310.16301*, 2023. 3

[78] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[79] Kimi Team. Kimi-VL technical report, 2025. 2

[80] Meta Llama Team. The llama 3 herd of models, 2024.

[81] V Team. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.

[82] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025. 2

[83] Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Object-aware gaze target detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21860–21869, 2023. 2

[84] Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-RTS: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28114–28128, Suzhou, China, 2025. Association for Computational Linguistics. 2

[85] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 2

[86] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 2

[87] Jianan Wei, Tianfei Zhou, Yi Yang, and Wenguan Wang. Nonverbal interaction detection. In *European Conference on Computer Vision*, pages 277–295. Springer, 2024. 2

[88] Eric Zhongcong Xu, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye, and Mike Zheng Shou. Ava-avd: Audio-visual speaker diarization in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3838–3847, 2022. 2

[89] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024. 2

[90] 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. 2, 3, 5, 6, 7

[91] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 5, 6, 7

[92] Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025. 2, 5, 6, 7

[93] 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*, pages 8807–8817, 2019. 2

[94] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2

[95] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning, 2025. 2

[96] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 2

[97] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 2

[98] Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*, 2025. 2

[99] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning, 2025. 2

[100] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34: 17616–17627, 2021. 2

[101] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research (TMLR)*, 2024. 2

[102] Chenxi Zhao, Jinglei Shi, Liqiang Nie, and Jufeng Yang. To err like human: Affective bias-inspired measures for visual emotion recognition evaluation. *Advances in Neural Information Processing Systems*, 37:134747–134769, 2024. 2

[103] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025. 2, 5, 6

[104] Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025. 2, 5, 6

[105] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024. 5

[106] Hengji Zhou, Lingxuan Huang, Si Wu, Lianghao Xia, Chao Huang, et al. Videoagent: All-in-one agentic framework for video understanding and editing. 2

[107] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019. 4

[108] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2