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

Abstract

In this paper, we present a new task, Omni-MMSI, which requires systems to understand multi-party multi-modal social interaction (MMSI) from raw audio-video. Unlike prior MMSI work, Omni-MMSI does not assume oracle identity-attributed cues such as speaker labels or participant boxes. The key difficulty is identity attribution: the system must determine who says what and to whom. We address this challenge with Omni-MMSI-R, a reference-guided pipeline that anchors participants with paired voice-image references, extracts identity-attributed verbal and non-verbal cues using lightweight tools, and performs two-step chain-of-thought reasoning. On Ego4D and YouTube subsets of Werewolf Among Us, Omni-MMSI-R improves average social-interaction accuracy over prior MMSI pipelines by 12.1% and 15.1%, respectively, while achieving much stronger identity attribution than large omni-LLMs.

1. Introduction

Multi-modal Multi-party Social Interaction Understanding (MMSI), which aims to interpret human behaviors in social situations, is fundamental for advancing socially intelligent AI systems [13, 11, 12, 14, 6]. As illustrated in Fig. 1, given raw audio-video input, the system must extract identity-attributed verbal and non-verbal social cues. For example, chronological utterances *[Player2]: All right. [Player4]: Okay. Do you need the script?*, with participant locations *[0.018, 0.736, 0.186, 0.992]* and *[0.668, 0.742, 0.875, 0.989]*, form the evidence needed to infer the underlying social interaction. These abilities are important for AI assistants that can perceive, reason over, and respond to human interactions in natural social scenarios [7, 5, 2].

Recent computer vision studies have improved MMSI with better representation alignment and conversation forecasting [12, 15]. However, they remain limited in scope because they assume identity-attributed social cues are available through oracle preprocessing. In realistic deployment, an assistant must instead operate directly on raw audio-video signals. Here, we introduce **Omni-MMSI**, which requires models to perceive who says what and where, and then infer the relevant social target or referent.

In Omni-MMSI, identity attribution is difficult in multi-party scenes with subtle motion, similar voices, and overlapping speech. First, off-the-shelf extractors such as Whisper and YOLO can provide transcriptions or detections,

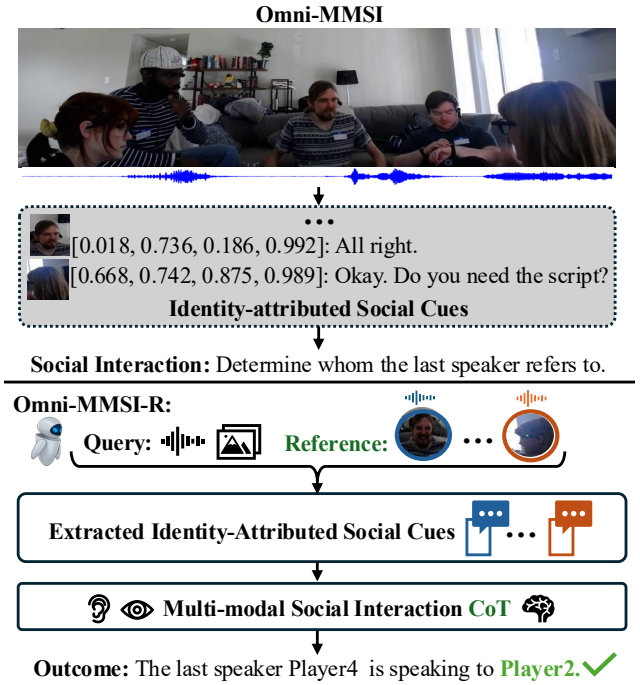

Figure 1. Omni-MMSI explores social interaction understanding in a multi-party social scene using only raw audio and video. To address the task, we propose Omni-MMSI-R, a reference-guided pipeline that anchors participants with paired voice-image references, extracts identity-attributed verbal and non-verbal cues using lightweight tools, and performs chain-of-thought reasoning.

but they do not solve person-level cross-modal binding in crowded scenes [16, 10]. Likewise, strong Omni-modal Large Language Models (Omni-LLMs) often transcribe speech or describe people correctly in isolation, yet still mismatch utterances and visible participants across modalities. Therefore, prior pipelines and Omni-LLMs degrade when transitioning from oracle input to raw input.

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 [9, 3]. As shown in Fig. 1, task-specific tools associate cues with references to generate accurate identity-attributed social cues. 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

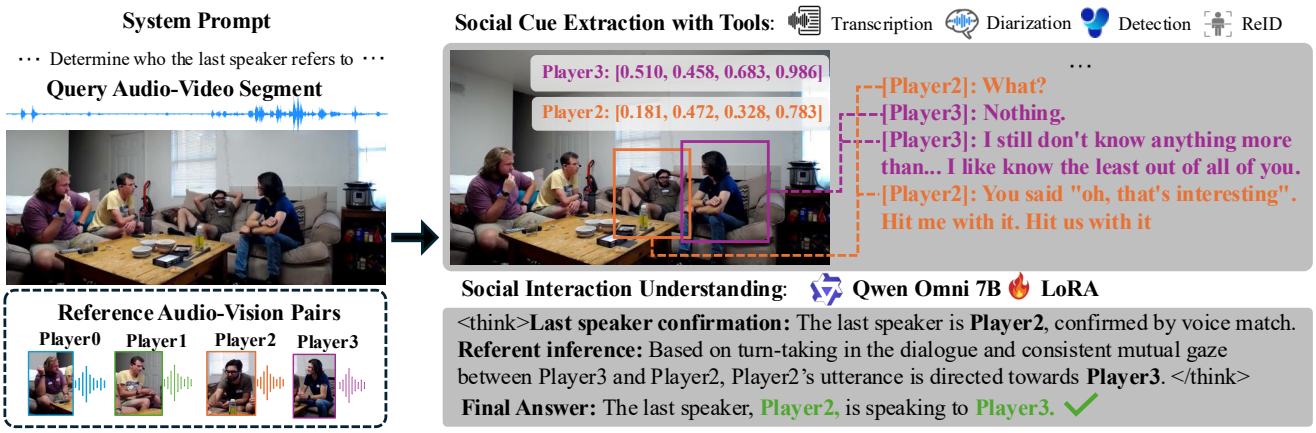

Figure 2. Omni-MMSI-R pipeline. Participant-specific voice-image references guide tool-based extraction of identity-attributed verbal and non-verbal cues, which are then combined with the raw query clip for two-step social reasoning.

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 [12]. Our method improves social interaction understanding over prior MMSI pipelines by 12.1% on Ego4D and 15.1% on YouTube, and exceeds advanced omni-LLMs by 23.7% on Ego4D and 18.9% on YouTube in identity attribution. These results show that reference guidance is effective for social interaction understanding on raw data.

2. Related Work

Multi-modal social interaction understanding. Recent MMSI benchmarks and models study social reasoning in conversations, games, and embodied settings by combining verbal and non-verbal evidence [13, 11, 12, 14, 6, 15]. However, these formulations typically assume identity-attributed transcripts, keypoints, boxes, or other oracle cues are already available. Omni-MMSI differs in that attribution is part of the task: the model must perceive who spoke what and where and then reason about to whom.

Omni-LLMs and multimodal reasoning. Recent omni-LLMs have made rapid progress on speech, image, and video understanding, including Qwen2.5-Omni, Phi-4-Multimodal, HumanOmni, OmniVinci, Qwen3-Omni, R1-Omni, and Gemini [18, 1, 22, 20, 19, 21, 4]. Yet strong general multimodal perception does not automatically solve cross-modal identity binding in crowded social scenes. Our results suggest that explicit references, tool-based attribution, and proper reasoning traces are a stronger recipe for Omni-MMSI than direct end-to-end prompting alone.

3. Approach

Task. Omni-MMSI targets MMSI on raw audio-visual input instead of relying on oracle cues. Specifically, we study two typical MMSI tasks [12, 15]: 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 goal of Omni-MMSI is to build a system f:

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

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

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

Pipeline. Figure 2 summarizes Omni-MMSI-R. The system augments the query with participant-specific reference pairs. First, the system extracts identity-attributed social cues using tools. For audio, Whisper transcribes the clip and SpeechBrain verifies each utterance against reference voices to assign speaker identity [16, 17]. For vision, YOLO detects participants in the final frame and OSNet matches each crop to the reference images [10, 24]. The result is a set of identity-attributed verbal and non-verbal social cues. Second, an Omni-LLM performs CoT reasoning on the query clip, reference pairs, and extracted cues to predict the social referent. The model is Qwen2.5-Omni-7B [18] fine-tuned with LoRA [8] in LLaMA-Factory [23]. The CoT supervision is structured with: (1) confirm the last speaker using voice, transcript, and visible evidence, and (2) infer the referent from verbal and non-verbal signals.

Data Curation. For each participant, we manually prepare a reference image and short voice clips that capture

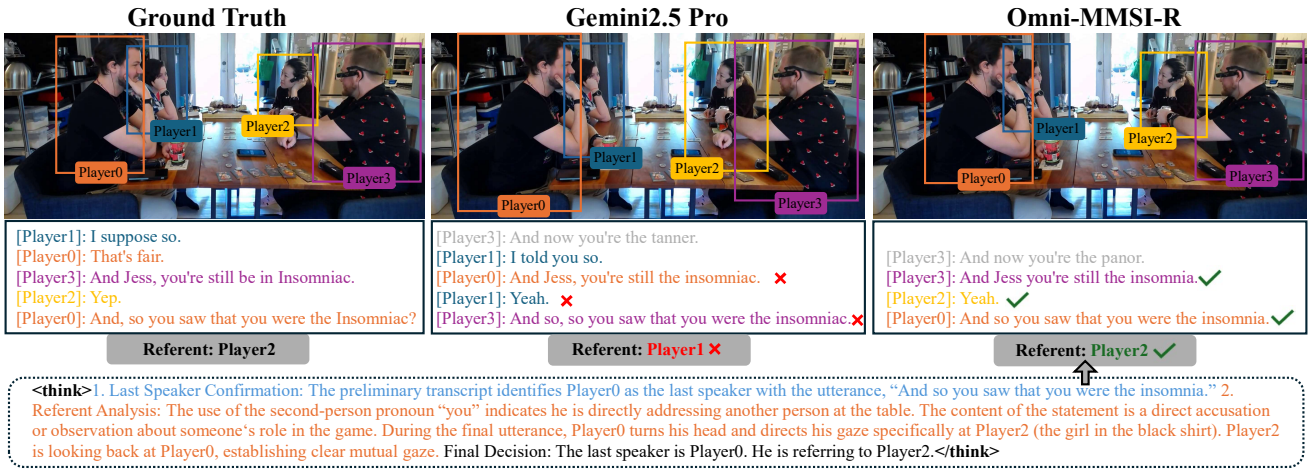

Figure 3. Qualitative comparison with Gemini 2.5 Pro. The baseline often mismatches utterances and visible participants, while Omni-MMSI-R better aligns verbal and non-verbal cues to the correct identity references and therefore predicts the referent more reliably.

Table 1. Average social-interaction accuracy (%) on Omni-MMSI.

Method	Ego4D	YouTube
Prior MMSI pipeline* [12, 15]	31.00	31.91
Gemini 2.5 Pro [4]	37.70	44.80
Gemini 2.5 Pro + references [4]	42.61	**48.72**
Omni-MMSI-R	**43.06**	47.04

representative appearance and vocal characteristics. Across the datasets used in this paper, the full benchmark contains 69 audio-visual identity profiles. For CoT training, we further curate short reasoning traces with a generate-filter-review process: a strong model, Gemini 2.5 Pro, proposes reasoning, we retain samples whose final answer matches the ground truth, and lightweight human review removes implausible traces. This gives interpretable and reliable CoT traces averaging 220 words per sample.

4. Experiments

Setup. We evaluate on the YouTube and Ego4D subsets of Werewolf Among Us [11]. Following prior MMSI work, we report average accuracy over STI and PCR. We also report identity attribution performance. The YouTube subset contains 3,255 STI samples and 2,679 PCR samples, while Ego4D contains 832 STI and 503 PCR samples. Query clips contain five dialogue turns on average and last about 14 seconds; reference audio clips are trimmed to five seconds. The omni-LLM backbone is fine-tuned for three epochs with LoRA rank 8 and learning rate 1×10^{-4}.

Main results. Table 1 shows that Omni-MMSI-R substantially improves social interaction accuracy over prior MMSI pipelines, gaining 12.1% on Ego4D and 15.1% on YouTube. For prior MMSI pipelines, we remove oracle information, such as speaker identity, from the input. Compared with Gemini 2.5 Pro without references, our method

Table 2. Identity attribution accuracy (%). Omni-MMSI-R is consistently stronger on both verbal and non-verbal attribution.

Method	Ego4D Avg.	YouTube Avg.
Gemini 2.5 Pro	35.64	58.04
Gemini 2.5 Pro + references	47.17	63.03
Omni-MMSI-R	**78.79**	**76.95**

Table 3. Ego4D ablations (%). Each component is helpful, and the full reference-guided two-step pipeline performs best.

Setting	Avg. Acc.
Query clip only	33.97
+ CoT supervision only	35.45
+ Reference pairs only	35.98
+ Tool-extracted cues only	39.44
+ References + cues + 2-step CoT	**43.06**

improves by 5.4% and 2.2%. When Gemini also receives references, social interaction accuracy becomes comparable, but its identity attribution remains notably weaker. Table 2 shows our method reaches 78.79% average attribution accuracy on Ego4D and 76.95% on YouTube, versus 47.17% and 63.03% for reference-enabled Gemini. Figure 3 shows the qualitative result: Gemini can bind speech and boxes to the wrong person, leading to incorrect final referent predictions even if the model broadly understands the scene. This gap supports the central claim: reliable social reasoning requires reliable identity attribution, and tools anchored by references provide that much more consistently than end-to-end prompting alone.

Ablations. Table 3 summarizes the main ablation trends on Ego4D. CoT reasoning alone helps over direct prediction, suggesting that explicit decomposition benefits complex social understanding even without references. Reference pairs alone also help, but extracted identity-attributed cues are stronger because they expose who-spoke-what and

where-each-person-is directly to the reasoning model. The full system is best, showing that references, tools, and structured reasoning play complementary roles.

5. Conclusion

Omni-MMSI reframes MMSI as social interaction understanding from raw audio-video, where identity attribution is part of the task rather than an oracle preprocessing step. Our results show that reference-guided attribution plus short structured reasoning is a strong recipe for realistic MMSI. An important next step is to move beyond the current game-based setting toward more open-ended multi-party videos with camera changes, partial visibility, and dynamically changing participants.

References

[1] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. 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] C. Breazeal, K. Dautenhahn, and T. Kanda. Social robotics. *Springer handbook of robotics*, pages 1935–1972, 2016.

[3] J. Clarke, Y. Gotoh, and S. 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.

[4] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. 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.

[5] M. Elsherbini, O. M. Aly, D. Alhussien, O. Amr, M. Fahmy, M. Ahmed, M. Adel, M. Fetian, M. Hatem, M. 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.

[6] X. Feng, L. Dou, M. Li, Q. Wang, H. Wang, Y. Guo, C. Ma, and L. Kong. A Survey on Large Language Model-Based Social Agents in Game-Theoretic Scenarios. *Transactions on Machine Learning Research (TMLR)*, 2025.

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

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[9] Y. Jiang, R. Tao, Z. Pan, and H. Li. Target active speaker detection with audio-visual cues. *arXiv preprint arXiv:2305.12831*, 2023.

[10] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, Z. Yifu, C. Wong, D. Montes, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*, 2022.

[11] B. Lai, H. Zhang, M. Liu, A. Pariani, F. Ryan, W. Jia, S. A. Hayati, J. Rehg, and D. Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. *Association for Computational Linguistics: ACL 2023*, 2023.

[12] S. Lee, B. Lai, F. Ryan, B. Boote, and J. M. Rehg. Modeling multimodal social interactions: New challenges and baselines with densely aligned representations. In *CVPR*, pages 14585–14595, 2024.

[13] S. Lee, M. Li, B. Lai, W. Jia, F. Ryan, X. Cao, O. Kara, B. Boote, W. Shi, D. Yang, et al. Towards social ai: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*, 2024.

[14] W. Li, Z. Meng, J. Zhou, D. Wei, C. Gan, and H. Pfister. Socialgpt: Prompting llms for social relation reasoning via greedy segment optimization. *Advances in Neural Information Processing Systems*, 37:2267–2291, 2024.

[15] X. Li, S. Deng, B. Lai, W. Pian, J. M. Rehg, and Y. Tian. Towards online multi-modal social interaction understanding. *arXiv preprint arXiv:2503.19851*, 2025.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[17] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

[18] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[19] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[20] H. Ye, C.-H. H. Yang, A. Goel, W. Huang, L. Zhu, Y. Su, S. Lin, A.-C. Cheng, Z. Wan, J. Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025.

[21] J. Zhao, X. Wei, and L. Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

[22] J. Zhao, Q. Yang, Y. Peng, D. Bai, S. Yao, B. Sun, X. Chen, S. Fu, X. Wei, L. Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025.

[23] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

[24] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019.