SOCIAL MICRO-STRATEGY UNDERSTANDING FROM MULTIMODAL BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding social interaction requires not only recognizing people's emotions, intentions, and beliefs, but also interpreting how people influence, maintain, and adjust ongoing interactions. However, existing social understanding benchmarks largely focus on mental states or coarse-grained behaviors, leaving fine-grained interactional functions and their multimodal behavioral cues underexplored. We introduce a structured framework for social micro-strategies that characterizes fine-grained interactional functions and grounds their interpretation in observable multimodal behavior. Based on this framework, we construct SMSBench, a benchmark spanning text, audio, and visual modalities, with detailed annotations across diverse social scenarios. A comprehensive evaluation of state-of-the-art multimodal large language models reveals substantial deficiencies in social micro-strategy understanding. We further fine-tune a 7B model using the training data to investigate task-specific adaptation. Experimental results show that non-expert humans slightly outperform current models overall, while fine-tuning yields clear improvements. We will release our dataset and code upon acceptance.

1 INTRODUCTION

Understanding human social interaction is a fundamental capability for intelligent systems operating in social environments (Zadeh et al., 2019). Beyond inferring people's beliefs, intentions, desires, and emotions, humans also interpret how others regulate ongoing interactions through speech, facial expressions, gaze, body movements, and other behavioral signals (Vinciarelli et al., 2009; Müller et al., 2023). A seemingly simple utterance or gesture can serve very different purposes depending on how it is expressed and situated in the surrounding interaction (Eo & Lim, 2026; Lai et al., 2023). We define a social micro-strategy as a fine-grained functional unit of social interaction, realized through observable verbal and non-verbal behavior, through which an actor locally regulates an ongoing interaction.

Recent advances in large language models and multimodal large language models have substantially improved emotion recognition (Zhang et al., 2026), intention inference (Zhang et al., 2024), theory-of-mind reasoning (Villa-Cueva et al., 2025), and social-event understanding (Niu et al., 2026). Recent benchmarks have further expanded toward grounded social reasoning, identity-aware interaction understanding, and broader social-behavior understanding (Li et al., 2026b; Ong et al., 2026). However, these tasks primarily characterize what people feel, believe, intend, or do, rather than how their behavior functions within an ongoing interaction. Related work on dialogue acts and communicative functions moves closer to this goal, but largely focuses on the linguistic function of individual utterances (Qamar et al., 2025). Real-world interaction instead unfolds through multimodal behavior spanning language, prosody, facial expressions, body movements, and responses from other participants (Vinciarelli et al., 2009; Poria et al., 2019). This leaves an important question underexplored: can models understand what a person's behavior is doing to the interaction?

Social micro-strategies are particularly challenging because they are fine-grained, context dependent, and behaviorally diverse (Lai et al., 2023; Qamar et al., 2025). Similar surface behaviors may realize different interactional functions, while the same function can be expressed through different combinations of verbal and non-verbal cues (Eo & Lim, 2026; Lai et al., 2023). Understanding them therefore requires jointly interpreting interactional context and multimodal behavioral evidence rather than recognizing isolated words and actions (Mathur et al., 2025; Poria et al., 2019).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

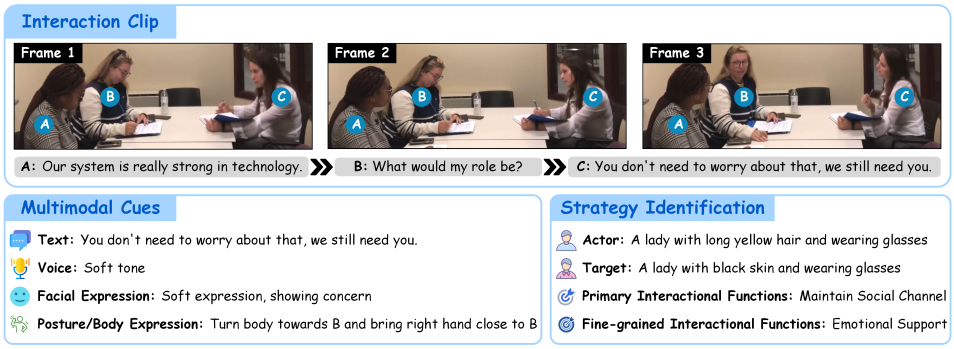


Figure 1: Example of Social Micro-Strategy Annotation.

To address this gap, we introduce the Social Micro-Strategy Benchmark (SMSBench), a fine-grained multimodal benchmark for social micro-strategy understanding. SMSBench contains 996 source videos, 4,051 interaction segments, and 5,600 annotated social micro-strategy instances across diverse social scenarios, together with multimodal behavioral evidence. Based on these annotations, as shown in Figure 1, we construct two complementary tasks: social micro-strategy understanding, which requires structured recognition of Actor, Target, Primary Interactional Function (PIF), and Fine-grained Interactional Function (FIF), and multimodal evidence grounding, which requires identifying the observable behavioral evidence supporting a specified strategy.

SMSBench is complementary to related benchmarks rather than subsuming them. MIntRec2.0 provides multimodal, multi-party dialogue with hierarchical fine-grained intent annotations (Zhang et al., 2024), whereas SMSBench represents each strategy instance as a directed Actor–Target event, allows multiple strategy instances within the same interaction segment, and associates each instance with supporting behavioral evidence. Social Genome provides grounded social reasoning with multimodal evidence traces (Mathur et al., 2025), while SMSBench focuses on structured recognition under a predefined interaction-function taxonomy together with channel-specific evidence grounding for a specified strategy instance. The resulting evaluation asks whether a model can jointly identify who acts on whom, determine the interactional function of that behavior, and locate the observable evidence supporting that interpretation. Table 1 summarizes these distinctions.

Our contributions are threefold: (i) we formulate social micro-strategy understanding through a layered interactional framework that captures fine-grained social functions and grounds their interpretation in observable multimodal behavior; (ii) we introduce SMSBench, a large-scale dataset with hierarchical social micro-strategy annotations and multimodal behavioral evidence; and (iii) we establish a comprehensive evaluation framework and systematically evaluate state-of-the-art LLMs to characterize their capabilities and limitations in social micro-strategy understanding.

2 Related Work

Multimodal Social Understanding and Reasoning. Multimodal social reasoning infers affect, mental states, interpersonal relations, and social events from linguistic, acoustic, and visual evidence (Kong et al., 2026). Existing studies have progressed from emotion recognition (Sun et al., 2026; Lian et al., 2025a;b) to Theory-of-Mind reasoning over beliefs, intentions, and perspectives (Villa-Cueva et al., 2025; Lee et al., 2025). Recent benchmarks further model grounded reasoning traces (Mathur et al., 2025; Kim et al., 2026a), mental–physical causal chains (Niu et al., 2026), and identity-attributed interactions in complex social scenes (Li et al., 2026b). However, these studies largely focus on participant states, interpersonal relations, or social events, and rarely treat the fine-grained interactional functions enacted through observable multimodal behavior as the primary object of understanding.

Social Behavior Understanding. Social Behavior Understanding focuses on recognizing observable social behaviors from facial expressions, gaze, gestures, and body movements (Müller et al., 2021; 2022; 2023; Kim et al., 2026b). Existing studies have progressed from subtle micro-gesture recognition in natural interviews (Liu et al., 2021b) to multi-person gesture understanding (Cao et al., 2025) and temporally localized micro-action detection (Li et al., 2025a). These works enable

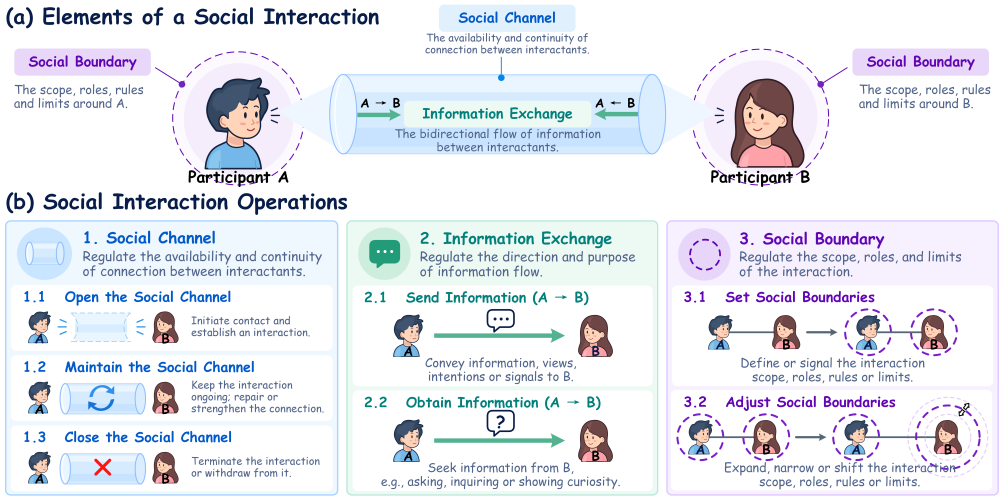


Figure 2: Social interaction elements and operations.

Table 1: Comparison with representative benchmarks for social strategy, intent, pragmatic, and grounded social understanding. T, A, and V denote text, audio, and visual modalities available in the benchmark data. Predefined Label Taxonomy denotes a fixed categorical label space for interaction-related functions, intents, strategies, dialogue acts, or related social behaviors.

Benchmark	Primary Task	Modality	Predefined Label Taxonomy	Actor–Target Annotation	Evidence Annotation
Werewolf Among Us	Persuasion strategy prediction	T/A/V	✓	✗	✗
PSAV	Persuasion strategy recognition	T/A/V	✓	✗	✗
AdsQA	Advertisement video understanding	T/A/V	✗	✗	✗
MentalManip	Mental manipulation analysis	T	✓	✗	✓
Persuasion Strategies in Advertisements	Persuasion strategy prediction	T/V	✓	✗	✓
Do LLMs Understand Dialogues?	Dialogue act classification	T	✓	✗	✗
PragMega+	Figurative language understanding	T	✗	✗	✗
MIntRec2.0	Multimodal intent recognition	T/A/V	✓	✗	✗
Social Genome	Grounded social reasoning	T/A/V	✗	✗	✓
SMSBench (Ours)	**Social micro-strategy understanding**	**T/A/V**	✓	✓	✓

increasingly fine-grained identification of what behavior occurs and when it occurs (Li et al., 2026a). However, they primarily characterize the form and temporal occurrence of observable behaviors, rather than the interactional functions these behaviors serve within an unfolding social interaction.

Social Interaction Strategies. Research on social interaction strategies studies how communicative behaviors serve strategic or pragmatic functions in interaction (Wang et al., 2019). Existing work has examined persuasion strategy prediction in interactive settings, as in Werewolf Among Us (Lai et al., 2023; Chawla et al., 2021; Jin et al., 2024), and extended strategy recognition to multimodal advertisement videos, as in PSAV (Bhattacharyya et al., 2023) and related persuasion benchmarks (Kumar et al., 2023). Closely related studies on dialogue acts (Qamar et al., 2025; Saha et al., 2020) and pragmatic meaning (Spinoso-Di Piano et al., 2025) further characterize the communicative functions conveyed by utterances. However, existing datasets (Liu et al., 2021a) typically focus on a specific family of strategies or communicative functions, such as persuasion or dialogue acts, rather than providing a unified account of fine-grained social micro-strategies across diverse natural interactions and multimodal behavioral cues.

3 Social Micro-Strategy Benchmark

3.1 Social Micro-Strategy Framework

Social interaction involves more than identifying what a person intends to do. The same social function can be expressed through different combinations of verbal and non-verbal behavior. We therefore conceptualize a social micro-strategy as a structured interactional event that connects who

acts on whom, what social function is being performed, and what observable multimodal evidence supports its interpretation. Each strategy instance is anchored by an Actor–Target relation and characterized by a hierarchical interactional function consisting of a PIF and a FIF. Together, Actor, Target, PIF, and FIF form the structured representation used for social micro-strategy understanding, while observable multimodal evidence grounds the interpretation in the behavior present in the interaction.

The interactional-function hierarchy captures what an actor is doing to shape the ongoing interaction. As illustrated in Figure 2, we organize social functions according to three basic elements of social interaction: the availability of the social channel, the direction and purpose of information exchange, and the boundaries governing participation. These give rise to seven PIFs, which are further refined into FIFs describing more specific interactional moves. This hierarchy provides a functional description of how participants open, maintain, redirect, constrain, and close social interaction. The complete interactional-function taxonomy is provided in Appendix A.

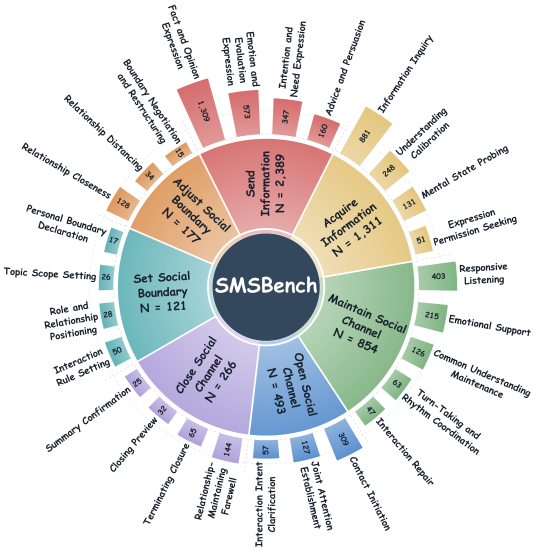


Figure 3: Social micro-strategy taxonomy and distribution in SMSBench.

Observable Multimodal Evidence further grounds each strategy interpretation in the behavior through which it becomes observable. We consider five evidence channels: linguistic content, prosody, facial behavior, body movement, and the target's response. These channels capture complementary behavioral cues that may jointly support the interpretation of an interactional function.

3.2 Dataset Construction

Video Collection. We collect publicly available human interaction videos from online platforms such as YouTube, covering seven diverse social interaction settings, including daily life, healthcare, counseling, business, interrogation, dating, and mediation. We prioritize videos with clear interpersonal interactions, identifiable participants, and observable verbal and non-verbal behaviors. Videos dominated by narration or monologue, severe editing, low-quality audio, unclear visual content, or insufficient interaction context are excluded. The resulting collection contains context-rich social interactions with observable linguistic, acoustic, facial, and bodily behaviors.

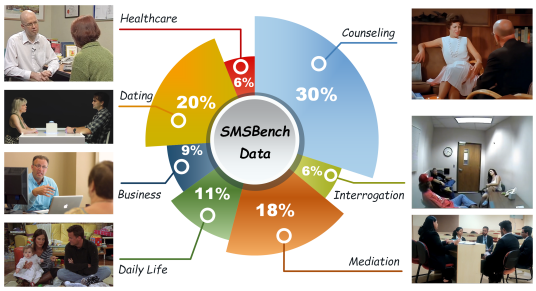


Figure 4: Overview of SMSBench, covering seven diverse real-world social contexts.

Video Processing. The collected long-form videos are divided into relatively self-contained interaction segments according to changes in participants, locations, topics, or interaction events. We transcribe the speech and align transcripts with video timestamps, while removing openings, endings, narration, and other non-interactive portions. Each resulting segment preserves sufficient surrounding context for interpreting the ongoing interaction. Within each segment, candidate regions containing observable social micro-strategies are identified for subsequent annotation. Interaction segments and strategy instances therefore operate at different temporal scales: the former provide interaction context, whereas the latter correspond to temporally localized social behaviors. A segment may contain one or multiple strategy instances. Further details are provided in Appendix B.

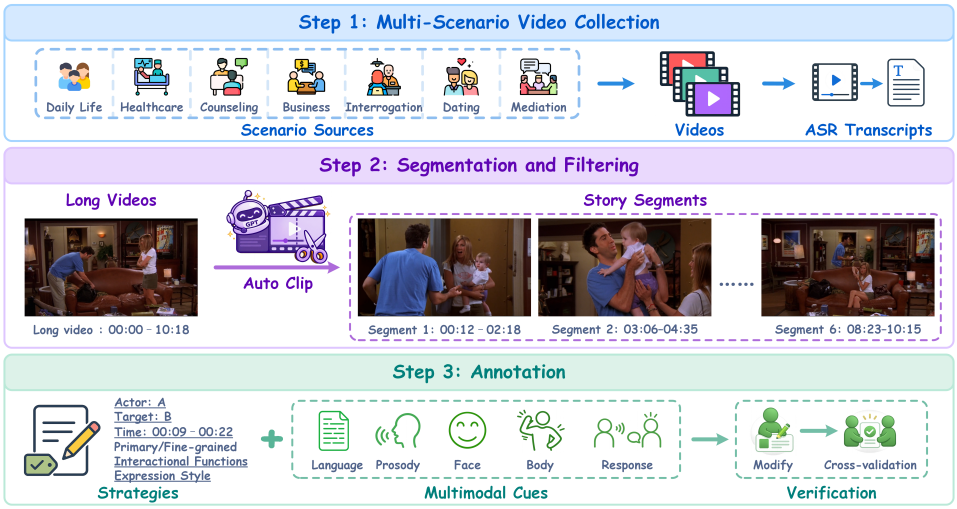


Figure 5: Annotation pipeline of SMSBench.

Data Annotation. For each interaction segment, annotators identify social micro-strategy instances and annotate their temporal boundaries, actor, target, PIF, FIF, IRS, and potential social risk. They additionally record observable evidence from five multimodal channels: linguistic content, prosody, facial expressions, body movements, and the target's response. Strategy labels are determined by their interactional function in the ongoing interaction rather than by isolated words, expressions, or physical actions. Each strategy span is defined as the shortest continuous interval preserving its complete realization, and multiple strategy instances may occur or overlap within the same interaction segment when distinct interactional functions co-occur. Given the fine-grained and context-dependent nature of the task, we employ a multi-stage quality-control procedure involving 12 annotators: nine primarily perform first-pass annotation, while three additional annotators also serve as senior reviewers responsible for verification and revision. All annotators receive guideline training and practice annotation before formal annotation, and rejected samples are returned for revision before final approval. Detailed quality-control statistics and an independent reproducibility analysis are provided in the appendix. SMSBench contains 996 source videos, 4,051 interaction segments and 5,600 annotated instances. Further dataset statistics are provided in Appendix C.

4 Experiments

4.1 Evaluation Settings

We evaluate a broad set of representative open- and closed-source multimodal large language models on SMSBench. All models receive the target segment together with timestamp-aligned subtitles. All models use a 1-fps visual sampling rate: frame-based models receive the sampled frames, whereas native audiovisual models receive the corresponding target video clip with synchronized audio and are configured to process its visual stream at 1 fps. We split the data at the source-video level into training, validation, and test sets with a target ratio of 7:1:2, ensuring no source-video overlap across splits. Unless otherwise stated, zero-shot evaluations and ablation studies use the full evaluation set, defined as the union of all three splits; model-adaptation experiments use the training and validation sets for optimization and model selection and report results on the held-out test set. Detailed split statistics are provided in the supplementary material.

For social micro-strategy understanding, models predict Actor, Target, PIF, and FIF for each strategy instance. We provide the annotated number of strategy instances per segment and require the same number of predictions, which are aligned with Gold instances using deterministic maximum-weight bipartite matching. Actor and Target are evaluated by accuracy, while PIF and FIF are evaluated by accuracy and Macro-F1. Overall denotes event-level four-field exact match.

For multimodal evidence grounding, each query is conditioned on a Gold strategy instance and asks the model to identify supporting evidence across Text, Prosody, Face, Body, and Target Re-

Table 2: Results on Social Micro-Strategy Understanding over the full evaluation set. Actor and Target are evaluated using accuracy, while PIF and FIF are evaluated using both accuracy and Macro-F1 (F1). Overall denotes event-level exact match over Actor, Target, PIF, and FIF. Best results are shown in bold, and second-best results are underlined. All values are reported as percentages.

Category	Model	Actor	Target	PIF		FIF		Overall
		Acc.	Acc.	Acc.	F1	Acc.	F1	
Open Source	Qwen2.5-Omni-3B	64.73	58.45	16.00	11.68	5.52	2.25	2.71
	Baichuan-Omni-1.5-7B	34.91	32.23	13.27	15.85	7.73	4.14	4.11
	Gemma 3n E4B-IT	57.86	54.54	47.38	27.65	24.11	9.81	11.86
	Qwen2.5-Omni-7B	65.18	59.05	25.11	22.68	15.62	6.44	9.45
	InternVL3.5-8B	67.59	60.20	43.21	29.19	24.79	8.37	14.62
	Qwen3.5-9B	70.43	65.27	60.00	46.19	43.21	22.71	28.54
	MiniCPM-o-4.5	70.46	64.54	40.39	28.47	29.55	13.40	17.23
	Qwen3-VL-235B	72.61	67.38	60.23	49.29	36.46	24.99	25.04
	Qwen3-Omni-30B	76.21	69.34	59.82	44.36	38.55	22.55	27.98
Closed Source	GPT-4o	72.38	67.38	57.54	47.43	39.66	24.71	28.62
	GPT-5.6-Sol	78.00	74.39	63.70	53.74	45.70	33.85	36.57
	GPT-6 Astra	<u>81.07</u>	**77.48**	<u>65.04</u>	**55.65**	<u>48.20</u>	**36.44**	<u>40.82</u>
	Gemini-3.1-Pro	**81.64**	<u>77.18</u>	**66.16**	<u>54.63</u>	**50.43**	<u>35.13</u>	**42.34**

sponse, independently of strategy recognition. Predicted evidence is evaluated against the annotated atomic Gold evidence by DeepSeek-V4.1-Flash using an LLM-as-a-judge protocol, with each applicable channel scored from 0 to 100. Channels without Gold evidence or unavailable under the corresponding input setting are excluded, whereas missing predictions for Gold-supported channels receive 0. We report the five channel-level scores and their average over applicable channels. The complete scoring protocol, handling rules, and evaluator validation are provided in the supplementary material. Details are provided in Appendix D.

4.2 Benchmark Results

Social Micro-Strategy Understanding. As shown in Table 2, current MLLMs exhibit substantial difficulty in recovering complete social micro-strategy events. Stronger models identify the interacting participants relatively well, with the best Actor and Target accuracies reaching 81.64 and 77.48, respectively. Performance decreases for interactional-function recognition, with the best PIF and FIF accuracies reaching 66.16 and 50.43. The corresponding best Macro-F1 scores are 55.65 for PIF and 36.44 for FIF, further indicating that less frequent interactional functions remain difficult to recognize. These results suggest that identifying who participates in an interaction is easier than determining the specific social function served by their behavior.

Closed-source models achieve the strongest Overall performance, with Gemini-3.1-Pro reaching 42.34, followed by GPT-6 Astra at 40.82 and GPT-5.6-Sol at 36.57. Among open-source models, Qwen3.5-9B achieves the highest Overall score of 28.54. Even the strongest model correctly recovers Actor, Target, PIF, and FIF jointly for only 42.34% of strategy events, showing that complete structured social micro-strategy understanding remains challenging despite relatively strong performance on individual fields. Detailed per-category results are provided in Appendix F.

Multimodal Evidence Grounding. As shown in Table 3, grounding social micro-strategies in observable multimodal evidence remains challenging, especially for non-textual behavioral cues. Text is consistently the strongest evidence channel across most models, with the best score reaching 79.79, whereas performance is substantially lower for facial behavior, body movement, and particularly Target Response, whose best score is only 57.32. Prosodic grounding also varies considerably across audiovisual models, ranging from 15.94 to 70.03, indicating large differences in the ability to exploit acoustic behavioral cues. Among closed-source models, Gemini-3.1-Pro achieves the highest Average point estimate of 66.64 and the strongest performance on Prosody, Face, and Target Response, while GPT-6 Astra performs best on Text and Body. Among open-source models with access to all five evidence channels, Qwen3-Omni-30B achieves the highest Average point es-

Table 3: Multimodal Evidence Grounding results over the full evaluation set. "–" indicates that the corresponding modality is unavailable. Average denotes the unweighted mean of the available channel-level scores and is not weighted by channel coverage.

Category	Model	Text	Prosody	Face	Body	Target Response	Average
Open Source	Qwen2.5-Omni-3B	67.02	46.36	38.97	30.46	27.39	42.04
	Baichuan-Omni-1.5-7B	32.41	15.94	17.50	18.65	11.02	19.10
	Gemma 3n E4B-IT	68.61	32.13	30.44	14.13	14.96	32.05
	Qwen2.5-Omni-7B	70.40	43.77	36.52	28.50	27.62	41.36
	InternVL3.5-8B	38.38	–	17.92	17.72	12.68	21.68
	Qwen3.5-9B	76.48	–	64.17	53.92	45.93	60.12
	MiniCPM-o-4.5	74.79	53.28	45.80	43.14	24.52	48.31
	Qwen3-VL-235B	73.01	–	58.93	56.48	48.70	59.28
	Qwen3-Omni-30B	75.93	<u>67.41</u>	59.51	52.89	44.43	60.03
Closed Source	GPT-4o	<u>78.43</u>	–	58.16	56.36	48.68	60.41
	GPT-5.6-Sol	76.83	–	<u>66.15</u>	<u>64.59</u>	<u>50.11</u>	64.42
	GPT-6 Astra	**79.79**	–	65.43	**64.97**	48.11	<u>64.58</u>
	Gemini-3.1-Pro	76.27	**70.03**	**66.85**	62.75	**57.32**	**66.64**

Table 4: Model adaptation results on the held-out test set.

Model	Actor	Target	PIF		FIF		Overall
	Acc.	Acc.	Acc.	F1	Acc.	F1	
Qwen2.5-Omni-7B	64.62	57.96	27.29	24.50	16.18	6.22	9.69
+ SFT	**71.11**	**65.96**	**68.18**	**50.79**	**50.76**	**25.79**	**33.42**
+ GRPO	64.89	58.22	48.44	28.49	29.07	5.52	16.18
+ SFT + GRPO	<u>70.58</u>	<u>65.07</u>	<u>67.73</u>	<u>47.55</u>	<u>49.51</u>	<u>21.77</u>	<u>31.20</u>
Qwen3-Omni-30B	76.00	66.67	61.87	46.28	39.38	20.14	27.38
Gemini-3.1-Pro	81.78	76.53	68.27	56.96	51.38	35.75	42.49

timate at 60.03. Overall, even the strongest models remain far from consistently grounding social interpretations across all behavioral channels.

Model Adaptation. We further investigate task-specific adaptation with Qwen2.5-Omni-7B. As shown in Table 4, SFT yields the largest gains, increasing Overall from 9.69 to 33.42 and substantially improving both PIF and FIF recognition. PIF accuracy increases from 27.29 to 68.18, while FIF accuracy rises from 16.18 to 50.76. GRPO also improves over the pretrained model, but provides considerably smaller gains than supervised training. Initializing GRPO from the selected SFT checkpoint preserves most of the SFT improvement, but does not further improve Overall performance.

Our validation diagnostics (Appendix H.1) show that the GRPO reward remains broadly aligned with strict correctness, while its dense partial-credit structure provides weaker discrimination among partially correct outputs. Meanwhile, token-level policy entropy decreases by 25.96% over GRPO optimization, indicating progressively reduced exploration. These observations help explain why continued reward optimization does not translate into further validation improvement. Notably, SFT enables the 7B model to outperform the zero-shot Qwen3-Omni-30B on PIF accuracy, FIF accuracy, and Overall, highlighting the benefit of task-specific supervision.

Human Comparison. Table 5 compares model performance with pooled non-expert human performance on the same evaluation subset. Seven university students participate, with each segment independently evaluated by three participants; the reported Human result pools all aligned event-level decisions before metric computation rather than using majority voting. Humans obtain the highest PIF and FIF accuracies, at 71.32 and 54.57, respectively, as well as the highest Overall point estimate of 44.34. Evaluated MLLMs, however, achieve higher scores on several individual metrics. Gemini-3.1-Pro obtains the highest Actor and Target accuracies, at 84.19 and 80.47, while GPT-6

Table 5: Comparison with human performance on 150 test segments.

Model	Actor	Target	PIF		FIF		Overall
	Acc.	Acc.	Acc.	F1	Acc.	F1	
Qwen3-Omni-30B	76.28	73.95	63.72	53.58	40.93	28.84	28.37
GPT-6 Astra	<u>82.79</u>	<u>80.00</u>	69.77	**57.81**	49.30	**37.57**	41.40
Gemini-3.1-Pro	**84.19**	**80.47**	68.84	55.82	50.23	34.66	<u>42.79</u>
Qwen2.5-Omni-7B + SFT	68.84	66.98	<u>70.23</u>	54.98	<u>54.42</u>	31.82	35.81
Human	81.09	75.97	**71.32**	<u>56.56</u>	**54.57**	<u>35.40</u>	**44.34**

Table 6: Modality ablation of Qwen3-Omni-30B over the full evaluation set. T, A, and V denote text, audio, and visual inputs, respectively.

Input	Actor	Target	PIF		FIF		Overall
	Acc.	Acc.	Acc.	F1	Acc.	F1	
T	74.84	66.83	56.76	43.07	35.39	20.74	24.73
A	75.99	68.78	58.58	**45.49**	35.57	21.63	25.36
V	71.36	63.49	44.53	29.37	25.93	8.33	16.72
V+T	70.61	63.89	57.13	43.00	35.52	19.45	23.21
V+A	**77.24**	**70.11**	<u>59.71</u>	44.05	<u>38.57</u>	21.45	<u>27.94</u>
A+T	75.68	68.06	57.92	<u>44.63</u>	36.52	**23.44**	26.09
V+A+T	<u>76.22</u>	<u>69.35</u>	**59.91**	44.41	**38.58**	<u>22.58</u>	**27.99**

Astra achieves the highest PIF and FIF Macro-F1 scores, at 57.81 and 37.57. These results show that models can match or exceed pooled non-expert human performance on particular dimensions, while the pooled Human result remains highest on the four-field Overall metric. We therefore interpret the Human result as non-expert performance under the benchmark interface rather than as a human ceiling. Further details are provided in Appendix I.

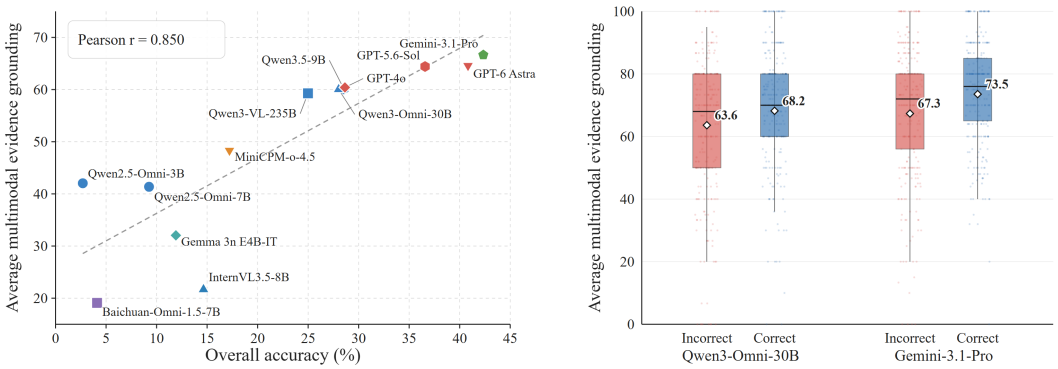
4.3 Modality Ablation

To examine how different modalities contribute to social micro-strategy understanding, we evaluate Qwen3-Omni-30B under seven input settings, as shown in Table 6. Text and audio alone already provide strong signals for interactional-function recognition, with PIF accuracies of 56.76 and 58.58, respectively, whereas visual input alone performs substantially worse on PIF, FIF, and Overall. The strongest participant identification is obtained with V+A, reaching 77.24 on Actor and 70.11 on Target, indicating that visual and acoustic cues provide complementary signals for identifying the interacting participants.

Combining all three modalities yields the highest PIF and FIF accuracies of 59.91 and 38.58, respectively, as well as the highest Overall score of 27.99. However, the class-balanced metrics exhibit somewhat different modality preferences: audio alone achieves the highest PIF Macro-F1 of 45.49, while A+T achieves the highest FIF Macro-F1 of 23.44. These results indicate that no single modality configuration dominates across all metrics and that different aspects of social micro-strategy understanding benefit from complementary multimodal cues.

4.4 Recognition–Gold-Conditioned Grounding Association

We further examine the relationship between strategy recognition and Gold-conditioned multimodal evidence grounding. At the model level, we compare Overall recognition performance with average grounding performance across all evaluated models. At the event level, for each canonical Gold strategy event, we define $R_i = 1$ if the aligned social micro-strategy prediction exactly matches the Gold event across Actor, Target, PIF, and FIF, and $R_i = 0$ otherwise. Let $H_i \in [0, 100]$ denote the judge-derived average grounding score obtained for the same Gold event. Importantly, grounding is evaluated independently of social micro-strategy understanding and is conditioned on the Gold strategy annotation rather than the model's predicted strategy. The analysis therefore measures whether



(a) Across-model association between Overall recognition performance and average Gold-conditioned multimodal evidence grounding. Each point represents one model; the dashed line shows the linear trend.

(b) Grounding distributions for strategy events grouped by whether they are incorrectly or correctly recognized under the four-field Overall criterion. White diamonds denote group means.

Figure 6: Association between social micro-strategy recognition and Gold-conditioned multimodal evidence grounding at the model and event levels. Recognition and grounding are strongly associated across models but remain only partially coupled at the event level.

recognition and grounding tend to co-vary, rather than whether a model can ground its own strategy prediction end to end.

Figure 6 reveals different association patterns at the model and event levels. Across models (Figure 6a), Overall recognition performance shows a strong positive association with average Gold-conditioned grounding performance, with a Pearson correlation of $r = 0.850$. Thus, models that are stronger at social micro-strategy recognition also tend to receive higher grounding scores when the target strategy is specified. At the event level (Figure 6b), however, the relationship is less pronounced. For Qwen3-Omni-30B, correctly recognized events obtain an average grounding score of 68.2, compared with 63.6 for incorrectly recognized events; for Gemini-3.1-Pro, the corresponding scores are 73.5 and 67.3. Despite these differences, the two distributions overlap substantially for both models. This indicates that an event that is easier to recognize is not necessarily easier to ground, and that a high grounding score for a specified Gold strategy does not imply successful inference of the complete Actor–Target–PIF–FIF structure. Recognition and Gold-conditioned evidence grounding therefore capture related but partially distinct capabilities.

5 Conclusion and Limitations

In this work, we introduce SMSBench, a multimodal benchmark for fine-grained social micro-strategy understanding in real-world interactions. We develop a structured framework that characterizes interaction participants and hierarchical interactional functions, together with the observable multimodal evidence supporting their interpretation. Extensive evaluations show that current MLLMs remain limited in recognizing fine-grained interactional functions, recovering complete strategy events, and grounding social interpretations in non-textual behavioral evidence. Modality ablations reveal complementary contributions from text, audio, and visual information, while task-specific adaptation substantially improves a smaller open-source model. Human evaluation further indicates a remaining gap in fine-grained interactional understanding. We hope SMSBench can support future research toward more reliable and evidence-grounded multimodal social intelligence.

Beyond the primary benchmark evaluation, we additionally explore Social Risk and Interactional Realization Style (IRS) as two complementary attributes of social interaction. Social Risk captures potential interactional risks, while IRS describes variation in how social functions are realized. We treat these analyses as supplementary rather than part of the primary benchmark evaluation, and report their annotation distributions, per-class performance, annotation reliability, and related analyses in the appendix.

AI USE STATEMENT

In this work, we used generative AI tools to code development and debugging, and drafting and editing portions of the manuscript. AI-assisted suggestions were reviewed and revised by the authors. All reported experiments were executed by the authors using the described experimental pipeline, and AI-generated code was inspected and tested before use. Literature-related claims and citations were checked against the corresponding original sources. Generative AI was not used to replace human annotators or to generate the final ground-truth annotations. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Our work adheres to the ICLR Code of Ethics. SMSBench is constructed from publicly accessible online videos depicting interpersonal interactions, including potentially sensitive scenarios such as healthcare, counseling, dating, and interrogation. We therefore carefully consider issues related to copyright, privacy, human annotation, and potential downstream misuse.

Source videos are acquired from publicly available YouTube content. At dataset construction time, all source videos are downloaded and stored locally, and all subsequent segmentation, annotation, preprocessing, and evaluation are performed on these fixed local copies rather than by accessing the live YouTube URLs. All media used in the reported experiments were downloaded during dataset construction and frozen before annotation and evaluation.

SMSBench focuses on observable interactional behavior rather than identifying individuals or inferring private personal attributes. We do not release uploader identities, account identifiers, contact information, or other platform metadata unnecessary for the research task. Because the source videos may contain identifiable individuals and sensitive interpersonal situations, users should treat the underlying media with appropriate care and comply with the relevant platform policies and applicable rights.

Human annotators participate in dataset annotation and human evaluation. Annotators are informed of the nature and research purpose of the tasks, and their annotation records are used only for research purposes. No personally identifying information about annotators is included in the released benchmark.

SMSBench is intended for research and benchmarking of multimodal social interaction understanding. We recognize that fine-grained social understanding technologies may have dual-use risks, including profiling, surveillance, or overconfident inference about individuals. The benchmark is not intended for high-stakes decision making about real people, and model predictions should not be treated as reliable assessments of individuals outside controlled research settings.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the SocialMicroBench construction process, annotation protocol, task definitions, evaluation metrics, prompts, data splits, model configurations, and training procedures in the main paper and appendices. Upon publication, we will release the benchmark annotations, data splits, source identifiers and temporal metadata, evaluation code, model prompts, and training configurations. For source videos whose redistribution is restricted, we provide the information necessary to retrieve the original content rather than redistributing the raw media. These materials are intended to enable reconstruction of the benchmark and reproduction of the reported experiments, subject to the continued availability and applicable terms of the original video sources.

REFERENCES

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

Aanisha Bhattacharyya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. A video is worth 4096 tokens: Verbalize videos to understand them in zero shot. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9822–9839, 2023.

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

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3167–3185, 2021.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Sugyeong Eo and Heuiseok Lim. Unveiling the limits of large language models in inferring pragmatic meaning from non-verbal responses. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 45304–45316, 2026.

Gemma Team. Gemma 3n. Google DeepMind, 2025. URL `https://ai.google.dev/gemma/docs/gemma-3n`.

Google. Gemini 3.1 Pro Preview. Gemini API Documentation, 2026. URL `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview`.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. Persuading across diverse domains: A dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1678–1706, 2024.

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

Min-Jae Kim, Jun-Yeong Moon, Mujeen Sung, and Gyeong-Moon Park. Open your model's eyes: Video and context-aware multimodal backchannel prediction. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 3738–3755, 2026b.

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

Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. Persuasion strategies in advertisements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 57–66, 2023.

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

Seungbeen Lee, Jinhong Jeong, Donghyun Kim, Yejin Son, and Youngjae Yu. Mind the motions: Benchmarking theory-of-mind in everyday body language. *arXiv preprint arXiv:2511.15887*, 2025.

Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. MMAD: Multi-label micro-action detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13225–13236, 2025a.

Kun Li, Jihao Gu, Fei Wang, Zhiliang Wu, Hehe Fan, and Dan Guo. MA-bench: Towards fine-grained micro-action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20118–20128, 2026a.

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

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-Omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025b.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014, 2025a.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. OV-MER: Towards open-vocabulary multimodal emotion recognition. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37015–37050, 2025b.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 3469–3483, 2021a.

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*, pp. 10626–10637, 2021b.

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

Philipp Müller, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Michael Dietz, Patrick Gebhard, Elisabeth André, and Andreas Bulling. Multimediate: Multi-modal group behaviour analysis for artificial mediation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4878–4882, 2021.

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

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

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

Keane Ong, Wei Dai, Carol Li, Dewei Feng, Hengzhi Li, Jingyao Wu, Jiaee Cheong, Rui Mao, Gianmarco Mengaldo, Erik Cambria, and Paul Pu Liang. Human Behavior Atlas: Benchmarking unified psychological and social behavior understanding. In *International Conference on Learning Representations*, 2026.

OpenAI. GPT-4o model. OpenAI API Documentation, 2024. URL https://developers.openai.com/api/docs/models/gpt-4o.

OpenAI. GPT-5.6 Sol model. OpenAI API Documentation, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol.

OpenAI. GPT-6 Astra model. OpenAI API Documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-6-astra.

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

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. Do LLMs understand dialogues? a case study on dialogue acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 26219–26237, 2025.

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

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4361–4372, 2020.

Cesare Spinoso-Di Piano, David Eric Austin, Pablo Piantanida, and Jackie CK Cheung. $(\text{RSA})^2$: A rhetorical-strategy-aware rational speech act framework for figurative language understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 20898–20938, 2025.

Yueru Sun, Yimeng Zhang, Haoyu Gu, Nuo Chen, Dong She, Xianrong Yao, Yang Gao, and Zhanpeng Jin. EmoMM: Benchmarking and steering MLLM for multimodal emotion recognition under conflict and missingness. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20351–20371, 2026.

Emilio Villa-Cueva, S. M. Masrur Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. MoMentS: A comprehensive multimodal benchmark for theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22591–22611, 2025.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5635–5649, 2019.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

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

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, Jieping Ye, and Pheng-Ann Heng. MME-Emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *International Conference on Learning Representations*, 2026.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *International Conference on Learning Representations*, 2024.

A Social Micro-Strategy Definition Details

A.1 Interactional Function Taxonomy

We organize social micro-strategies hierarchically into 7 PIFs and 27 FIFs. The central annotation question is which interactional state is primarily changed by the observed behavior. PIFs identify this broad interactional function, while FIFs specify the concrete local function through which it is realized. Table 7 summarizes the seven PIFs, while Table 8 provides the complete set of 27 FIFs.

Table 7: Primary interactional function taxonomy.

PIF	Definition
Open Social Channel	Establishes an interactional channel by bringing participants from a non-interacting or weakly interacting state into active interaction.
Set Social Boundary	Defines the framework within which the interaction proceeds, including interaction rules, participant roles, topic scope, and personal boundaries.
Send Information	Changes another participant's informational, evaluative, or action-related state by conveying facts, opinions, emotions, intentions, needs, advice, or persuasion.
Acquire Information	Changes the actor's own information state by eliciting facts, interpretations, emotions, intentions, concerns, or other information from another participant.
Maintain Social Channel	Preserves an ongoing interaction by maintaining conversational continuity, mutual understanding, emotional stability, and willingness to continue interacting.
Adjust Social Boundary	Changes an existing social boundary or relational configuration, including relational distance and the negotiation or restructuring of interactional boundaries.
Close Social Channel	Moves an ongoing interaction toward termination while managing conclusions, responsibilities, farewell, and possible future connection.

Table 8: Fine-grained interactional function taxonomy.

PIF	FIF	Definition
Open Social Channel	Contact Initiation	Initiates contact and establishes a minimal interactional connection with another participant.
	Joint Attention Establishment	Directs participants toward a shared object, event, topic, task, or other common focus of attention.
	Interaction Intent Clarification	Clarifies why the interaction is being initiated and what response or activity is expected.
Set Social Boundary	Interaction Rule Setting	Establishes procedures, participation rules, turn order, or other constraints governing how the interaction should proceed.
	Role and Relationship Positioning	Establishes the roles, identities, relationship positions, or interactional status under which participants engage.
	Topic Scope Setting	Defines what may be discussed, how far a topic may be pursued, or which content should remain outside the current interaction.
	Personal Boundary Declaration	States the actor's limits regarding privacy, time, emotion, resources, authority, or willingness to participate.
Send Information	Fact and Opinion Expression	Communicates facts, explanations, judgments, beliefs, positions, or other informational content to another participant.
	Emotion and Evaluation Expression	Communicates the actor's feelings, preferences, approval, concern, disappointment, or subjective evaluation.
	Intention and Need Expression	Communicates intended actions, requests, needs, desired support, or expected outcomes.
	Advice and Persuasion	Provides recommendations, reasons, or framing intended to influence another participant's judgment, choice, or behavior.
Acquire Information	Information Inquiry	Seeks facts, experiences, reasons, preferences, constraints, or other externally grounded information from another participant.
	Understanding Calibration	Checks whether the actor has correctly understood another participant through clarification, paraphrasing, or confirmation.
	Mental State Probing	Seeks another participant's emotions, intentions, concerns, preferences, hidden needs, or attitudes.
	Expression Permission Seeking	Seeks permission before entering a sensitive topic, requesting personal information, or continuing a potentially intrusive line of interaction.

PIF	FIF	Definition
Maintain Social Channel	Responsive Listening	Uses brief verbal or non-verbal responses to signal attention and support the other participant's continued expression.
	Emotional Support	Responds to another participant's emotional state through empathy, validation, reassurance, encouragement, or acknowledgment.
	Common Understanding Maintenance	Maintains shared understanding of the current topic, task, conclusion, responsibility, or next step.
	Turn-Taking and Rhythm Coordination	Coordinates turn exchange, pauses, continuation, interruptions, speaking order, and interaction pace.
	Interaction Repair	Restores interaction after misunderstanding, conflict, offense, interruption, or other breakdowns in interactional coordination.
Adjust Social Boundary	Relationship Closeness	Moves the relationship toward greater intimacy, trust, familiarity, or interpersonal involvement.
	Relationship Distancing	Reduces relational involvement, emotional investment, or responsibility while leaving some interactional connection available.
	Boundary Negotiation and Restructuring	Renegotiates or restructures existing interactional boundaries, responsibilities, permissions, or authority arrangements.
Close Social Channel	Summary Confirmation	Consolidates conclusions, agreements, responsibilities, or next steps as the interaction approaches closure.
	Closing Preview	Signals that the interaction is approaching its end while leaving room for final contributions or transition.
	Terminating Closure	Explicitly stops a topic, interaction, or subsequent engagement and prevents further continuation.
	Relationship-Maintaining Farewell	Ends the current interaction through thanks, acknowledgment, wishes, or farewell while preserving a positive future relationship.

A.2 Social Risk Taxonomy

Social Risk is annotated independently from PIF and FIF. It characterizes how an interactional function is realized rather than what function the behavior performs. The same FIF may therefore occur with No Apparent Risk or with one risk label. Risk is assigned only when observable evidence indicates that the interaction constrains autonomy, violates boundaries, distorts informed judgment, exploits relational or power asymmetries, undermines personal dignity, or uses interaction itself as a form of punishment. Firmness, disagreement, negative emotion, or interpersonal discomfort alone is not sufficient for assigning a risk label.

We define eight risk mechanisms. Boundary Intrusion captures continued or escalated interaction despite an expressed or implied boundary. Autonomy Compression reduces another participant's ability to refuse, negotiate, or choose independently. Emotional Manipulation uses affective burdens such as guilt, fear, shame, or anxiety as leverage, while Information Manipulation distorts judgment through concealment, selective presentation, exaggeration, or misleading framing. Relationship Coercion exploits relational obligations or intimacy, whereas Power Suppression exploits differences in status, authority, resources, or interactional control. Derogation and Humiliation attacks personal dignity through ridicule, contempt, or insults, and Interactional Punishment uses withdrawal, silence, cold treatment, or disconnection as a penalty. Table 9 summarizes these categories.

Table 9: Social Risk taxonomy.

Risk	Core Mechanism	Definition
Boundary Intrusion	Boundary violation	Continues or escalates an interaction despite an expressed or implied personal, topical, emotional, temporal, or physical boundary.
Autonomy Compression	Restricted choice	Reduces another participant's ability to refuse, delay, negotiate, or choose independently through pressure or constrained options.
Emotional Manipulation	Affective leverage	Uses guilt, fear, shame, anxiety, indebtedness, or related emotional burdens to influence another participant.
Information Manipulation	Distorted judgment	Influences judgment through concealment, selective presentation, exaggeration, misleading framing, or obscured intentions.
Relationship Coercion	Relational leverage	Exploits intimacy, loyalty, relational obligations, or prior favors to pressure another participant.
Power Suppression	Power asymmetry	Exploits status, authority, resources, evaluation power, or interaction control to restrict disagreement or refusal.
Derogation and Humiliation	Dignity attack	Undermines personal dignity through ridicule, contempt, insults, or related behavior.
Interactional Punishment	Withdrawal as punishment	Uses silence, withdrawal, cold treatment, or disconnection to penalize another participant.

A.3 Interactional Realization Styles Taxonomy

Beyond identifying what social micro-strategy is performed, we further characterize how the corresponding interactional function is realized in observable behavior. We refer to this dimension as Interactional Realization. Rather than representing additional social functions, realization styles capture systematic differences in how the same function is enacted. We distinguish five styles according to whether the social move is fully enacted, whether its function is explicitly recoverable, whether its interactional force is attenuated, and whether its progression depends on the recipient's response.

Table 10: Interactional realization styles for social micro-strategies.

Style	Definition
Direct	The social function is clearly and fully enacted without requiring substantial pragmatic inference or deliberate attenuation.
Mitigated	The function remains explicit, but its interactional force or pressure is deliberately softened.
Tentative	The actor probes the recipient's willingness or reaction before fully advancing the social move.
Implicit	The social function is not fully expressed on the surface and must be recovered through contextual or pragmatic inference.
Avoidant	The actor avoids fully enacting the social move through withholding, vagueness, deflection, minimal response, or silence.

The five styles are not treated as positions on a single directness continuum. Instead, they reflect different interactional properties of realization: Avoidant primarily concerns reduced enactment, Implicit reduced explicitness, Tentative recipient-contingent commitment, and Mitigated attenuated interactional force, while Direct corresponds to a fully enacted and explicitly recoverable realization without these additional mechanisms. Realization style is always judged relative to the annotated social micro-strategy and jointly grounded in linguistic, prosodic, facial, bodily, and response evidence; no individual cue, such as a soft tone or interrogative form, is sufficient by itself to determine the label.

B Dataset Composition and Statistics

The resulting corpus contains 996 source videos and 4,051 annotated segments, covering 15.39 hours of interaction. The average segment duration is 13.65 s and the median is 10.80 s. Most segments are relatively short: 78.77% are no longer than 20 seconds. Although single-strategy segments constitute the majority, nearly one third contain multiple strategies, reflecting the multifunctional nature of social interaction. Table 11 summarizes the temporal distribution and the number of strategies per segment.

Table 11: Segment length and number of strategies per segment.

Duration	Count	Ratio	# Strategies	Count	Ratio
$\leq$5 s	804	19.85%	1	2787	68.80%
5–10 s	1117	27.57%	2	1023	25.25%
10–20 s	1270	31.35%	3	206	5.09%
20–30 s	560	13.82%	4	28	0.69%
30–60 s	282	6.96%	5	6	0.15%
$>$60 s	18	0.44%	7	1	0.02%

Dataset split. SMSBench is partitioned into training, validation, and test sets with a target ratio of 7:1:2. The partition is performed at the source-video level, such that all segments and strategy instances derived from the same source video are assigned to the same split, thereby preventing content leakage across subsets. Because source videos contain different numbers of annotated segments and strategy instances, the resulting instance-level proportions are close to, but not exactly equal to, the target ratio. Table 12 summarizes the resulting split statistics.

Table 12: Dataset statistics for the source-video-level split.

Split	Source Videos	Segments	Strategy Instances
Train	697	2,835	3,910
Validation	100	406	565
Test	199	810	1,125
Total	996	4,051	5,600

The source-video proportions are 69.98%, 10.04%, and 19.98% for the training, validation, and test sets, respectively. The corresponding segment proportions are 69.98%, 10.02%, and 20.00%, while the strategy-instance proportions are 69.82%, 10.09%, and 20.09%. These small deviations from the target 7:1:2 ratio arise because source videos are kept intact during partitioning rather than splitting their derived segments or strategy instances across subsets.

The interactional functions exhibit a naturally imbalanced distribution. As shown in Table 13, Send Information is the most frequent PIF, followed by Acquire Information and Maintain Social Channel. At the FIF level, Fact and Opinion Expression and Information Inquiry are the two largest categories, while the remaining functions form a long-tailed distribution. We retain this naturally occurring distribution rather than artificially balancing the taxonomy.

Table 13: Hierarchical distribution of PIFs and FIFs.

PIF	Count	Ratio (%)	FIF	Count	Ratio (%)
Open Social Channel	493	8.80	Contact Initiation	309	5.52
			Joint Attention Establishment	127	2.27
			Interaction Intent Clarification	57	1.02
Set Social Boundary	121	2.16	Interaction Rule Setting	50	0.89
			Role and Relationship Positioning	28	0.50
			Topic Scope Setting	26	0.46
			Personal Boundary Declaration	17	0.30
Send Information	2,380	42.50	Fact and Opinion Expression	1,305	23.30
			Emotion and Evaluation Expression	569	10.16
			Intention and Need Expression	347	6.20
			Advice and Persuasion	159	2.84
Acquire Information	1,311	23.41	Information Inquiry	881	15.73
			Understanding Calibration	248	4.43
			Mental State Probing	131	2.34
			Expression Permission Seeking	51	0.91
Maintain Social Channel	854	15.25	Responsive Listening	403	7.20
			Emotional Support	215	3.84
			Common Understanding Maintenance	126	2.25
			Turn-Taking and Rhythm Coordination	63	1.12
			Interaction Repair	47	0.84
Adjust Social Boundary	177	3.16	Relationship Closeness	128	2.29
			Relationship Distancing	34	0.61
			Boundary Negotiation and Restructuring	15	0.27
Close Social Channel	264	4.71	Summary Confirmation	25	0.45
			Closing Preview	32	0.57
			Terminating Closure	65	1.16
			Relationship-Maintaining Farewell	142	2.54

Social Risk is substantially more imbalanced. Most strategy instances contain no apparent risk, while individual risk mechanisms occur only sparsely. Table 14 reports the detailed distribution. This reflects our decision to annotate risk as an independent property of social interaction rather than using risky interactions as a sampling criterion.

Interactional realization styles are also highly skewed toward explicit realizations. As shown in Table 15, Direct Expression accounts for the majority of strategy instances, followed by Mitigated Expression, whereas Implicit, Tentative, and Avoidant expressions are substantially less frequent.

Table 14: Distribution of Social Risk labels.

Risk	Count	Ratio
No Apparent Risk	5,497	98.16%
Autonomy Compression	36	0.64%
Derogation and Humiliation	22	0.39%
Boundary Intrusion	20	0.36%
Interactional Punishment	9	0.16%
Power Suppression	5	0.09%
Emotional Manipulation	4	0.07%
Relationship Coercion	4	0.07%
Information Manipulation	3	0.05%

Table 15: Distribution of IRS categories.

IRS	Count	Ratio
Direct Expression	4,748	84.79%
Mitigated Expression	680	12.14%
Implicit Expression	90	1.61%
Tentative Expression	34	0.61%
Avoidant Expression	48	0.86%

Finally, Table 16 reports the coverage of the five multimodal evidence channels. Text provides annotated evidence for nearly all strategy instances, while Body, Prosody, and Face provide complementary non-verbal evidence for approximately half of the data. Target Response is less frequent but still appears in more than one third of the instances. Since evidence channels are non-exclusive, their coverage rates are computed independently and therefore do not sum to 100%.

Table 16: Coverage of the five multimodal evidence channels.

Evidence	Instances	Coverage
Text	5,410	96.61%
Body	3,255	58.13%
Prosody	3,223	57.55%
Face	2,850	50.89%
Target Response	2,028	36.21%

C Annotation Protocol and Quality Control

We collect videos containing salient interpersonal interactions, prioritizing scenes with clear social functions and informative multimodal behaviors such as facial expressions, prosody, gestures, body movements, and partner responses. Annotators first watch the complete video to establish the overall context and participating social actors, and then select interactionally meaningful segments rather than exhaustively annotating every utterance. A segment is defined as a temporally localized interaction unit containing sufficient evidence for one Social Micro-Strategy. Its boundary covers the behavior realizing the strategy. Purely transitional, repetitive, or insufficiently informative content is not prioritized. Segments and strategy instances are not necessarily one-to-one: a segment may contain multiple simultaneous or sequential interactional functions, which are retained as separate strategy instances.

For each strategy instance, annotators identify the Actor and Target and assign the corresponding PIF, FIF, IRS, and Social Risk labels. The Actor is the participant who initiates the relevant interactional change, while the Target is the individual or group toward whom it is primarily directed. Each annotation is further grounded in up to five evidence channels: Text, Prosody, Face, Body, and Target Response. Annotators record only cues that meaningfully support the strategy interpretation, with at least one evidence channel required. Text captures diagnostic linguistic content; Prosody covers intonation, loudness, speaking rate, stress, and pauses; Face includes facial expressions and gaze; Body includes posture, orientation, gestures, and movement; and Target Response captures verbal or non-verbal reactions from the recipient.

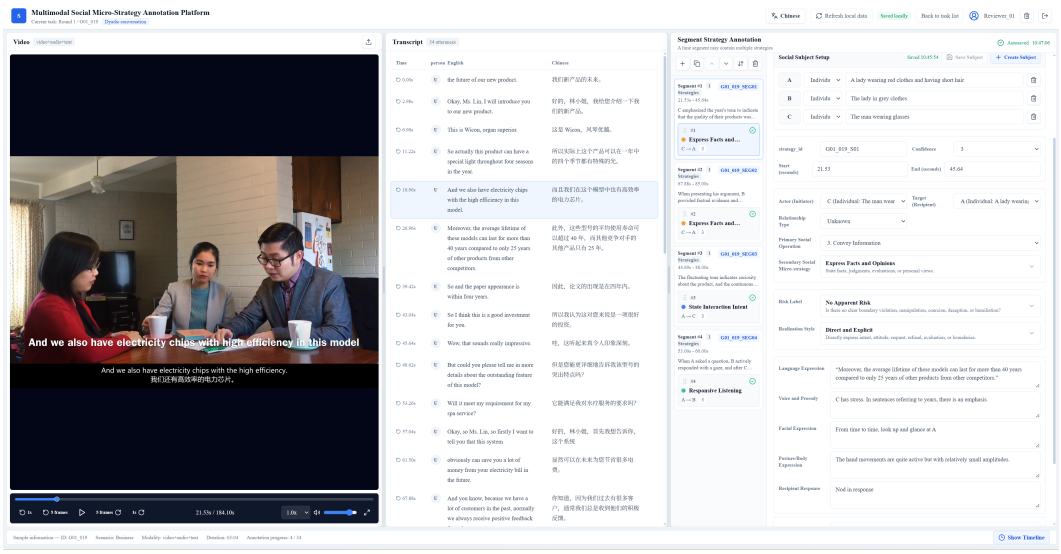


Figure 7: Annotation interface used for SMSBench.

Annotation follows a context-first quality-control procedure. Before formal annotation, all annotators complete a small set of practice samples, which are reviewed to clarify task definitions, decision rules, and commonly confused categories. During formal annotation, each sample is first annotated by one annotator and then undergoes an initial quality check by the authors. Among the 12 annotators, nine primarily perform first-pass annotation, while three additional annotators also serve as senior reviewers and verify the annotations, revising incorrect, incomplete, ambiguous, or insufficiently supported labels when necessary. The review stage is used for verification and adjudication rather than consensus-based sample filtering, and ambiguous cases are resolved according to the annotation guidelines instead of being discarded. Reviewers also check temporal coverage, Actor–Target consistency, functional compatibility, and whether the recorded multimodal evidence sufficiently supports the assigned labels. The annotation interface is shown in Figure 7.

The production workflow provides additional evidence of the intensity of this quality-control process. Among the 996 source samples, 995 contain review records, with 2,019 review decisions in total. In particular, 784 samples (78.71%) were rejected at least once and returned for revision before final approval. These statistics characterize reviewer-based verification and adjudication rather than independent inter-annotator agreement.

Table 17: Audit statistics for the annotation review process.

Audit Item	Count / Rate
Source samples	996
Samples with review history	995 (99.90%)
Total review decisions	2,019
Rejection decisions	938
Approval decisions	1,081
Samples rejected at least once	784 (78.71%)
Samples with reviewer comments	560 (56.22%)
Samples receiving final approval	996 (100.00%)

Because reviewer verification is not equivalent to independent double annotation, we additionally assess taxonomy reproducibility using the human evaluation subset, in which each segment is independently evaluated by three participants. After aligning predictions to common Gold event slots, we compute pairwise agreement and nominal Krippendorff's α. As shown in Table 18, PIF achieves $\alpha = 0.591$ and FIF achieves $\alpha = 0.473$, providing moderate evidence of independent reproducibility for the core functional taxonomy. In contrast, Risk and IRS obtain substantially lower chance-corrected agreement ($\alpha = 0.091$ and -0.015, respectively). Their highly skewed distributions and sparse minority categories contribute to this instability, but the current evidence is insufficient to

establish reliable independent reproducibility for these attributes. We therefore treat Risk and IRS as complementary annotations rather than core functional taxonomy dimensions. Accordingly, six-field Overall should be interpreted as a strict composite measure that additionally requires agreement on these complementary attributes, rather than as a standalone measure of the core functional taxonomy.

Table 18: Independent human-annotation reliability on the human-evaluation subset. Krippendorff's α uses nominal distance, with 95% confidence intervals obtained by segment-level cluster bootstrap.

Field	Pairwise Agreement	Krippendorff's α	95% CI
Actor	78.76	0.645	[0.562, 0.716]
Target	74.88	0.599	[0.516, 0.677]
PIF	73.33	0.591	[0.516, 0.659]
FIF	57.21	0.473	[0.415, 0.534]
Risk	97.21	0.091	[-0.014, 0.325]
IRS	95.35	-0.015	[-0.023, -0.008]

D Task Definitions and Prompt Templates

We use a fixed evaluation protocol across all models. Social micro-strategy understanding and multimodal evidence grounding are issued independently, such that evidence grounding does not depend on the model's prediction in social micro-strategy understanding.

Social micro-strategy understanding. Given a target interaction segment, the model predicts structured social micro-strategy instances consisting of Actor, Target, PIF, and FIF. The Gold number of strategy instances K is provided, and the model is required to return exactly K predictions.

Social Micro-strategy Understanding Prompt Template

Analyze the provided media for **Social Micro-Strategy Understanding**.
Social subjects: {{SOCIAL_SUBJECT_LIST}}
Available modalities: {{AVAILABLE_MODALITIES}}
Context interval: {{CONTEXT_INTERVAL}}
Target interval: {{TARGET_INTERVAL}}
Number of strategies: {{STRATEGY_SET_COUNT}}
Return exactly this number of strategy instances. Each instance must contain Actor, Target, PIF, and FIF. Return JSON only.

Multimodal evidence grounding. Each request is conditioned on a fixed Gold strategy instance, including its Actor, Target, PIF, FIF, and FIF definition. The model identifies observable evidence supporting the given strategy without revising its labels.

Multimodal Evidence Grounding Prompt Template

Analyze the provided media for **Multimodal Evidence Grounding**.
Social subjects: {{SOCIAL_SUBJECT_LIST}}
Available modalities: {{AVAILABLE_MODALITIES}}
Context interval: {{CONTEXT_INTERVAL}}
Target interval: {{TARGET_INTERVAL}}
Gold strategy: Actor={{ACTOR_ID}}, Target={{TARGET_ID}}, PIF={{PIF}}, FIF={{FIF}}.
FIF definition: {{FIF_DEFINITION}}
Extract observable evidence from five channels: `text`, `prosody`, `face`, `body`, and `target_response`. Return JSON only.

The complete prompts additionally include the full PIF and FIF label definitions, decision boundaries, modality constraints, and JSON validation rules, which are released with the benchmark implementation.

Evidence grounding scoring protocol. Evidence grounding is evaluated independently for five channels: Text, Prosody, Face, Body, and Target Response. For each strategy instance, the evaluator receives the predicted evidence, the annotated atomic Gold evidence, the channels available under the corresponding model input setting, and the channels for which the Gold annotation contains non-empty evidence. The evaluator does not inspect the original video or audio and assesses each prediction solely against the provided Gold evidence. A channel is scored only when it is both available under the corresponding input setting and supported by non-empty Gold evidence; all other channels receive a null score and are excluded from aggregation. Thus, the absence of Gold evidence for a channel indicates only that the benchmark does not evaluate that channel for the given instance and does not imply that the corresponding behavior is absent from the original video. For every scored channel, the evaluator assigns a score from 0 to 100 based on five criteria. Semantic consistency measures whether the prediction describes the same core behavior, utterance, state, or interactional cue as the Gold evidence; semantically equivalent paraphrases and cross-language expressions are accepted without requiring lexical overlap. Coverage measures whether the prediction captures the important and independently meaningful information contained in the Gold evidence, with partial credit assigned for partial coverage. Contradictions and attribution errors are penalized when a prediction reverses the meaning of the Gold evidence, attributes an action to the wrong participant, misidentifies the direction of an interaction, or otherwise conflicts with the reference. Channel consistency requires the evidence to belong to the channel currently being evaluated, such that correct information from another channel does not compensate for missing or incorrect evidence in the current channel. Finally, only evidence supported by the Gold annotation can receive credit. Details that are not supported by the Gold evidence are treated as unsupported content and reduce the score according to their extent and severity; incorrect attribution, cross-channel content, and irrelevant details are penalized accordingly.

The evaluator uses the score anchors in Table 19. Intermediate scores between adjacent anchors are permitted when the quality of a prediction falls between their corresponding descriptions.

Table 19: Score anchors for evidence grounding evaluation.

Score	Interpretation
100	The key evidence is correct and complete, with no substantive error.
80	The prediction is mostly correct, with only minor omissions or imprecision.
60	The prediction contains clearly valid evidence but has important omissions or partial errors.
40	Only limited evidence matches the Gold evidence, with substantial omissions or errors.
20	The prediction contains very little relevant evidence.
0	The prediction is empty for a supported channel, completely unrelated, or contradicts the core Gold evidence.

In addition to these score anchors, the deterministic rules in Table 20 are applied to missing, unsupported, invalid, or contradictory predictions.

Table 20: Deterministic rules for evidence grounding evaluation.

Situation	Scoring Rule
The Gold annotation contains no evidence for the channel	The channel receives a null score and is excluded from aggregation.
The channel is unavailable under the corresponding model input setting	The channel receives a null score and is excluded from aggregation.
The Gold annotation contains evidence, but the prediction is empty	The channel receives a score of 0.
The prediction contains details not supported by the Gold evidence	Unsupported details do not receive credit and reduce the channel score according to their extent and severity. Incorrect attribution, cross-channel content, and irrelevant details are penalized accordingly.
The prediction describes evidence from another channel	The misplaced information does not count as valid evidence for the channel being evaluated and lowers the score according to its severity.
The prediction contradicts the Gold evidence	The score is reduced according to the severity of the contradiction; a contradiction of the core meaning or interaction direction may receive a score of 0.
The model response is missing or fails output-format validation	Every Gold-supported and input-available channel receives a score of 0.
Evaluation fails because of an infrastructure error	The instance is excluded pending rerun rather than being assigned a score of 0.

For each model and evidence channel, we first compute a channel-level grounding score by averaging over all instances for which that channel is both available under the model input setting and supported by non-empty Gold evidence. Null instances are excluded from this channel-level denominator. The aggregate grounding score reported in Table 3 is then computed as the unweighted arithmetic mean of the resulting channel-level scores over the evidence channels available to that model. Thus, channels contribute equally to the aggregate regardless of how many Gold-supported instances they contain; the coverage counts in Table 16 determine the number of instances used to estimate each channel-level score but are not used as aggregation weights.

Cross-model grounding aggregation. Let G_m^c denote the channel-level grounding score of model m for evidence channel c, computed as described above. Our primary availability-aware aggregate is the unweighted mean of the channel-level scores available to that model:

$$G_m^{\mathrm{Avg}} = \frac{1}{|\mathcal{C}_m|} \sum_{c \in \mathcal{C}_m} G_m^c,$$

where $\mathcal{C}_m$ is the set of evidence channels available under the model's input setting. Consequently, Prosody contributes as one equally weighted channel for models with audio input but is omitted for models evaluated without audio. This is a macro-average over channel-level scores rather than an instance-weighted average across all scored channel–instance pairs.

Because the set of contributing channels differs across models, we additionally construct a modality-consistent Common-4 Average as a sensitivity measure. The Common-4 Average is the unweighted mean of Text, Face, Body, and Target Response, the four evidence channels available to all evaluated models:

$$G_m^{\mathrm{C4}} = \frac{1}{4} \left(G_m^{\mathrm{Text}} + G_m^{\mathrm{Face}} + G_m^{\mathrm{Body}} + G_m^{\mathrm{TargetResponse}} \right).$$

We retain the availability-aware average for the primary cross-model analysis and use the Common-4 Average only as a complementary sensitivity check.

Evidence grounding evaluator validation. Since Evidence Grounding relies on an LLM-based semantic evaluator, we additionally examine its agreement with independent human raters and its stability under repeated evaluation. We construct a stratified subset of 200 channel-level prediction–reference pairs, with 40 examples from each of the five evidence channels: Text, Prosody, Face, Body, and Target Response. Three independent raters, denoted as Person 1, Person 2, and Person 3, score exactly the same 200 examples using the 0–100 evidence-grounding scoring protocol defined above. The identity of the evidence-generating model, the DeepSeek-V4.1-Flash score, and the scores assigned by the other raters are hidden during evaluation.

The three raters exhibit substantial agreement. Their single-rater reliability is $\mathrm{ICC}(2,1) = 0.784$ (95% CI: $[0.728, 0.830]$), while the reliability of their aggregated ratings reaches $\mathrm{ICC}(2,k) = 0.916$ (95% CI: $[0.889, 0.936]$). Table 21 reports both pairwise agreement among the three raters and agreement between their consensus score and DeepSeek-V4.1-Flash.

Table 21: Agreement among the three independent raters and between their consensus and DeepSeek-V4.1-Flash on the 200-sample Evidence Grounding validation subset. Person Consensus denotes the mean score of Persons 1–3.

Comparison	Pearson r	Spearman ρ	MAE
Person 1 vs. Person 2	0.895	0.904	10.65
Person 1 vs. Person 3	0.818	0.841	20.40
Person 2 vs. Person 3	0.843	0.840	15.55
Person Consensus vs. DeepSeek	0.829	0.840	13.93

The three raters show consistently positive pairwise agreement, with Pearson correlations ranging from 0.818 to 0.895 and Spearman correlations ranging from 0.840 to 0.904. More importantly, their consensus shows strong agreement with the DeepSeek-V4.1-Flash evaluator used in the benchmark, with Pearson $r = 0.829$ (95% CI: $[0.771, 0.877]$) and Spearman $\rho = 0.840$ (95% CI: $[0.780, 0.885]$). The corresponding MAE is 13.93 points. Across the 200 examples, 51.5% of DeepSeek scores fall within 10 points of the person consensus and 84.0% fall within 25 points. These results indicate

that the primary evaluator is strongly aligned with independent human judgments in its relative assessment of evidence-grounding quality.

We further examine agreement separately for each evidence channel in Table 22.

Table 22: Channel-level validation of the Evidence Grounding evaluator. Correlations and MAE compare the consensus of the three independent raters with DeepSeek-V4.1-Flash.

Channel	ICC$(2,1)$	Pearson r	Spearman ρ	MAE
Text	0.750	0.911	0.902	9.33
Prosody	0.725	0.825	0.878	15.25
Face	0.786	0.769	0.772	14.96
Body	0.707	0.793	0.782	13.21
Target Response	0.692	0.654	0.728	16.88

Agreement remains positive across all five evidence channels. Text shows the strongest correspondence between the person consensus and DeepSeek-V4.1-Flash, with Spearman $\rho = 0.902$, followed by Prosody with $\rho = 0.878$. Face and Body also show substantial positive agreement, while Target Response exhibits the lowest agreement, with $\rho = 0.728$ and an MAE of 16.88. The positive correlations across all channels indicate that the overall agreement is not driven by a single evidence modality.

We separately assess the repeatability of DeepSeek-V4.1-Flash by rescoring the same 200-sample validation subset three times under a fixed evaluation setting that is held constant across the three runs. Results are reported in Table 23.

Table 23: Repeatability of DeepSeek-V4.1-Flash over three independent evaluations of the same 200 channel-level samples.

Comparison	Pearson r	Spearman ρ	MAE	Exact Match (%)
Run 1 vs. Run 2	0.981	0.975	1.90	92.0
Run 1 vs. Run 3	0.962	0.957	3.13	88.5
Run 2 vs. Run 3	0.969	0.963	2.53	90.5

The three repeated runs yield mean scores of 40.45, 40.60, and 40.58, respectively, with ICC$(2,1) = 0.971$. Pairwise Pearson correlations range from 0.962 to 0.981, Spearman correlations range from 0.957 to 0.975, and MAEs remain between 1.90 and 3.13 points. Exact score agreement ranges from 88.5% to 92.0%, and the nearly identical run-level means indicate negligible score drift across repeated evaluations. These results demonstrate high repeatability of DeepSeek-V4.1-Flash under a fixed evaluation setting.

Overall, the validation provides complementary evidence for the reliability of the Evidence Grounding evaluator. DeepSeek-V4.1-Flash shows strong agreement with the consensus of three independent raters in relative evidence-quality assessment and remains highly stable across repeated evaluations. We therefore use the resulting scores primarily for comparative analysis across models and evidence channels rather than interpreting individual 0–100 scores as an absolutely calibrated semantic-quality scale.

E Model and Evaluation Reproducibility Details

This section provides the implementation details that can affect reproduction of the reported evaluation results, including the exact model variants, failure handling, and deterministic alignment procedure.

Model versions and failure handling. Table 24 reports the exact model checkpoints or API identifiers together with prediction coverage for social micro-strategy understanding. All reported metrics use the complete evaluation denominator of 4,051 target segments and 5,600 Gold strategy instances. A prediction is considered invalid when no usable response is returned, the response cannot be parsed into the required structured schema, the predicted number of strategy instances violates the provided Gold-count requirement, or a predicted label falls outside the permitted taxonomy.

Invalid responses are represented by empty prediction sets rather than removed from evaluation; consequently, all associated Gold strategy instances remain in the denominator and are counted as incorrect. We do not selectively retry failed examples or compute metrics only over successful requests.

Table 24: Exact model variants and prediction coverage for social micro-strategy understanding on the 4,051-segment evaluation set. Invalid or missing predictions remain in the formal evaluation denominator.

Model	Checkpoint / API ID	Valid Predictions	Failed / Missing
Qwen2.5-Omni-3B (Xu et al., 2025a)	Qwen2.5-Omni-3B	4,048	3
Baichuan-Omni-1.5-7B (Li et al., 2025b)	Baichuan-Omni-1.5-7B	3,888	163
Gemma 3n E4B-IT (Gemma Team, 2025)	Gemma-3n-E4B-it	4,046	5
Qwen2.5-Omni-7B (Xu et al., 2025a)	Qwen2.5-Omni-7B	4,051	0
InternVL3.5-8B (Wang et al., 2025)	InternVL3.5-8B-HF	3,960	91
Qwen3.5-9B (Qwen Team, 2026)	Qwen3.5-9B	4,048	3
MiniCPM-o-4.5 (Cui et al., 2026)	MiniCPM-o-4.5	3,929	122
Qwen3-VL-235B (Bai et al., 2025)	Qwen3-VL-235B-A22B	4,028	23
Qwen3-Omni-30B (Xu et al., 2025b)	Qwen3-Omni-30B-A3B-Instruct	4,051	0
GPT-4o (OpenAI, 2024)	gpt-4o-2024-11-20	3,927	124
GPT-5.6-Sol (OpenAI, 2026a)	gpt-5.6-sol	3,987	64
GPT-6 Astra (OpenAI, 2026b)	gpt-6-astra	4,005	46
Gemini-3.1-Pro (Google, 2026)	gemini-3.1-pro-preview	4,029	22

For multimodal evidence grounding, a missing model response or an output-format failure receives a score of zero for every Gold-supported and input-available evidence channel. Infrastructure failures are tracked separately and excluded pending rerun rather than silently converted into model errors.

Deterministic bipartite alignment. For segments containing multiple strategy instances, predicted and Gold instances are aligned using deterministic maximum-weight bipartite matching. Let $G = \{g_1, \ldots, g_n\}$ denote the Gold strategy instances and $P = \{p_1, \ldots, p_m\}$ the predicted instances. For each Gold–prediction pair, the matching weight is defined over the four fields Actor, Target, PIF, and FIF:

$$w(g_i, p_j) = \mathbb{I}[A_i = A_j] + \mathbb{I}[T_i = T_j] + \mathbb{I}[\mathrm{PIF}_i = \mathrm{PIF}_j] + \mathbb{I}[\mathrm{FIF}_i = \mathrm{FIF}_j]. \tag{1}$$

Each edge therefore receives an integer weight between 0 and 4. Social Risk and IRS do not contribute to the alignment weight and are not used for tie breaking. Before matching, the legacy FIF labels Boundary Testing, Exception Negotiation, and Power Restructuring are mapped to the canonical category Boundary Negotiation and Restructuring; all other labels are compared using their canonical taxonomy forms. The bipartite graph is padded to a square matrix of size $\max(n, m)$ using dummy nodes, with all edges incident to dummy nodes assigned weight zero. We then select a one-to-one assignment that maximizes the total edge weight. Gold instances retain their annotation order, while predicted instances retain their order in the parsed strategy list. If multiple assignments obtain the same maximum total weight, we select the assignment with the lexicographically smallest vector of prediction indices in Gold-instance order, with real prediction indices preceding padded dummy indices, making the alignment deterministic and independent of solver-specific tie breaking. After alignment, Actor, Target, PIF, and FIF are evaluated on the aligned Gold–prediction pairs, and Overall is correct only when all four fields match simultaneously. Unmatched Gold instances are counted as incorrect. Under the main evaluation setting, the provided Gold strategy count normally gives $n = m$; if a response is missing or invalid, its prediction set is empty, and the corresponding Gold instances are aligned to dummy predictions and remain incorrect in the formal denominator.

F Fine-grained Results

We conduct a fine-grained analysis using Gemini-3.1-Pro, which achieves the highest Overall performance among the evaluated models. The analysis is performed on the full evaluation set, containing 4,051 target segments and 5,600 Gold strategy events. For segments containing multiple strategies, predictions are aligned with Gold events using the same deterministic maximum-weight bipartite matching procedure as in the main evaluation. For class-wise PIF and FIF analysis, Recall denotes the proportion of Gold instances of the corresponding class that are correctly predicted.

PIF and FIF. Performance varies substantially across both PIF and FIF categories, as shown in Table 25. At the PIF level, Acquire Information, Close Social Channel, and Send Information achieve relatively high recalls of 76.20%, 75.00%, and 73.74%, respectively, whereas Maintain Social Channel and Adjust Social Boundary are considerably more difficult, reaching 40.05% and 16.95%. Variation is even more pronounced at the FIF level. For example, within Open Social Channel, Contact Initiation reaches 81.88% recall whereas Joint Attention Establishment reaches only 7.87%. Similarly, within Maintain Social Channel, Interaction Repair reaches 61.70%, while Common Understanding Maintenance reaches 11.11%. The aggregate PIF and FIF accuracies are 66.16% and 50.43%, respectively, while their Macro-F1 scores decrease to 54.63% and 35.13%. These results indicate that recognizing broad interactional functions is generally easier than resolving the specific fine-grained function expressed by the observed behavior.

Table 25: Per-class PIF and FIF recall of Gemini-3.1-Pro on the full evaluation set. Support denotes the number of Gold instances. Class-wise recall is the proportion of Gold instances of the corresponding class that are correctly predicted.

PIF	Support	Recall (%)	FIF	Support	Recall (%)
Open Social Channel	493	60.24	Contact Initiation	309	81.88
			Joint Attention Establishment	127	7.87
			Interaction Intent Clarification	57	19.30
Set Social Boundary	121	69.42	Role and Relationship Positioning	28	67.86
			Interaction Rule Setting	50	70.00
			Topic Scope Setting	26	30.77
			Personal Boundary Declaration	17	52.94
Send Information	2380	73.74	Fact and Opinion Expression	1305	56.09
			Intention and Need Expression	347	44.38
			Emotion and Evaluation Expression	569	53.25
			Advice and Persuasion	159	50.31
Acquire Information	1311	76.20	Information Inquiry	881	66.29
			Understanding Calibration	248	54.84
			Mental State Probing	131	41.22
			Expression Permission Seeking	51	33.33
Maintain Social Channel	854	40.05	Responsive Listening	403	28.04
			Turn-Taking and Rhythm Coordination	63	23.81
			Common Understanding Maintenance	126	11.11
			Emotional Support	215	38.60
			Interaction Repair	47	61.70
Adjust Social Boundary	177	16.95	Relationship Closeness	128	18.75
			Boundary Negotiation and Restructuring	15	0.00
			Relationship Distancing	34	11.76
Close Social Channel	264	75.00	Closing Preview	32	12.50
			Summary Confirmation	25	36.00
			Relationship-Maintaining Farewell	142	61.97
			Terminating Closure	65	55.38
PIF aggregate accuracy		66.16	FIF aggregate accuracy		50.43
PIF Macro-F1		54.63	FIF Macro-F1		35.13

Social Risk. Social Risk exhibits severe class imbalance, as shown in Table 26. No Apparent Risk accounts for 5,497 of the 5,600 Gold events, whereas only 103 events contain an explicit risk mechanism. Consequently, Gemini-3.1-Pro obtains a high raw accuracy of 93.77%, but only 22.08% Macro-F1. The dominant No Apparent Risk class reaches an F1 score of 96.96%, while performance on the individual risk mechanisms is substantially lower. Information Manipulation and Interaction Punishment receive 0.00% F1, and even the strongest explicit-risk class, Relational Coercion, reaches only 26.09% F1. Among the 103 events containing an explicit risk mechanism, only 31 are assigned the correct risk class. The large discrepancy between accuracy and Macro-F1 therefore shows that high aggregate accuracy does not imply reliable recognition of fine-grained social risks.

Table 26: Per-class Social Risk performance of Gemini-3.1-Pro. Macro-F1 is computed over the fixed nine-class Gold label space.

Risk class	Support	Precision (%)	Recall (%)	F1 (%)
No Apparent Risk	5497	99.05	94.96	96.96
Boundary Intrusion	20	12.07	35.00	17.95
Autonomy Constraint	36	21.05	11.11	14.55
Emotional Manipulation	4	5.26	50.00	9.52
Information Manipulation	3	0.00	0.00	0.00
Relational Coercion	4	15.79	75.00	26.09
Power Suppression	5	6.67	40.00	11.43
Demeaning and Humiliation	22	13.68	59.09	22.22
Interaction Punishment	9	0.00	0.00	0.00
Accuracy				93.77
Macro-F1				22.08
Balanced accuracy				40.57

Interactional Realization Style. IRS exhibits a similarly uneven class distribution, as shown in Table 27. Direct Expression accounts for 4,748 of the 5,600 Gold events and achieves an F1 score of 88.43%, whereas the remaining realization styles are substantially more difficult. This imbalance leads to a substantial gap between the raw accuracy of 78.96% and the Macro-F1 of 28.31%. Across the 852 non-direct events, only 154 are correctly recognized, corresponding to 18.08% accuracy. The results indicate that the model recognizes dominant direct realizations considerably more reliably than less frequent indirect or interactionally nuanced realization styles.

Table 27: Per-class IRS performance of Gemini-3.1-Pro. Macro-F1 is computed over the fixed five-class Gold label space.

IRS class	Support	Precision (%)	Recall (%)	F1 (%)
Direct Expression	4748	87.01	89.89	88.43
Mitigated Expression	680	30.02	18.24	22.69
Implicit Expression	90	15.74	18.89	17.17
Tentative Expression	34	0.00	0.00	0.00
Avoidant Expression	48	8.78	27.08	13.27
Accuracy				78.96
Macro-F1				28.31
Balanced accuracy				30.82

Strategy-Count Analysis. Finally, we examine performance as a function of the number of Gold strategy instances contained in each segment. Since this analysis focuses on changes in event-level recognition as strategy cardinality increases, Actor and Target are reported using accuracy, while PIF and FIF are reported using both accuracy and Macro-F1. Overall denotes exact agreement over the four evaluated fields: Actor, Target, PIF, and FIF. Performance does not decrease monotonically with the number of strategies: segments containing one and two strategies obtain Overall scores of 40.80% and 42.91%, respectively, while three-strategy segments reach 48.22% and also show relatively strong PIF and FIF performance. Under the Gold strategy-count setting and event-level bipartite matching, strategy cardinality therefore does not appear to be the primary source of recognition difficulty. Results for segments containing four or more strategies should nevertheless be interpreted cautiously, since the five-strategy and seven-strategy groups contain only six and one segment, respectively, making their scores highly sensitive to individual examples.

Table 28: Performance grouped by the number of Gold strategy instances per segment. Actor and Target are evaluated using accuracy, while PIF and FIF are evaluated using accuracy and Macro-F1, denoted as F1 in the table. Overall denotes four-field exact match. Groups with very small support should be interpreted cautiously. Best results are shown in bold, and second-best results are underlined. All values are percentages.

Gold strategy count	Actor	Target	PIF		FIF		Overall
	Acc.	Acc.	Acc.	F1	Acc.	F1	
1	81.56	75.85	65.05	53.16	48.65	32.20	40.80
2	80.89	<u>77.42</u>	66.67	49.73	51.71	<u>34.26</u>	42.91
3	<u>84.63</u>	**82.20**	<u>68.77</u>	**57.75**	<u>54.21</u>	**36.99**	**48.22**
4	80.36	76.79	64.29	<u>54.21</u>	46.43	31.49	38.39
5	80.00	76.67	60.00	44.62	53.33	17.77	<u>43.33</u>
7	**100.00**	42.86	**85.71**	14.29	**85.71**	7.41	28.57
Overall	81.64	77.11	66.04	54.71	50.41	35.22	42.34

Sensitivity of the Six-field Overall Metric. Because Social Risk and IRS exhibit lower independent reproducibility than the core functional annotations, we examine how their inclusion affects the six-field Overall metric. We select the four strongest models according to the original six-field Overall score and recompute event-level exact match after excluding Risk, IRS, or both fields. All variants retain the original six-field maximum-weight alignment, with fields removed only after alignment rather than rematching predictions to Gold events.

As shown in Table 29, excluding Risk produces only a small increase of 0.94–1.69 percentage points, whereas excluding IRS yields a larger increase of 4.85–10.98 points. Excluding both attributes increases exact match by 6.73–12.34 points. Nevertheless, the relative ordering of the four models remains unchanged under all variants: Gemini-3.1-Pro, GPT-6 Astra, GPT-5.6-Sol, and Qwen3-Omni-30B. These results indicate that Risk and IRS, particularly IRS, make the six-field Overall criterion substantially stricter in absolute terms, while the ordering among the strongest models is stable to their inclusion.

Table 29: Comparison between the previous six-field Overall and the new four-field Overall. The four models with the highest previous six-field Overall scores are shown. Difference denotes the absolute increase from six-field to four-field Overall. Best results are shown in bold, and second-best results are underlined. All values are percentages.

Model	Previous Six-field Overall	New Four-field Overall	Difference
Qwen3-Omni-30B	21.20	28.02	+6.82
GPT-5.6-Sol	25.27	36.57	<u>+11.30</u>
GPT-6 Astra	<u>28.32</u>	<u>40.82</u>	**+12.50**
Gemini-3.1-Pro	**32.88**	**42.34**	+9.46

These results confirm that Risk and IRS, particularly IRS, make the six-field Overall criterion substantially stricter in absolute terms. However, removing either or both attributes leaves the ordering of the four strongest models unchanged. We therefore retain Overall as a strict composite measure while interpreting PIF and FIF separately as the primary indicators of core interactional-function recognition.

Sensitivity to grounding aggregation. As a sensitivity analysis, we examine whether the relationship between social micro-strategy understanding and evidence grounding is affected by differences in modality availability across models. Our primary analysis uses availability-aware aggregation, which averages over the evidence channels available to each model and therefore includes Prosody for audio-capable models but excludes it for models without audio input. As a complementary check, we construct a Common-4 Average using only Text, Face, Body, and Target Response, which are available to all 13 evaluated models. As shown in Table 30, the resulting cross-model association with four-field Overall remains highly stable: Pearson correlation changes from $r = 0.850$ under the primary availability-aware aggregation to $r = 0.858$ under the Common-4 Average, while Spearman correlation changes from $\rho = 0.923$ to $\rho = 0.918$. These small differences indicate that

the observed relationship between strategy understanding and evidence-grounding performance is not materially driven by the asymmetric availability of Prosody.

Table 30: Sensitivity of the cross-model correlation to the grounding aggregation scheme. Availability-aware aggregation is used in the primary analysis, while Common-4 Average provides a modality-consistent sensitivity check. Correlations are computed between four-field Overall and the corresponding grounding aggregate across the 13 evaluated models.

Grounding Aggregation	Models	Pearson r	Spearman ρ
Availability-aware Average	13	0.850	0.923
Common-4 Average	13	0.858	0.918

The near-identical correlations under the Common-4 sensitivity analysis provide additional evidence that the cross-model relationship reported in the primary analysis is robust to the grounding aggregation scheme. We therefore retain the availability-aware average for the main analysis and use the Common-4 Average only as a complementary robustness check.

G Additional Ablation Studies

G.1 Unknown-Count Strategy-Set Prediction

The main Task 1 evaluation provides the Gold number of strategy instances K for each target segment and requires the model to return exactly K structured predictions. This oracle-count setting isolates strategy recognition from cardinality estimation, but removes the need to infer how many distinct strategies are present. To quantify the effect of this assumption, we additionally evaluate Qwen3-Omni-30B without providing K, allowing the model to return an arbitrary number of strategy instances. We use the same 4,051 target segments and the same Gold annotations as in the oracle-count evaluation. Overall correctness requires Actor, Target, PIF, and FIF to be simultaneously correct for a strategy instance, and we report event-level precision, recall, and F1 under this criterion. Set Exact is a stricter segment-level metric that requires the complete predicted strategy set to exactly match the complete Gold strategy set. Among the 4,051 segments, 3,993 produce valid structured outputs, corresponding to a valid-output rate of 98.57%. On these valid outputs, exact-count accuracy is 42.95%, count MAE is 0.890, and the mean signed count error is $+0.706$. The model under-predicts the strategy count in 8.44% of segments and over-predicts it in 48.61%, indicating a strong tendency toward over-generation when the Gold count is removed.

Table 31: Effect of removing the oracle strategy count on Qwen3-Omni-30B. Both settings are evaluated on the same 3,993 segments with valid structured outputs using the same open-cardinality scoring procedure. Overall requires Actor, Target, PIF, and FIF to be simultaneously correct for a strategy instance. Set Exact requires the complete predicted strategy set of a segment to exactly match the Gold set. All values except Count MAE are percentages.

Setting	Count Acc.	Count MAE	Overall P.	Overall R.	Overall F1	Set Exact
Oracle K	100.00	0.000	28.06	28.06	28.06	21.01
Unknown K	42.95	0.890	21.15	32.02	25.47	10.64

Removing the Gold strategy count produces a clear precision–recall trade-off. Overall recall increases from 28.06% to 32.02%, while Overall precision decreases from 28.06% to 21.15%, showing that the model recovers additional correct strategy instances at the cost of substantial over-generation. Overall F1 decreases more moderately from 28.06% to 25.47%, whereas Set Exact drops from 21.01% to 10.64%. The much larger degradation in Set Exact indicates that Gold K primarily removes the difficulty of cardinality estimation and complete strategy-set recovery: individual correct strategies can still be recognized without the oracle count, but determining how many distinct strategies are present and recovering the complete set remains substantially more difficult.

For an end-to-end evaluation over all 4,051 target segments, we treat the 58 invalid structured outputs as containing no valid predictions and do not selectively retry them. Under this setting, Overall precision, recall, and F1 are 21.15%, 31.38%, and 25.27%, respectively, while Set Exact is

10.49%. This experiment still assumes pre-segmented target clips, and every target segment in the current evaluation set contains at least one annotated strategy. It therefore evaluates unknown-count strategy-set prediction rather than temporal strategy localization or strategy-presence detection for $K = 0$ segments.

G.2 Oracle Information Ablation

To better understand the sources of error in social micro-strategy understanding, we conduct controlled experiments with Qwen3-Omni-30B by additionally providing gold Actor and Target identities, gold multimodal evidence, or both. These oracle settings help distinguish errors caused by participant identification and evidence grounding from those arising in social micro-strategy reasoning itself. Results are shown in Table 32.

Table 32: Oracle information ablation on Qwen3-Omni-30B. Gold A/T denotes Gold Actor and Target identities, while Gold Evidence denotes Gold multimodal behavioral evidence. Actor and Target are evaluated using accuracy, while PIF and FIF are evaluated using accuracy and Macro-F1, denoted as F1 in the table. Overall denotes four-field exact match. Best results are shown in bold, and second-best results are underlined. All values are percentages.

Setting	Actor	Target	PIF		FIF		Overall
	Acc.	Acc.	Acc.	F1	Acc.	F1	
Original	76.77	69.64	58.96	43.81	38.87	22.89	28.10
+ Gold Evidence	85.18	78.19	<u>67.01</u>	<u>51.86</u>	<u>44.75</u>	**27.83**	33.91
+ Gold A/T	**100.00**	**100.00**	58.55	43.44	37.90	22.09	<u>37.40</u>
+ Gold A/T + Evidence	<u>99.87</u>	<u>99.91</u>	**67.20**	**52.27**	**45.14**	<u>27.46</u>	**44.50**

Providing gold Actor and Target identities substantially improves overall performance, while producing only limited gains in PIF and FIF recognition. In contrast, gold evidence leads to clearer improvements in PIF and FIF, suggesting that evidence grounding remains an important bottleneck for fine-grained social micro-strategy understanding. Combining both sources of oracle information yields the highest overall performance, while substantial errors remain in strategy recognition.

G.3 Effect of Visual Sampling Rate

Our main evaluation standardizes the visual sampling rate to 1 fps across models, including Qwen3-Omni-30B. Since the official configuration of Qwen3-Omni-30B uses a higher visual sampling rate of 2 fps, we additionally evaluate the model at 2 fps to examine whether constraining it to the benchmark-wide 1-fps setting affects its performance. To ensure a direct comparison, both settings are evaluated on the same common subset, with audio, subtitles, video duration, and all other evaluation settings kept unchanged. Results are reported in Table 33.

Table 33: Effect of visual sampling rate on Qwen3-Omni-30B over the common evaluation subset. The 1-fps setting is used in the main evaluation, while 2 fps corresponds to the model's official visual sampling configuration. The two settings differ only in visual sampling rate. F1 denotes Macro-F1, and Overall denotes four-field exact match. The better result in each metric is shown in bold. All values are percentages.

Visual Sampling	Actor	Target	PIF		FIF		Overall
	Acc.	Acc.	Acc.	F1	Acc.	F1	
1 fps (Main; common subset)	76.22	69.35	**59.91**	**44.41**	38.58	22.58	27.99
2 fps (Official; common subset)	**76.77**	**69.64**	58.96	43.81	**38.87**	**22.89**	**28.10**

The two sampling rates yield highly similar performance across all evaluation dimensions. Increasing the visual sampling rate from 1 fps to 2 fps produces small improvements in Actor, Target, FIF, and Overall, while both PIF accuracy and PIF Macro-F1 decrease slightly. In particular, Overall changes only marginally from 27.99 to 28.10. These results indicate that constraining Qwen3-Omni-30B to the benchmark-wide 1-fps sampling protocol does not materially affect its performance, supporting the use of a standardized 1-fps setting in the main comparison.

H Model Adaptation Details

We investigate task-specific adaptation using Qwen2.5-Omni-7B under three settings: SFT, GRPO, and SFT+GRPO. The first two are initialized from the pretrained model, whereas the third continues optimization from the selected SFT checkpoint. All adaptation experiments are conducted only for social micro-strategy understanding, whose structured output contains Actor, Target, PIF, and FIF. Hyperparameters and checkpoint selection are determined exclusively on the validation split, and the test split is evaluated only after the training configuration and checkpoint have been fixed.

For supervised fine-tuning, let $x = (V, A, T, P)$ denote the multimodal input, consisting of video V, audio A, subtitles T, and the task prompt P, and let $y = (y_1, \ldots, y_L)$ denote the Gold four-field structured output. We perform autoregressive fine-tuning by minimizing

$$\mathcal{L}_{\mathrm{SFT}}(\phi) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{train}}}\left[\sum_{\ell=1}^{L} \log p_{\theta,\phi}\left(y_\ell \mid y_{<\ell}, x\right)\right], \tag{2}$$

where θ denotes the pretrained model parameters and ϕ denotes the trainable LoRA parameters. LoRA is applied to the language-model component while the pretrained multimodal encoders are kept fixed. We tune the LoRA rank and learning rate on the validation split while fixing LoRA $\alpha = 32$, dropout $= 0.05$, BF16 precision, an effective batch size of 16, and a maximum of three training epochs. Intermediate checkpoints are evaluated using validation FIF accuracy. The selected configuration uses LoRA rank 8 and a learning rate of 2×10^{-4}, with the checkpoint at step 400 used for final evaluation and as the initialization for the subsequent reinforcement-learning stage.

For GRPO, we compare two initialization strategies. One starts directly from the pretrained model with a newly initialized LoRA adapter, while the other starts from the selected supervised checkpoint. Apart from initialization, both use the same training data, reward formulation, and optimization configuration, thereby isolating the effect of supervised initialization before reinforcement learning. For each training instance, a group of candidate structured outputs is sampled and optimized according to their relative rewards. To align reinforcement learning directly with the four-field evaluation target, the reward contains only structured-output validity and the four evaluated fields:

$$\begin{aligned} R = {} & 0.05R_{\mathrm{fmt}} + 0.10R_{\mathrm{Actor}} + 0.10R_{\mathrm{Target}} + 0.15R_{\mathrm{PIF}} \\ & + 0.30R_{\mathrm{FIF}} + 0.30R_{\mathrm{Overall}}. \end{aligned} \tag{3}$$

Here, R_{fmt} measures structured-output validity, R_{Actor} and R_{Target} evaluate participant identification, and R_{PIF} and R_{FIF} evaluate interactional-function prediction. The FIF reward is assigned only when both the corresponding PIF and FIF predictions are correct, preserving their hierarchical dependency. R_{Overall} rewards an exact match of all four evaluated fields, Actor, Target, PIF, and FIF. We retain the original field-level weights for Actor, Target, PIF, and FIF and reassign the reward mass previously associated with auxiliary attributes to the four-field Overall term, increasing its weight to 0.30. This keeps the dense field-level learning signal while more directly aligning the optimization objective with the primary evaluation criterion: partially correct predictions receive credit through the field-level terms, whereas jointly correct four-field predictions receive an additional reward through R_{Overall}. We analyze the resulting reward–metric alignment and optimization dynamics in Appendix H.1.

Both reinforcement-learning settings use LoRA rank 8, LoRA $\alpha = 32$, dropout $= 0.05$, a learning rate of 1×10^{-6}, BF16 precision, an effective batch size of 4, and one training epoch. We use four generations per prompt, temperature 0.9, top-$p = 1.0$, KL coefficient $\beta = 0.04$, clipping coefficient $\epsilon = 0.20$, and a maximum completion length of 512. Intermediate checkpoints are evaluated exclusively on the validation split using FIF accuracy, with the earlier checkpoint selected in the case of a tie. Final test evaluation is performed only after checkpoint selection.

H.1 GRPO Optimization Diagnostics

To better understand why GRPO provides only limited additional gains over supervised adaptation, we further analyze the optimization behavior of the SFT-initialized GRPO run. All diagnostics in this subsection are conducted exclusively on the validation split and do not affect checkpoint selection or test evaluation. We examine three complementary aspects: validation performance across checkpoints, alignment between the training reward and strict structured correctness, and the evolution of token-level policy entropy.

Validation dynamics. Validation performance improves rapidly during the early stage of GRPO but does not continue to increase with the training reward. FIF accuracy reaches its highest value of 41.63% at step 300, while Overall reaches its maximum of 26.35% at step 600. At the final checkpoint (step 2606), the corresponding values are 40.39% and 25.86%, respectively. In contrast, the mean training reward increases from 0.732 in the first half of training to 0.752 in the second half. Over the same two periods, mean validation FIF changes from 40.72% to 40.33%, and Overall changes from 25.99% to 25.76%. The fraction of sampled groups with zero reward variance also increases from 0.330 to 0.430, indicating that a larger proportion of sampled groups provide no within-group reward discrimination later in training. The average KL divergence remains small throughout training, increasing from 7.36×10^{-4} to 1.47×10^{-3} between the two halves. These observations indicate that continued reward optimization does not translate into corresponding improvements in validation accuracy.

Reward–metric alignment. We next evaluate the reward in Eq. 3 on the 371 validation instances retained by the GRPO length filter, using exactly the same reward implementation as during training. The reward remains strongly associated with strict four-field correctness: its Spearman correlation with Overall correctness is 0.778, and the mean reward is 1.000 for completely correct predictions. However, incorrect predictions still receive a mean reward of 0.502. Among the 271 predictions that fail Overall exact match, 149 (55.0%) receive a reward of at least 0.5 and 81 (29.9%) receive a reward of at least 0.6.

This behavior arises from the deliberately dense partial-credit design. Format validity is already saturated on the validation set (1.000), while the field-level reward components provide substantial credit for partially correct structured predictions. Consequently, a prediction can obtain a relatively high reward while still making an error in one of the Actor, Target, PIF, or FIF fields required for Overall correctness. The reward is therefore broadly aligned with the target metric, but its dense partial-credit structure provides weaker discrimination among predictions that are correct on most, but not all, of the four evaluated fields.

Policy entropy. Finally, we measure whether GRPO reduces policy exploration over the course of training. We fix a deterministic subset of 32 instances sampled from the 371 GRPO validation instances with seed 42 and evaluate the same subset at the selected SFT initialization and three GRPO checkpoints. For each generated sequence, we compute the mean entropy of the raw next-token policy distribution along the greedy decoding trajectory,

$$H = -\frac{1}{T} \sum_{t=1}^{T} \sum_{v} p_t(v) \log p_t(v), \tag{4}$$

where $p_t(v)$ is the model probability of token v at decoding step t. Using raw next-token probabilities makes this diagnostic independent of the sampling temperature used during GRPO training.

Table 34: Token-level policy entropy during SFT-initialized GRPO optimization. All checkpoints are evaluated on the same fixed 32-instance validation subset. Relative change is computed with respect to the selected SFT checkpoint.

Model State	Mean Entropy	Relative Change	Below SFT
SFT initialization	0.04744	–	–
GRPO step 300	0.04276	-9.87%	32/32
GRPO step 1400	0.03591	-24.32%	32/32
GRPO step 2606	0.03513	-25.96%	31/32

The reduction is consistent across validation instances. Comparing the final GRPO checkpoint with the SFT initialization, the paired mean difference is -0.01232, with a bootstrap 95% confidence interval of $[-0.01384, -0.01075]$. A paired Wilcoxon signed-rank test also confirms a systematic difference ($p < 10^{-8}$). Policy entropy decreases most strongly during the early and middle stages of GRPO and largely stabilizes after step 1400. We therefore characterize this behavior as a contraction of policy entropy rather than a complete entropy collapse.

Taken together, these diagnostics suggest that the limited additional gain from GRPO is not explained by a gross mismatch between the reward and strict correctness. Instead, two effects appear simultaneously: dense field-level partial credit reduces discrimination among predictions that differ only in a small number of fields, while policy entropy progressively contracts during optimization. This combination is consistent with increasing training reward without a corresponding improvement in validation FIF or Overall accuracy.

I HUMAN EVALUATION

We conduct human evaluation on 150 segments of the test set. Seven university students participate, with each segment independently evaluated by three participants, yielding 450 segment-level responses and 645 aligned strategy-event observations. Participants are assigned partially overlapping subsets rather than an identical subset. The reported Human result is computed by pooling all aligned individual event-level decisions before metric computation; it is neither a majority-vote consensus nor an unweighted average of annotator-level scores. Annotators are provided with the target video, subtitles, social-subject descriptions, and the predefined number of strategy instances, and their responses are evaluated using the same protocol as model predictions. Figure 8 shows the evaluation interface.

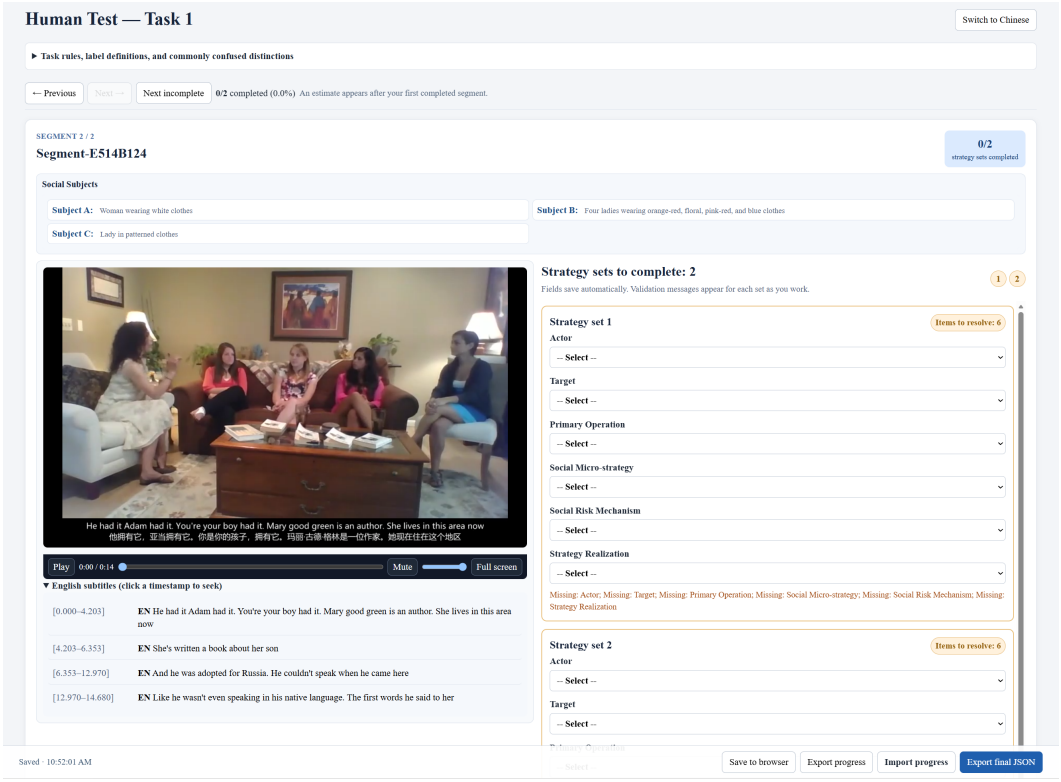


Figure 8: Human evaluation interface for social micro-strategy understanding. Annotators watch the target video with subtitles and social-subject descriptions, and complete the six fields for each predefined strategy instance.

As a sensitivity analysis, we additionally aggregate the three responses for each event using field-wise majority voting. This yields an Overall score of 44.19%, compared with 44.34% under pooled individual scoring, indicating that the human–model comparison is not sensitive to the aggregation scheme.

Table 35: Individual human performance on Social Micro-Strategy Understanding. Actor and Target are evaluated using accuracy, while PIF and FIF are evaluated using accuracy and Macro-F1, denoted as F1 in the table. Overall denotes four-field exact match.

Human	Actor	Target	PIF		FIF		Overall
	Acc.	Acc.	Acc.	F1	Acc.	F1	
Human 1	**88.35**	79.61	71.84	54.40	52.43	20.11	43.69
Human 2	87.06	**81.18**	**78.82**	**71.96**	**60.00**	**34.47**	**50.59**
Human 3	85.29	76.47	67.65	53.69	49.02	19.57	38.24
Human 4	78.95	76.84	73.68	46.90	55.79	19.32	42.11
Human 5	72.73	71.59	72.73	51.61	57.95	29.59	46.59
Human 6	81.40	75.58	70.93	43.38	56.98	21.68	47.67
Human 7	72.09	69.77	63.95	44.67	51.16	18.85	43.02
Pooled individual decisions	81.09	75.97	71.32	56.56	54.57	35.40	44.34

Table 36: Sensitivity of human performance to response aggregation. F1 denotes Macro-F1, and Overall denotes four-field exact match. All values are percentages.

Aggregation	PIF		FIF		Overall
	Acc.	F1	Acc.	F1	
Pooled individual decisions	71.32	**56.56**	**54.57**	**35.40**	**44.34**
Three-person majority vote	**72.56**	54.34	53.49	26.29	44.19