MISID: A Multimodal Multi-turn Dataset for Complex Intent Recognition in Strategic Deception Games

Shufang Lin*
123090335@link.cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen
Shenzhen, GuangDong, China

Muyang Chen*
123090028@link.cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen
Shenzhen, GuangDong, China

Xiabing Zhou
zhouxiabing@cnu.edu.cn
Capital Normal University
Beijing, China

Rongrong Zhang
zhangrr@cnu.edu.cn
Capital Normal University
Beijing, China

Dayou Zhang†
zhangdayou@cnu.edu.cn
Capital Normal University
Beijing, China

Fangxin Wang
wangfangxin@cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen
Shenzhen, GuangDong, China

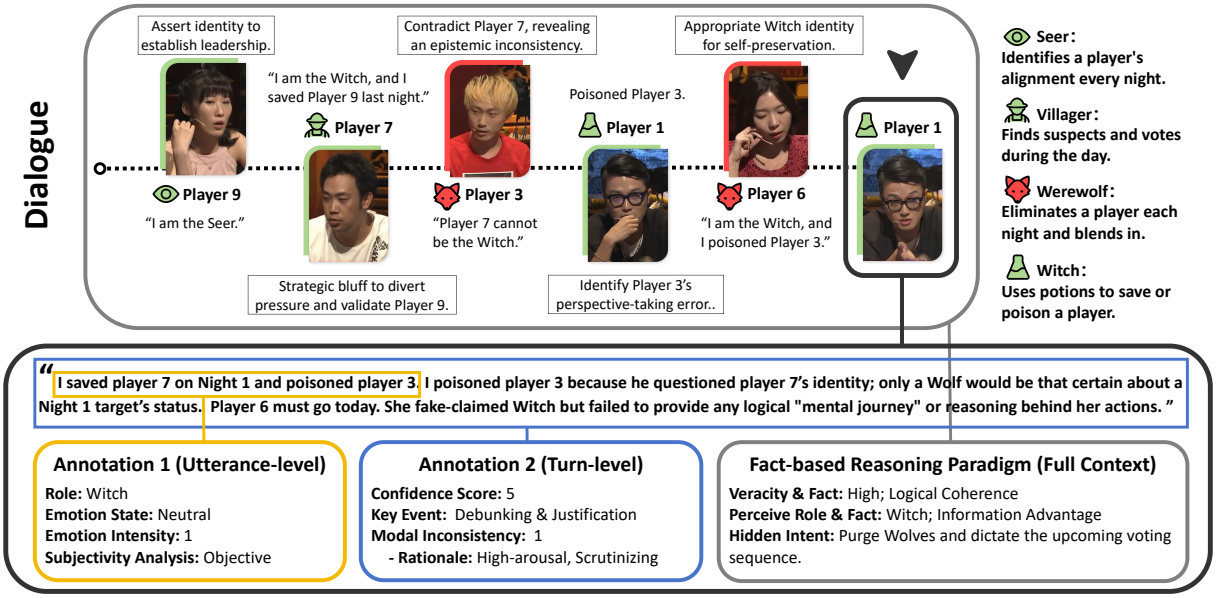

Figure 1: An overview of the MISID benchmark. (Top) A multi-participant strategic dialogue timeline exhibiting hidden tactics. (Bottom) Our multi-dimensional annotation scheme and fact-based reasoning paradigm for deducing hidden intents.

Abstract

Understanding human intent in complex multi-turn interactions remains a fundamental challenge in human-computer interaction and behavioral analysis. While existing intent recognition datasets focus mainly on single utterances or simple dialogues, real-world scenarios often involve sophisticated strategic interactions where participants must maintain complex deceptive narratives over extended periods. To address this gap, we introduce MISID, a comprehensive multimodal, multi-turn, and multi-participant benchmark for intent recognition. Sourced from high-stakes social strategy games, MISID features a fine-grained, two-tier multi-dimensional annotation scheme tailored for long-context discourse analysis and evidence-based causal tracking. Our systematic evaluation of state-of-the-art Multimodal Large Language Models (MLLMs) on MISID reveals critical deficiencies in complex scenarios, including text-prior visual hallucination, impaired cross-modal synergy, and limited capacity in chaining causal cues. Consequently, we propose FRACTAM as a baseline framework. Using a "Decouple-Anchor-Reason" paradigm, FRACTAM reduces text bias by extracting pure unimodal factual representations, employs two-stage retrieval for long-range factual anchoring, and constructs explicit cross-modal evidence chains. Extensive experiments demonstrate that FRACTAM enhances mainstream models' performance in complex strategic tasks, improving hidden intent detection and inference while maintaining robust perceptual accuracy. Our dataset is available at https://naislab.cn/datasets/MISID.

CCS Concepts

• **Computing methodologies** → **Language resources**.

*Both authors contributed equally to this research.
†Corresponding author.

Keywords

Intent Recognition, Deception Detection, Multimodal Dataset, Strategic Games

1 Introduction

"Real life consists of bluffing, of little tactics of deception, of asking yourself what is the other man going to think I mean to do."

— John von Neumann

Human social interaction is fundamentally a game of incomplete information [19, 51]. From high-stakes political negotiations to everyday workplace dynamics, people strategically conceal their true intentions behind euphemisms [7, 39], leverage multimodal signals to deceive or persuade, and continuously adapt their strategies across prolonged exchanges [8]. While current Artificial Intelligence (AI) excels at processing transparent, explicit instructions, it falters when confronted with this defining feature of human sociality: the pervasive gap between what people say and what they truly mean [34, 44]. As AI advances toward genuine social intelligence [63], the benchmark for evaluating model capabilities can no longer rely on surface semantic comprehension [5]. Instead, the critical challenge lies in penetrating strategic expressions to perceive hidden intentions, tracking how these intentions evolve over long-term interactions, and capturing the subtle cross-modal leaks, such as a fleeting micro-expression belying verbal agreement, that expose genuine psychological states.

To endow AI with such sophisticated social intelligence, comprehensive benchmarks are imperative. However, existing intent recognition datasets exhibit severe limitations when addressing dynamic and covert interactions [32, 45, 46]. Most benchmarks are predominantly "transparent" and static, focusing on instantaneous, single-sentence interactions where speech perfectly aligns with thought [60, 61]. By stripping away the strategic concealment and long-term causal chains prevalent in real-world settings, these datasets force models to over-rely on superficial textual signals [13]. Even the few datasets that address intent concealing often remain at the level of binary judgments or shallow "stimulus-response" patterns [25]. They lack fine-grained analysis of intent dynamics, hidden motivations, and the complex historical facts that drive dynamic interactions, fundamentally hindering models from developing fact-grounded reasoning capabilities [63].

To bridge this gap, we introduce MISID—a multi-turn, multimodal, and multi-participant intent recognition benchmark. Constructed from multimodal sources of high-pressure social strategy games involving deception, reasoning, and voting-based elimination, MISID naturally recreates realistic strategic scenarios with extreme information density. Featuring 3,962 high-quality utterance segments with temporally aligned audiovisual modalities, the dataset transcends traditional single-dimensional labeling by introducing a two-tier multi-dimensional annotation scheme. It scales from utterance-level micro-states (sentiment, intensity) to macro-level, long-range multimodal discourse analysis. By explicitly annotating cross-modal inconsistencies, cross-turn logical contradictions, and precisely localized historical key facts, MISID creates an evidence-based causal reasoning paradigm. It shifts the learning objective from superficial guessing to tracking complex derivation chains based on hard evidence.

To gauge the capability of current models in such strategic scenarios, we systematically evaluated mainstream Large Language Models (LLMs) alongside their multimodal counterparts (MLLMs) on MISID. We observed that the concealment and temporal dynamics of human intentions uniquely captured by MISID pose fundamental challenges to existing multimodal architectures. Specifically, their failures are characterized by three primary dilemmas: (1) *Text-prior Visual Hallucination*, where models force visual representations to align with dominant textual logic rather than relying on objective visual input; (2) *Causal Threading Limitations*, where models struggle to penetrate noisy, multi-turn contexts to connect scattered historical events into a coherent causal chain; and (3) *Impaired Modal Synergy*, where distributing causal chains across modalities paradoxically increases fact-related error rates compared to unimodal settings.

Addressing these multi-turn multimodal reasoning dilemmas, we propose the FRACTAM (**F**act-grounded **R**easoning **A**nd **C**ausal **T**hreading **A**cross **M**odalities) framework. Based on a structured "decoupling-anchoring-reasoning" paradigm, FRACTAM directly tackles the observed model failures. First, it extracts unimodal symbolic representations from audiovisual signals to mitigate textual prior dominance. Second, it employs a dual-stage, long-range fact anchoring mechanism to isolate causal variables from historical noise accurately. Finally, by constructing cross-modal causal chains, it effectively repairs impaired modality synergy, enabling the model to perform robust, evidence-based reasoning in complex gaming scenarios.

Our primary contributions are as follows:

- We propose MISID, the first multi-turn, multimodal, and multi-dimensionally labeled benchmark dataset designed within a complex, high-pressure strategic environment.
- We comprehensively benchmark state-of-the-art unimodal and multimodal models, exposing their critical bottlenecks (e.g., text-prior hallucination and impaired modal synergy) in multi-turn hidden intent recognition.
- We introduce the FRACTAM framework, a novel structured reasoning paradigm that effectively overcomes current limitations in long-range cross-modal causal threading, establishing a strong baseline for future research in artificial social intelligence.

2 Related Work

Dialogue Understanding and Intent Recognition Datasets. Recent advancements in affective computing and dialogue understanding depend heavily on multimodal benchmarks [54]. Mainstream datasets (e.g., SLURP [4], FSC [42], MIntRec [61] and MELD [41]) primarily target explicit emotion alignment and multi-dimensional intent classification. To capture hidden psychological states, various datasets have been developed to study deception and masking behaviors (e.g., Real-Life Trial Data [37], Bag-of-Lies [18], and MUStARD [9]), investigating phenomena from physiological leakage in constrained environments to intentional masking in the wild. Furthermore, recent works have begun extending into the LLM-driven multimodal domain (e.g., MECPE [53], IntentQA [27], Genesis [28]). However, these existing datasets have notable limitations: they primarily assume explicit expressions, are confined to highly

Table 1: Comparison of MISID with existing multimodal dialogue and intent recognition benchmarks. Depth: Literal (Explicit) or Underlying (Implicit) Intentions. CSE: Complex Strategic Environment. FCA: Fact-based Causal Annotation. Length: Dialogue Turn Ranges.

Dataset	Depth	Text	Audio	Video	CSE	FCA	Length
MCIC [57]	Explicit	✓	✗	✗	✗	✗	10-30
MSAIRS [48]	Explicit	✓	✗	✗	✗	✗	1-10
SLURP [4]	Explicit	✓	✓	✗	✗	✗	1-10
FSC [42]	Explicit	✓	✓	✗	✗	✗	1-10
MINDS-14 [14]	Explicit	✓	✓	✗	✗	✗	1-10
MIntRec [61]	Explicit	✓	✓	✓	✗	✗	1-10
MIntRec 2.0 [60]	Explicit	✓	✓	✓	✗	✗	10-20
EMOTyDA [43]	Explicit	✓	✓	✓	✗	✗	1-10
EmoInt-MD [49]	Explicit	✓	✓	✓	✗	✗	1-10
MC-EIU [31]	Explicit	✓	✓	✓	✗	✗	1-10
BID [33]	Explicit	✓	✓	✓	✗	✗	1-10
MECPE [53]	Explicit	✓	✓	✓	✗	✓	10-40
Genesis [28]	Explicit	✓	✓	✓	✗	✓	100-500
Open Domain [38]	Implicit	✓	✗	✗	✗	✗	1-10
CSC [21]	Implicit	✗	✓	✗	✗	✗	1-10
Bag-of-Lies [18]	Implicit	✓	✓	✓	✗	✗	1-10
Box of Lies [50]	Implicit	✓	✓	✓	✗	✗	10-20
Real Life Trials [37]	Implicit	✓	✓	✓	✗	✗	1-10
MELD [41]	Implicit	✓	✓	✓	✗	✗	10-100
MUStARD [9]	Implicit	✓	✓	✓	✗	✗	1-10
MultiMET [59]	Implicit	✓	✗	✗	✗	✗	1-10
MDID [24]	Implicit	✓	✗	✗	✗	✗	1-10
CraigslistBargain [20]	Implicit	✓	✗	✗	✓	✗	10-30
IntentQA [27]	Implicit	✓	✓	✓	✗	✓	1-10
Diplomacy [35]	Implicit	✓	✗	✗	✓	✗	100-600
MISID (OURS)	Implicit	✓	✓	✓	✓	✓	154-555

restricted contexts with single-dimensional features, and often restrict annotations to static, binary judgments. Consequently, they fail to track the dynamic evolution of intents grounded in key facts during real-world strategic concealment. To fill this gap, MISID provides hierarchical annotations, multi-turn multimodal synchronized data, and a fact-based reasoning paradigm.

Intent and Deception Detection Methods. To capture temporal cues and multi-party interactions, traditional methods typically employ sequence models, basic attention mechanisms and Graph Neural Networks (e.g., LSTMs [22], TCNs [3], DialogueGCN [15], DAG-ERC [47]). Advanced deep networks, such as CTNet [29], further integrate global contexts through deep attention frameworks. More recently, Large Language Models (LLMs) have been heavily leveraged to provide powerful zero-shot reasoning capabilities in multimodal social scenarios [10, 23, 52, 55]. However, these methodologies present critical limitations in high-pressure, long-range social games. Traditional sequence and graph-based models struggle to maintain long-range strategic contexts. They are prone to capturing superficial temporal correlations rather than robust causal dependencies, and often suffer from feature over-smoothing that obscures sparse triggers. Meanwhile, despite their strong reasoning potential, LLMs are hampered by long-term memory decay, modality bias, and a lack of rigorous causal frameworks to constrain hallucinations. To overcome these bottlenecks, our proposed FRACTAM framework integrates context-aware key-fact extraction, multimodal perception, and fact-based rigorous reasoning to effectively address these challenges.

Table 2: Dataset Statistics. Physical statistics (left) and multimodal annotation distributions (right) of the MISID dataset. VAD: Voice Activity Detection. Inconsistency: Cross-modal Incongruence During Interactions.

Dimension		Annotation Distribution	
Total Participants	15	Subjectivity: Subj	3,259
Avg. Participants	12	Subjectivity: Obj	703
Role Instances	120	Emotion: Calm	2802
Total Utterances	3,962	Emotion: Others	1160
Total Duration	9.15 hours	Deception: Truth	2191
Max Turns	555	Deception: Lie	1771
Avg. Turns	374.7	Inconsistency: High	1542
Mean VAD Ratio	92.8%	Inconsistency: Low	2420

3 Dataset

Dataset Overview We introduce MISID, a multi-turn multimodal dataset based on complex strategic games. The dataset contains 3,962 high-quality utterance segments, covering 120 role-specific utterance instances from 15 participants, with detailed statistics and annotations summarized in Table 2. Each segment includes video and audio modalities obtained from public online videos and precisely synchronized to ensure temporal alignment.

Data Processing Audio tracks were uniformly standardized to 16kHz mono PCM format. Speaker diarization was conducted with Pyannote 3.1 [6, 40], augmented by multi-dimensional filtering, heuristic merging and manual correction. Simultaneously, the visual modality utilized face detection and DBSCAN-based identity clustering techniques [12] to extract individual sequences, ensuring strict temporal alignment with the audio components.

Data Annotation MISID adopts a two-tier multi-dimensional annotation scheme: the first layer records foundational background (identity, basic emotion, intensity, and subjectivity); the second layer targets discourse analysis (anchoring key events, confidence, and modality inconsistency), establishing cross-modal and multi-turn factual reasoning chains. The annotation workflow follows a two-stage pattern: initial labels are first generated by Large Language Models, followed by rigorous manual verification by four human annotators, integrating iterative peer reviews and cross-modal consistency audits.

Fact-based Reasoning Paradigm We constructed an evidence-based causal reasoning paradigm aimed at guiding models to learn factual derivation chains. By leveraging the two-tier annotations and ground truth, this paradigm precisely locates key contextual cues to guide the model in reconstructing logical chains and inferring participants' deceptive behaviors and hidden intentions, serving as a standard benchmark for model training and evaluation.

4 Methodology

FRACTAM is designed to address intent concealment and multimodal signal conflicts in complex social games. Its core logic is to first achieve the objective decoupling of physical facts, then anchor

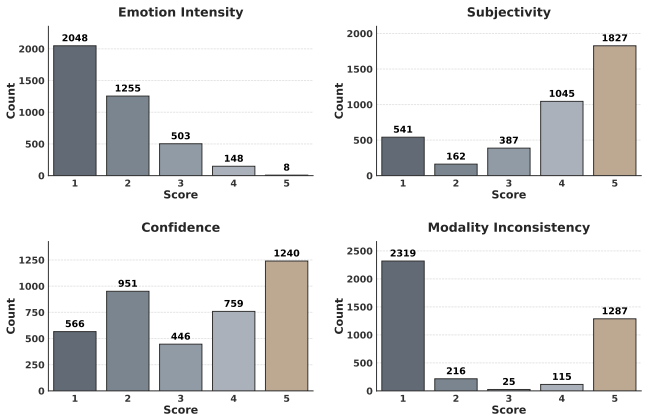

Figure 2: Distribution of multi-dimensional annotations in the MISID dataset. The x-axis denotes the annotation score, and the y-axis represents the occurrence frequency.

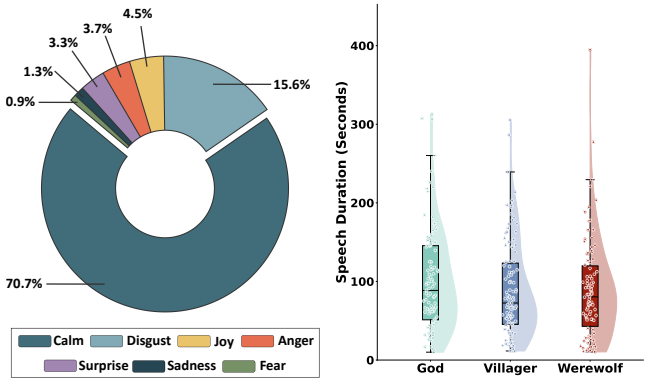

Figure 3: Emotion category distribution (left) and speech duration distribution (right) in MISID dataset.

historical evidence via long-range retrieval, and ultimately conduct evidence-based reasoning through the construction of cross-modal causal chains.

4.1 Modal Decoupling and Text Reconstruction

To eliminate modality feature contamination induced by textual logic alignment, FRACTAM utilizes Multimodal Large Language Models (MLLMs) to implement strict unimodal fact decoupling. For the input signal $\mathcal{X}_i = \{x_i^t, x_i^v, x_i^a\}$ at the i-th interaction round, we disable early cross-modal attention and independently input the non-textual modalities x_i^m into the MLLM Θ. By introducing a restrictive prompt $\mathbf{p}_m$, the model is constrained to output only objective descriptions of physical states. This decoding process can be abstracted as the maximization of the sequence joint probability under given conditions:

$$f_i^m = \arg\max_{\mathbf{y}} P_\Theta\big(\mathbf{y} \mid \Phi_m(x_i^m), \mathbf{p}_m\big), \quad m \in \{v, a\} \tag{1}$$

where the Φ_m represents the modality-specific projection mapping. Through Equation (1), both visual and audio signals are decoded into factual texts f_i^v and f_i^a, respectively. Ultimately, the multimodal signals are decoupled and uniformly constructed into a unified plain-text fact set $\mathcal{F}_i = \{x_i^t, f_i^v, f_i^a\}$.

4.2 Hybrid Long-range Fact Anchoring

Addressing the challenge of sparse causal variables in long-term interactions, FRACTAM introduces a two-stage long-range fact anchoring mechanism. Given the current fact set $\mathcal{F}_t$ and the historical memory bank $\mathcal{H} = \{\mathcal{F}_i\}_{i=1}^{t-1}$, we first conduct a dual-path recall in the lexical (lex) and semantic (sem) feature spaces, calculating the initial relevance via weighted Reciprocal Rank Fusion [11]:

$$\Omega(\mathcal{F}_i, \mathcal{F}_t) = \sum_{\rho \in \{\text{lex},\text{sem}\}} \frac{\omega_\rho}{\eta + \pi_\rho(\mathcal{F}_i, \mathcal{F}_t)} \tag{2}$$

where $\pi_\rho(\cdot,\cdot)$ is the descending ranking function in a specific retrieval space, and η is a smoothing term. Based on this score, we extract the top M facts with the highest Ω scores from $\mathcal{H}$ to form a high-confidence candidate subset $\mathcal{H}_{cand}$:

$$\mathcal{H}_{cand} = \text{Top-M}\Big\{\mathcal{F}_i \in \mathcal{H} \;\Big|\; \Omega(\mathcal{F}_i, \mathcal{F}_t)\Big\} \tag{3}$$

To further capture deep contextual dependencies, we introduce a cross-encoder (Qwen3-Reranker [62]) Φ_{ce} to re-rank the candidate set. For each historical fact $\mathcal{F}_j$ in the candidate set, the model calculates its deep semantic relevance score s_{ce} with the current fact $\mathcal{F}_t$ through a joint attention mechanism:

$$s_{ce}(\mathcal{F}_t, \mathcal{F}_j) = \sigma\Big(\mathbf{W}_{ce}^\top \Phi_{ce}\big([\mathcal{F}_t; \mathcal{F}_j]\big)\Big) \tag{4}$$

where $[\cdot;\cdot]$ denotes sequence concatenation, $\mathbf{W}_{ce} \in \mathbb{R}^{d_{ce}\times 1}$ is the linear projection weight, and σ is the activation function. Based on this score, we select the top K most relevant historical facts from $\mathcal{H}_{cand}$ to construct the final retrieval context bank $\mathcal{C}_{ret}$:

$$\mathcal{C}_{ret} = \text{Top-K}\Big\{\mathcal{F}_j \in \mathcal{H}_{cand} \;\Big|\; s_{ce}(\mathcal{F}_t, \mathcal{F}_j)\Big\} \tag{5}$$

4.3 Causal Threading and Intent Inference

To resolve the "modality synergy degradation" caused by black-box fusion, we introduce explicit modality fact evidence chains. Based on the current modality-decoupled fact set $\mathcal{F}_t$ and the retrieval context $\mathcal{C}_{ret}$, and under the constraint of the cross-modal alignment prompt $\mathbf{p}_{align}$, the reasoning model Ψ is forced to construct an explicit causal evidence chain $\mathcal{T}_c$:

$$\mathcal{T}_c = \arg\max_{\mathbf{z}} P_\Psi\big(\mathbf{z} \mid \mathcal{F}_t, \mathcal{C}_{ret}, \mathbf{p}_{align}\big) \tag{6}$$

Subsequently, under the logical hard constraint of the system inference prompt $\mathbf{p}_{sys}$, the model Ψ must strictly follow the evidence chain $\mathcal{T}_c$ and key facts to generate the final output:

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{y}} P_\Psi\big(\mathbf{y} \mid \mathcal{T}_c, \mathcal{F}_t, \mathcal{C}_{ret}, \mathbf{p}_{sys}\big) \tag{7}$$

Through this proposed paradigm, the model Ψ effectively coordinates various modalities with an interpretable evidence chain, yielding the evidence-based reasoning result $\hat{\mathbf{Y}}$ regarding fact determination and intent analysis threads.

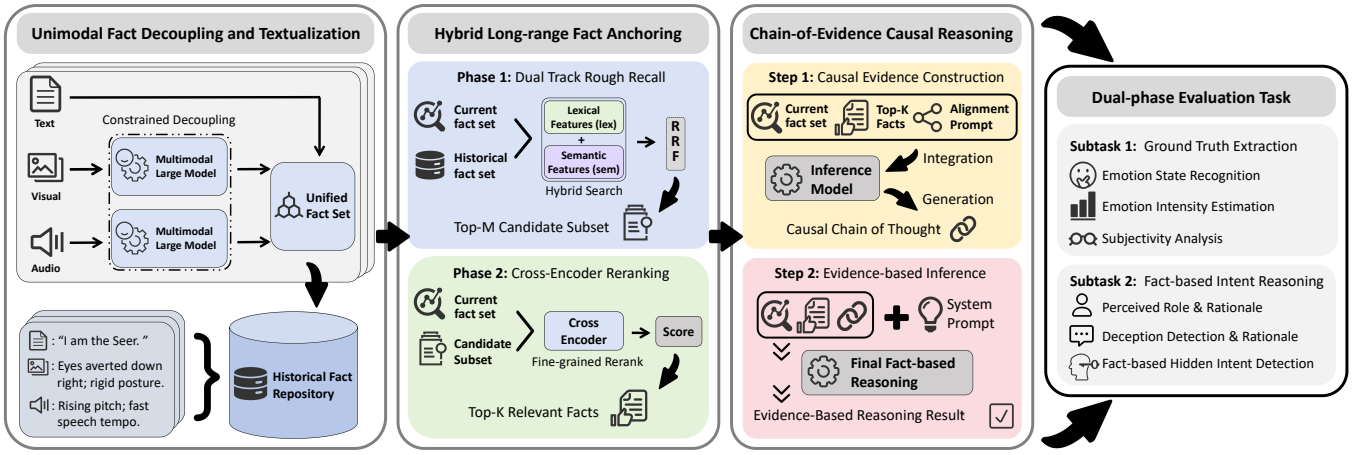

Figure 4: Overall architecture of the FRACTAM framework. The pipeline standardizes multimodal inputs into objective text, retrieves historical evidence via dual-stage hybrid search, and constructs explicit logical chains to deduce underlying intents for dual-phase evaluation.

5 Evaluation Metrics

For ground truth extraction, we adopt standard metrics along with a distance-decay scoring mechanism based on numerical gaps. For hidden intent extraction involving complex causal reasoning, n-gram metrics based on lexical overlap and cosine similarity calculations are no longer sufficient. Therefore, we employ the **LLM-as-a-Judge** evaluation paradigm, utilizing a Large Language Model evaluator $\mathcal{E}_{LLM}$ for quantification.

5.1 State Recognition Metrics

State recognition metrics are used to evaluate the model's precision in determining the key game states of participants and serve as factual grounding for downstream reasoning metrics. Among these, **Role Accuracy (RA)** measures the model's judgment regarding the participants' specific identities and their respective camps. We utilize a tiered scoring method:

$$RA = \frac{1}{N}\sum_{i=1}^{N} \mathcal{F}_{role}(y_i^{role}, \hat{y}_i^{role}), \tag{8}$$

where $\mathcal{F}_{role}$ is the tiered scoring function, which assigns a full score for a completely correct identity, a partial score if only the camp is correct, and 0 for an incorrect judgment.

Deception Binary Accuracy (DBA) evaluates the model's classification performance regarding "whether a lie was told":

$$DBA = \frac{100}{N}\sum_{i=1}^{N} I(y_i^{lie} = \hat{y}_i^{lie}), \tag{9}$$

where $I(\cdot)$ is the indicator function, which equals 1 when the prediction matches the ground truth and 0 otherwise.

5.2 Evidence-based Reasoning Metrics

Evidence-based reasoning metrics are used to evaluate the model's fact recall and logical argumentation capabilities. We quantify the model evaluation results across three dimensions: Identity Reasoning (IRS), Lie Details (LDS), and Hidden Intent (HIS). For any reasoning task $task \in \{IRS, LDS, HIS\}$, the evaluator $\mathcal{E}_{LLM}$ assesses the fact-grounded responsiveness (Φ_{FG}) and logical consistency (Φ_{LC}) of the predicted reasoning $\hat{R}$ using the following formula:

$$Score_{task} = \frac{1}{N}\sum_{i=1}^{N}\Big(\alpha \cdot \Phi_{FG}(F_i^*, \hat{R}_i) + \beta \cdot \Phi_{LC}(R_i^*, \hat{R}_i)\Big), \tag{10}$$

where F_i^* represents the key facts, R_i^* represents the standard logic, and α, β are balancing weights. Furthermore, to suppress fact-detached logical hallucinations, a pre-state hard truncation mechanism is introduced:

$$HIS_i = \min\Big(\tau,\ HIS_i\Big), \quad \text{if } RA_i = 0 \vee DBA_i = 0, \tag{11}$$

where τ is the penalty threshold. This mechanism ensures that when the ground truth determination is incorrect, the score for its corresponding reasoning explanation is constrained.

6 Experiments

6.1 Experimental Setup

We benchmark a diverse suite of mainstream models. These are categorized into VideoLLMs and Text-only LLMs. All models were evaluated in a zero-shot setting. For text-only LLMs, visual dynamics were represented through textual behavioral descriptions.

6.2 Model Limitations Analysis

Our empirical analysis identifies three critical bottlenecks hindering current foundation models in the MISID environment:

Text-prior Visual Hallucination. In deception scenarios, people often exhibit a distinct discrepancy between their verbal language and non-verbal cues. We observe that VideoLLMs heavily over-rely on the audio transcript. When textual claims conflict with visual evidence, they frequently ignores authentic visual cues and

Table 3: Performance of LLMs on various tasks. The scores are reported in percentage (%). The best and the second best results are denoted by pink and yellow. LLM implies the exclusive use of the text modality. FRACTAM refers to LLMs assisted by the FRACTAM framework.

	Model	Emotion State Accuracy	Emotion Intensity Score	Subjectivity Accuracy	Identity Judgment Accuracy	Identity Reasoning Score	Lie Detection Accuracy	Lie Reasoning Score	Hidden Intent Inference Score
VideoLLM	GPT-4o [23]	76.43	73.75	77.14	46.15	43.77	50.16	45.77	39.77
	Grok-4-Fast [55]	86.21	79.89	90.80	53.37	41.88	53.82	39.46	32.69
	Qwen3-VL-235B-A22B [2]	79.05	76.19	87.14	28.85	31.69	42.31	31.46	21.69
	Qwen3-VL-Flash	69.23	46.70	87.91	46.15	41.54	46.15	41.58	28.27
	Qwen3-VL-Max	74.05	75.57	91.22	30.77	34.85	46.15	35.54	23.42
	Qwen3-VL-Plus	72.83	61.78	90.58	15.38	27.69	34.62	34.77	23.27
	Qwen3.5-Plus [52]	83.78	80.28	94.37	13.46	26.65	30.77	29.65	22.12
	Gemini-2.5-flash [10]	82.55	78.55	91.64	21.15	31.08	50.16	36.42	26.85
	Gemini-3-flash [16]	87.09	85.10	94.04	53.85	52.73	53.88	57.58	49.52
	Average	79.02	73.09	89.43	34.35	36.88	45.34	39.14	29.73
LLM	Claude-Sonnet-4.5 [1]	89.01	81.87	82.42	50.48	45.12	57.69	52.38	43.35
	DeepSeek-R1 [17]	83.45	78.10	83.79	65.38	56.54	42.31	43.85	36.35
	DeepSeek-V3 [30]	66.91	71.40	80.94	51.92	49.04	50.33	49.52	44.23
	GLM-4.7 [58]	83.79	81.03	92.07	42.31	45.38	34.62	38.27	31.96
	GPT-4o	80.28	75.70	81.34	51.92	44.81	49.84	48.27	41.23
	GPT-5.1 [36]	82.73	74.32	92.73	44.23	48.54	53.85	53.15	36.46
	Grok-4-Fast	86.59	81.10	90.24	57.69	47.62	61.54	50.65	40.85
	MiniMax-M2.1 [26]	80.08	82.27	86.06	42.31	40.69	23.08	27.31	21.69
	Qwen3-Max [56]	77.18	77.52	87.58	26.92	30.15	34.62	29.04	22.85
	Qwen3.5-Plus	87.32	81.08	93.24	17.31	27.50	34.62	32.65	23.27
	Average	81.73	78.44	87.04	45.05	43.54	44.25	42.51	34.22
FRACTAM	Claude-Sonnet-4.5	$90.75_{\uparrow 1.74}$	$81.46_{\downarrow 0.41}$	$82.88_{\uparrow 0.46}$	$61.96_{\uparrow 11.48}$	$56.53_{\uparrow 11.41}$	$61.14_{\uparrow 3.45}$	$57.27_{\uparrow 4.89}$	$53.27_{\uparrow 9.92}$
	DeepSeek-R1	$83.73_{\uparrow 0.28}$	$80.18_{\uparrow 2.08}$	$85.66_{\uparrow 1.87}$	$72.22_{\uparrow 6.84}$	$63.87_{\uparrow 7.33}$	$50.25_{\uparrow 7.94}$	$50.86_{\uparrow 7.01}$	$45.59_{\uparrow 9.24}$
	DeepSeek-V3	$69.53_{\uparrow 2.62}$	$71.2_{\downarrow 0.20}$	$81.92_{\uparrow 0.98}$	$61.69_{\uparrow 9.77}$	$57.83_{\uparrow 8.79}$	$56.6_{\uparrow 6.27}$	$55.05_{\uparrow 5.53}$	$48.02_{\uparrow 3.79}$
	GLM-4.7	$83.39_{\downarrow 0.40}$	$81.3_{\uparrow 0.27}$	$93.34_{\uparrow 1.27}$	$55.01_{\uparrow 12.70}$	$55.0_{\uparrow 9.62}$	$42.56_{\uparrow 7.94}$	$45.97_{\uparrow 7.70}$	$43.04_{\uparrow 11.08}$
	GPT-4o	$79.87_{\downarrow 0.41}$	$75.9_{\uparrow 0.20}$	$83.11_{\uparrow 1.77}$	$54.85_{\uparrow 2.93}$	$50.16_{\uparrow 5.35}$	$53.85_{\uparrow 4.01}$	$51.61_{\uparrow 3.34}$	$46.09_{\uparrow 4.86}$
	GPT-5.1	$84.14_{\uparrow 1.41}$	$74.59_{\uparrow 0.27}$	$94.29_{\uparrow 1.56}$	$56.93_{\uparrow 12.70}$	$58.41_{\uparrow 9.87}$	$60.36_{\uparrow 6.51}$	$59.38_{\uparrow 6.23}$	$52.83_{\uparrow 16.37}$
	Grok-4-Fast	$88.92_{\uparrow 2.33}$	$80.62_{\downarrow 0.48}$	$92.56_{\uparrow 2.32}$	$55.56_{\downarrow 2.13}$	$49.75_{\uparrow 2.13}$	$59.8_{\downarrow 1.74}$	$45.1_{\downarrow 5.55}$	$40.01_{\downarrow 0.84}$
	MiniMax-M2.1	$82.02_{\uparrow 1.94}$	$82.96_{\uparrow 0.69}$	$86.1_{\uparrow 0.04}$	$43.29_{\uparrow 0.98}$	$43.13_{\uparrow 2.44}$	$31.59_{\uparrow 8.51}$	$32.58_{\uparrow 5.27}$	$29.72_{\uparrow 8.03}$
	Qwen3-Max	$80.03_{\uparrow 2.85}$	$78.2_{\uparrow 0.68}$	$87.4_{\downarrow 0.18}$	$33.76_{\uparrow 6.84}$	$38.65_{\uparrow 8.50}$	$41.09_{\uparrow 6.47}$	$35.85_{\uparrow 6.81}$	$30.44_{\uparrow 7.59}$
	Qwen3.5-Plus	$87.16_{\downarrow 0.16}$	$83.55_{\uparrow 2.47}$	$94.85_{\uparrow 1.61}$	$37.82_{\uparrow 20.51}$	$44.7_{\uparrow 17.20}$	$41.7_{\uparrow 7.08}$	$39.23_{\uparrow 6.58}$	$38.84_{\uparrow 15.57}$
	Average	$82.95_{\uparrow 1.22}$	$79.0_{\uparrow 0.56}$	$88.21_{\uparrow 1.17}$	$53.31_{\uparrow 8.26}$	$51.8_{\uparrow 8.26}$	$49.89_{\uparrow 5.64}$	$47.29_{\uparrow 4.78}$	$42.79_{\uparrow 8.57}$

generating hallucinated visual descriptions that spuriously align with the deceptive text. Consequently, the average Hidden Intent Inference Score for VideoLLMs languishes at a mere 29.73%.

Limitations in Threading Causal Clues. Both VideoLLMs and text-only LLMs demonstrated severe limitations in temporal causal reasoning. Their reasoning scores consistently trail their binary judgment accuracies. For instance, while text-only LLMs average 44.25% in Lie Detection Accuracy, their Lie Reasoning Score drops to 42.51%, indicating that even when models guess the correct label, their underlying causal justification is often flawed.

Impaired Modal Synergy. The integration of raw visual modalities degrades, rather than enhances, higher-order strategic reasoning. As shown in Table 3, text-only LLMs consistently outperform VideoLLMs across all strategic metrics. Specifically, the average Identity Judgment Accuracy for text-only LLMs is 45.05%, compared to a significantly lower 34.35% for VideoLLMs, which indicates compromised modal synergy. Rather than complementing, the unrefined visual modality acts as noise and distracts the model from logical deduction.

7 Conclusion

In this work, we address the limitations of existing intent recognition benchmarks by introducing MISID, a pioneering multi-turn, multimodal, and multi-participant dataset constructed from high-pressure strategic games. By offering precise audiovisual alignment and a novel multi-dimensional annotation scheme that captures cross-modal inconsistencies, MISID shifts the evaluation paradigm toward complex, fact-grounded causal reasoning. Our comprehensive evaluation of state-of-the-art models on MISID reveals critical bottlenecks in current architectures, notably text-prior visual hallucination and impaired modal synergy during complex social interactions. To establish a robust baseline, we propose the FRACTAM framework, which mitigates these multi-turn reasoning dilemmas through strict modality decoupling and explicit evidence chain construction. Ultimately, by capturing the pervasive gap between surface expressions and hidden psychological states, we envision MISID serving as a vital catalyst for the multimedia community, driving future research toward more interpretable, robust, and genuinely socially intelligent multimodal models.

References

[1] Anthropic. 2025. Claude Sonnet 4.5 System Card. https://www.anthropic.com/claude-sonnet-4-5-system-card.

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

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[4] Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 7252–7262.

[5] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 8718–8735.

[6] Hervé Bredin. 2023. pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proceedings of the Interspeech Conference*. ISCA, 1983–1987.

[7] Penelope Brown. 1987. *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press.

[8] David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. *Communication Theory* 6, 3 (1996), 203–242.

[9] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 4619–4629.

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[11] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the international ACM SIGIR Conference on Research and Development in Information Retrieval*. 758–759.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 226–231.

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[14] Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 7468–7475.

[15] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. 154–164.

[16] Google DeepMind. 2025. Gemini 3 Flash: Frontier Intelligence Built for Speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/.

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

[18] Viresh Gupta, Mohit Agarwal, Manik Arora, Tanmoy Chakraborty, Richa Singh, and Mayank Vatsa. 2019. Bag-of-lies: A multimodal dataset for deception detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 83–90.

[19] John C Harsanyi. 1967. Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model. *Management Science* 14, 3 (1967), 159–182.

[20] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2333–2343.

[21] Julia Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura A Michaelis, Bryan L Pellom, Elizabeth Shriberg, and Andreas Stolcke. 2005. Distinguishing deceptive from non-deceptive speech. In *Proceedings of the Annual Conference of the International Speech Communication Association*. 1833–1836.

[22] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

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

[24] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. 4622–4632.

[25] David La Barbera, Gian Carlo Milanese, Georgios Peikos, Gabriella Pasi, and Marco Viviani. 2025. Beyond binary classification: ranking for information access in misinformation contexts. In *Proceeding of the National Conference on Artificial Intelligence (CEUR Workshop Proceedings, Vol. 4121)*. 1–7.

[26] Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, et al. 2025. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313* (2025).

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

[28] Yulong Li, Yuxuan Zhang, Rui Chen, Feilong Tang, Zhixiang Lu, Ming Hu, Jianghao Wu, Haochen Xue, Mian Zhou, Chong Li, et al. 2025. Genesis: A Large-Scale Benchmark for Multimodal Large Language Model in Emotional Causality Analysis. In *Proceedings of the ACM International Conference on Multimedia*. 12651–12658.

[29] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 985–1000.

[30] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[31] Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. 2024. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751* (2024).

[32] Shuhua Liu, Lanting Li, Ming Fang, Chih-Cheng Hung, and Shihao Yang. 2022. Research on Implicit Intent Recognition Method Based on Prompt Learning. *Available at SSRN 4164522* (2022).

[33] Adyasha Maharana, Quan Hung Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. 2022. Multimodal intent discovery from livestream videos. In *Findings of the Association for Computational Linguistics*. 476–489.

[34] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences* 28, 6 (2024), 517–540.

[35] Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*. 1650–1659.

[36] OpenAI. 2025. GPT-5.1 Instant and GPT-5.1 Thinking System Card Addendum. https://openai.com/index/gpt-5-system-card-addendum-gpt-5-1/.

[37] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the ACM on International Conference on Multimodal Interaction*. 59–66.

[38] Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1120–1125.

[39] Steven Pinker, Martin A Nowak, and James J Lee. 2008. The logic of indirect speech. *Proceedings of the National Academy of Sciences* 105, 3 (2008), 833–838.

[40] Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proceedings of the Interspeech Conference*. ISCA, 3222–3226.

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

[42] Yao Qian, Ximo Bianv, Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao, and Michael Zeng. 2021. Speech-language pre-training for end-to-end spoken language understanding. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7458–7462.

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

[44] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 3762–3780.

[45] Vasanth Sarathy, Alexander Tsuetaki, Antonio Roque, and Matthias Scheutz. 2020. Reasoning requirements for indirect speech act interpretation. In *Proceedings of the International Conference on Computational Linguistics*. 4937–4948.

[46] Jocelyn Shen, Amina Luvsanchultem, Jessica Kim, Kynnedy Smith, Valdemar Danry, Kantwon Rogers, Sharifa Alghowinem, Hae Won Park, Maarten Sap, and Cynthia Breazeal. 2026. The Hidden Puppet Master: A Theoretical and Real-World Account of Emotional Manipulation in LLMs. *arXiv preprint arXiv:2603.20907* (2026).

[47] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*. 1551–1560.

[48] Yuanchen Shi, Fang Kong, and Longyin Zhang. 2025. Impact of Stickers on Multimodal Sentiment and Intent in Social Media: A New Task, Dataset and Baseline. In *Proceedings of the ACM International Conference on Multimedia*. 5637–5646.

[49] Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Emoint-trans: A multimodal transformer for identifying emotions and intents in social conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 290–300.

[50] Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. 1768–1777.

[51] Michael Spence. 1978. Job market signaling. In *Uncertainty in economics*. Elsevier, 281–306.

[52] Qwen Team. 2026. Qwen 3.5: Scaling Native Multimodal Agents with Efficient Architectures. https://qwen.ai/blog?id=qwen3.5.

[53] Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2022. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing* 14, 3 (2022), 1832–1844.

[54] Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. *arXiv preprint arXiv:2505.20511* (2025).

[55] xAI. 2025. Grok 4.1 Model Card. https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf.

[56] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[57] Shaozu Yuan, Xin Shen, Yuming Zhao, Hang Liu, Zhiling Yan, Ruixue Liu, and Meng Chen. 2022. MCIC: multimodal conversational intent classification for E-commerce customer service. In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 749–761.

[58] Z.ai. 2025. GLM-4.7 Model Card. https://build.nvidia.com/z-ai/glm4_7/modelcard.

[59] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A multimodal dataset for metaphor understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*. 3214–3225.

[60] Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. 2024. MIntRec2.0: A Large-scale Benchmark Dataset for Multimodal Intent Recognition and Out-of-scope Detection in Conversations. In *Proceedings of the International Conference on Learning Representations*.

[61] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the ACM International Conference on Multimedia*. 1688–1697.

[62] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176* (2025).

[63] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net. https://openreview.net/forum?id=mM7VurbA4r