Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Social Human Robot Embodied Conversation (SHREC) Dataset: Benchmarking Foundational Models' Social Reasoning

Dong Won Lee[♢] Yubin Kim[♢] Sooyeon Jeong[♡] Denison Guvenoz[♡]
Parker Malachowsky[♢] Louis-Philippe Morency[♠]
Cynthia Breazeal[♢] Hae Won Park[♢]
[♡]Purdue University [♠]Carnegie Mellon University
[♢]Massachusetts Institute of Technology

Dataset: https://huggingface.co/collections/MIT-personal-robots/shrec
Code: https://github.com/mitmedialab/SHREC

***Abstract*—Our work focuses on the social reasoning capabilities of foundational models for real-world human–robot interactions. We introduce the Social Human Robot Embodied Conversation (SHREC) Dataset, a benchmark of ∼400 real-world human-robot interaction videos and over 10,000 annotations, capturing robot social errors, competencies, underlying rationales, and corrections. Unlike prior datasets focused on human–human interactions, the SHREC Dataset uniquely highlights the social challenges faced by real-world social robots such as emotion understanding, intention tracking, and conversational mechanics. Moreover, current foundational models struggle to recognize these deficits, which manifest as subtle, socially situated failures. To evaluate AI models' capacity for social reasoning, we define eight benchmark tasks targeting critical areas such as (1) detection of social errors and competencies, (2) identification of underlying social attributes, (3) comprehension of interaction flow, and (4) providing rationale and alternative correct actions. Experiments with state-of-the-art foundational models, alongside human evaluations, reveal substantial performance gaps—underscoring the difficulty and providing directions in developing socially intelligent AI.**

I. Introduction

In this work, we aim to advance the social reasoning capabilities of physically embodied AI agents, particularly tabletop social robots, engaged in real-world, socially interactive conversations. To this end, we focus on understanding and modeling both social competencies (i.e. desirable behaviors) and social errors (i.e. norm violations or failures in interaction) that arise during natural human-robot interactions. While prior work in social intelligence has primarily centered on human–human interaction datasets [36, 60, 55], these settings do not capture the unique challenges that arise when robots interact with humans. Unlike humans, embodied AI agents may have varying and lacking socio-cognitive skills such as emotion understanding, belief tracking, or conversational coordination [31, 12, 5, 1], leading to failure cases that are subtle, socially situated, and poorly represented in existing resources.

To this end, we introduce the **Social Human Robot Embodied Conversation (SHREC) Dataset**, a dataset of 400+ videos capturing natural conversational interactions between a human and a physical tabletop social robot, making it, to the best of our knowledge, one of the largest real-world human–social robot interaction benchmark dataset available to date (see App. Tab. IV). We focus on a tabletop social robot, which is a widely used class of physically embodied agents, as a principled starting point for studying real-world social reasoning, and introduce a benchmark dataset that addresses the scarcity of real-world social robot interaction data available for advancing research in this area. The dataset is accompanied by 10K+ human annotations of the robot's social errors (undesirable behaviors) competencies (desirable

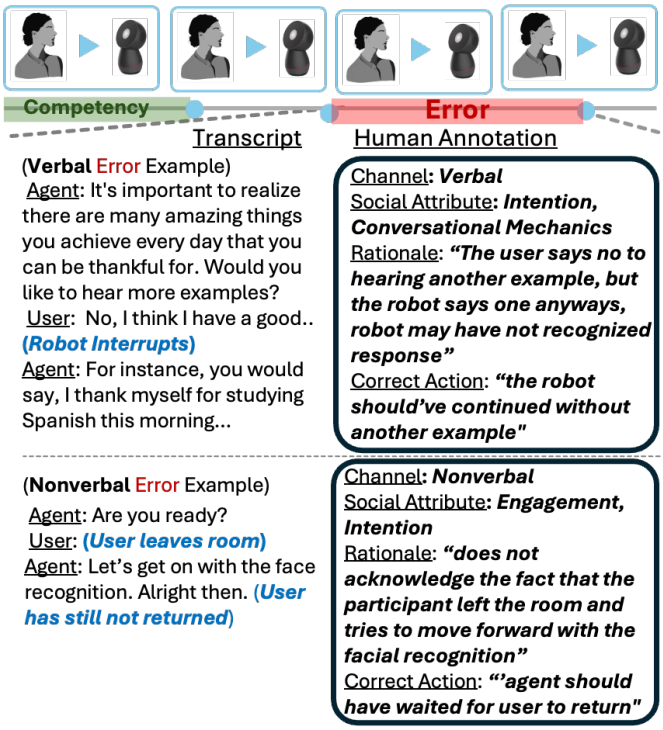

Fig. 1. SHREC Dataset dataset offers real-world **Social Human Robot Embodied Conversation** videos and annotations of errors and competencies, the channel and type of social attribute, along with rationale and possible corrective actions. (Top) Error sourced from verbal (audio) channel, (Bottom) Error sourced from non-verbal (visual) channel.

behaviors), rationale and corrective actions. These annotations are grounded from prior HRI taxonomies [52, 16] of seven key social attributes that underpin social interactions, emotion, engagement, conversational mechanics, knowledge state, intention, social relationships, and norms.

The SHREC Dataset offers a novel resource to assess foundational models' capabilties in identifying a social robot's social errors and competencies in real-world human-robot interactions, filling a gap not addressed by current social intelligence benchmarks. To systematically evaluate social reasoning of state-of-the-art AI models, we propose eight benchmark tasks, spanning four core dimensions: (1) errors and competency detection in the robot's behavior, (2) social attribute identification related to the errors and competencies, (3) interaction progression reasoning, and (4) rationale and correction reasoning (outlined in Sec. IV). Beyond assessing models' social reasoning, generalization, and robustness in real-world, multimodal, and socially grounded settings [67, 10, 31, 44, 43, 34], these tasks also serve as structured probes for fundamental representation learning challenges. They require multimodal alignment to integrate visual, auditory, and linguistic signals into representations for social understanding; reasoning to infer how specific behaviors influence downstream interaction trajectories; and reward-learning to align embodied agents to human social preferences. Our benchmark provides not only a principled testbed for advancing socially intelligent foundation models but also a new resource investigate these core learning problems.

We benchmarked 18 state-of-the-art LLMs and VLMs, including ChatGPT [23] and Gemini [50] variants, on these tasks. While some models excel in specific subtasks, none perform well across the board. Notably, the gap between model and human performance remains substantial, underscoring the challenge and novelty of our benchmark. These findings highlight Social Human Robot Embodied Conversation (SHREC) Dataset as a valuable testbed for diagnosing and improving the social reasoning abilities of embodied AI agents, and for guiding the development of reward models and evaluators aligned with social intelligence [38, 67, 33, 7, 66].

II. Related Work

A. Datasets for Social Interaction Analysis

Several datasets have been developed to analyze human social interactions, many of which focus on conversational data or multimodal behaviors. For example, the MELD (Friends TV Series) Dataset [42] and the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset [63] provide annotated multimodal data for studying emotions and sentiment in dialogue. More closely related to our work, the SocialIQA [45] dataset introduced 38,000 multiple-choice questions derived from the ATOMIC [44] knowledge graph, which is a large-scale graph of commonsense knowledge. These questions are aimed at testing models' understanding of social norms, intentions, and emotional responses in social scenarios. Furthermore, the Social-IQ Dataset [60] was designed to evaluate social intelligence in AI with human-to-human interaction videos, including multimodal question-answer pairs that assess the ability to understand and respond to social situations effectively. A recent benchmark, SOCIAL GENOME [36], evaluates multimodal models' ability to generate grounded social reasoning traces from videos, incorporating fine-grained cues and external knowledge. Our benchmark includes around 400 videos, consisting of 3600 minutes of real-world human–robot interaction footage, this scale is on par with or larger than many multimodal datasets in social interaction research (e.g., Social Genome [36]: 280 minutes, Social-IQ [61]: 1200 minutes, MEmoR [48]: 2800 minutes, CMU-MOSEI [62]: 3900 minutes). We also refer the reviewer to Appendix D, where we provide a comparative table of related works for clarity. Importantly, what distinguishes our dataset is that it is collected from real-world deployments of physically embodied social robots, offering naturalistic, longitudinal interactions, an extremely scarce setting, with, to the best of our knowledge, no publicly available data at this scale with equally comprehensive annotations.

B. Frameworks for Annotating Social Errors in HRI

Previous work in Human-Robot Interaction (HRI) has defined social errors as instances where atypical robot behavior results in violations of social norms [18]. These studies primarily focused on nonverbal cues, annotating both the robot's behavior and the human social signals, such as speech, head movements, gestures, facial expressions, and body posture, during error episodes. A taxonomy to systematically categorize social errors and competencies in HRI was developed, which forms the foundation for our own approach [52]. Similarly, the ASA questionnaire was proposed to assess social intelligence in artificial agents, leveraging human feedback collected across interaction trials [16]. While these efforts have proposed valuable conceptual frameworks, our work is the first to contribute a large-scale, aligned dataset of real-world HRI episodes annotated for social errors and competencies.

C. Social Interaction Analysis with Language Models

Several previous studies have focused on the social reasoning capabilities of language-model social agents. Theory of Mind (ToM) has been tested through a variety of tasks, often times through measuring an LLM's ability to understand others' mental states using a series of reasoning-specific tasks [53], or identifying social errors and understanding the perspectives of participants through faux pas tasks [47]. Frameworks such as COKE utilize contextual meanings of input entities to more accurately map knowledge graphs to be used in LLMs [56], and the SOCIALIQA benchmark similarly provides a collection of commonsense questions with appropriate and inappropriate responses to be used in LLMs [45]. In studying emotional understanding, psychometric assessments have been developed to measure emotional understanding of an LLM based on a given scenario [54]. Benchmarks such as Socratis measured emotional intelligence utilizing a repository of emotional reactions and appropriate scenarios [12].

III. SHREC: Dataset of Human Robot Social Embodied Conversation

Our dataset consists of 10,353 annotations from 403 interaction videos spanning over 3,500 minutes. Under a newly accepted IRB protocol, we annotated and anonymized data on three prior human-robot interaction studies [49, 28, 26] to be shared for dissemination. To enable public access while preserving participant privacy, we follow institutional IRB procedures and release the dataset under gated access. All personally identifiable information (PII) in transcripts is filtered. For video anonymization, we leverage FRESCO [58], a zero-shot video-to-video diffusion framework, to perform stylized face transfer. FRESCO's spatial-temporal consistency allows us to reliably replace participant faces while maintaining coherence across frames. This ensures that the social signals (e.g., gaze, affect) critical for interaction analysis are preserved while protecting individual identities through high-fidelity anonymization. In selecting FRESCO, we compared several anonymization techniques on the same interaction clips, generating matched variants from the raw videos and evaluating them with the identical pipeline (see Appendix C). Most vision-language models (VLMs) operate at 1 Hz, which is the default temporal resolution used in our benchmark. Hence, we release video frames sampled at 1 Hz to align with common VLM processing rates and ensure compatibility across models. [1]

We describe the three real-world studies that form the foundation of the SHREC dataset, all of which used the **Jibo social robot** [40] as the embodied interaction platform. Jibo is a tabletop social robot originally designed as a daily companion for the home. Its embodiment supports speech-based interaction, expressive gaze, animated idle behaviors such as blinking and looking around, and attention behaviors that orient the robot toward sound sources or people's faces. These capabilities enable Jibo to proactively initiate lightweight social interactions, offer greetings or activities, and maintain a sense of social presence during interaction. At the same time, Jibo's embodiment imposes important constraints: it is non-mobile, has no arms or object-manipulation capabilities, and expresses social behavior primarily through speech, gaze orientation, screen-based display, and body/head motion. Firstly, **Empathic++** [49]: a ChatGPT-powered social robotic agent acts as an empathic companion, facilitating the exchange of emotionally meaningful stories using narrative therapy techniques. The goal is to enhance users' feelings of connection and belonging through emotionally attuned interaction. **Wellness-Dorm** [26]: A socially assistive robot was deployed as a positive psychology coach for college students living in on-campus dormitories. The robot was manually scripted with seven intervention types grounded in established principles of positive psychology, such as gratitude, strengths-based reflection, and goal setting. **Wellness-Home** [28]: Positive psychology robots were deployed in participants' homes. All three datasets are ready for dissemination (We kindly recommend the reader to view the supplementary materials for examples of the dataset and demographic information).

[1]To support community interest in fine-grained temporal analysis, we will release an extended addendum with 15Hz videos (processing is underway and sample videos are included in the supplementary). We maintain a private hold-out set of ∼40 videos to prevent data contamination and enable future benchmarking and evaluations.

A. Human Annotations

As shown in Fig. 1 and App. C, annotators watch videos of conversational human-robot interactions and are asked to detect segments which manifest a ***social error or competency***. ***Social Competence*** is defined as the behaviors where the agent successfully conduct social interactions by being aware of and identifying social-emotional cues, processing such cues, and expressing a user-expected response to these cues [22]. ***Social Error*** is defined as the behaviors where the robot deviates from the desired behaviors expected by a user and degrades the user's perception of a robot's social competence [52]. Then, they are asked to identify which ***social attribute*** the segment is related to (described in Sec. III-A1) and offer a ***rationale*** of why they believe so. If the segment was an error, they are asked to suggest an alternative ***correct action***.

1) Social Attributes: Given a segment of social error or competencies, we are interested in the *explanatory factors* related to a social error or competency. To do so, we consider seven specific categories of social attributes that are related to social errors and competencies. The definitions for each of the attributes are as follows: **Emotions:** The ability to identify and interpret emotional expressions in oneself and others, allowing for empathetic responses and social awareness, e.g., recognizing that someone crying might mean they're sad [19]. **Engagement:** The skill to observe and assess levels of participation and involvement in social interactions, including cues that indicate interest or disinterest, e.g., continuing to tell a story when a listener is engaged [11]. **Conversational Mechanics:** Understanding the structure and flow of conversations, including turn-taking, interruptions, and cues for when to speak or listen, e.g., waiting for another person to finish speaking before taking a turn [17]. **Knowledge State:** The ability to assess what others know or believe, as well as being aware of one's own knowledge in social situations, e.g., make reference to a user's dog recalling that the user has a dog [3]. **User Intention:** The capacity to infer the goals or purposes behind the actions and words of others, facilitating better responses in social interactions, e.g., when the user says "I'll be right back", indicates that the user will vacate and then return [14]. **Social Context and Relationships:** The ability to identify and understand the dynamics of social relationships and the context in which they occur, influencing behavior and expectations, e.g., knowing how to act in front of a close friend vs. colleague at work [3]. **Social Norms and Routines:** The skill to identify accepted behaviors and attitudes within a social group, as well as recognizing negative or harmful interactions that violate these norms. e.g., understanding that waving hands at the beginning of the interaction is a sign of a greeting [51]

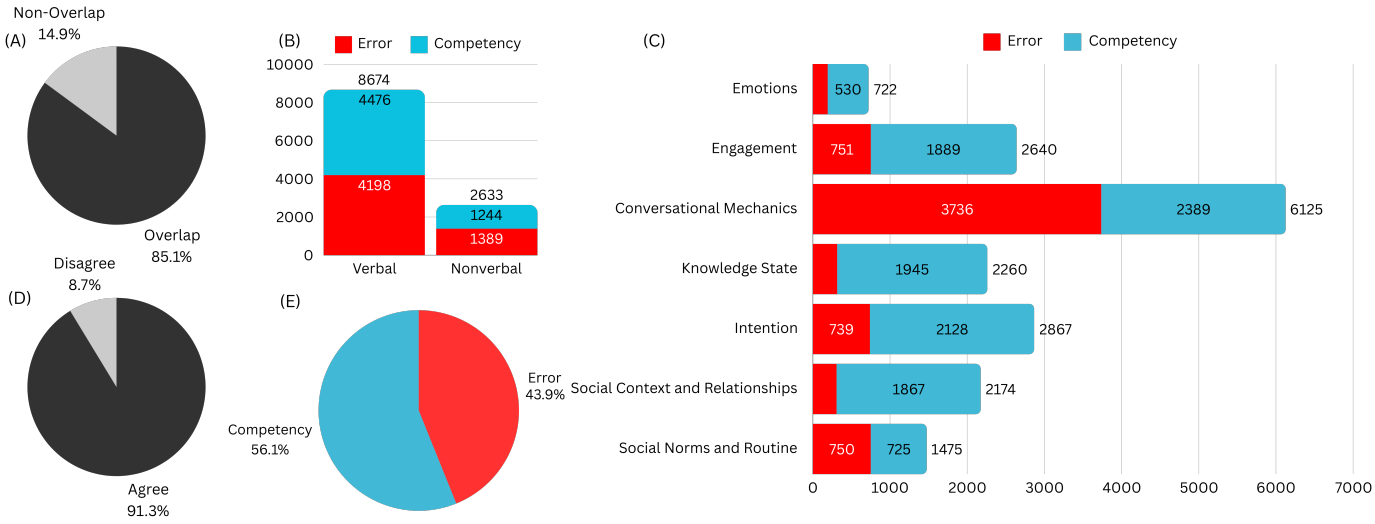

Fig. 2. SHREC Dataset contains **high overlapping** annotations with a **high level of agreement**. The dataset includes error and competency labels, and annotations for the source of evidence either from nonverbal cues, verbal cues, and explanatory factors in the form of seven key social attributes.

B. Dataset Statistics

We refer the reader to Figure 2 for the overall statistics of our Social Human Robot Embodied Conversation (SHREC) Dataset. As shown in Figure 2-(A), 85.1% of the dataset consists of overlapping annotations, where multiple annotators independently labeled the same segment of an interaction. Amongst the overlapping samples as shown in Figure 2-(D), we find a 91.3% overall agreement and a Cohen's κ = 0.7638 and Krippendorff's α = 0.76392. In Figure 2-(B), we find that more annotations come from verbal channel, rather than the non-verbal channel (For further analysis on which modality the human annotators relied on when annotating, we refer the reader to App. C). As shown in Figure 2-(E), we find 56% corresponds to competency labels and 43% corresponds to error labels. In Figure 2-(C), we display the social attributes the annotators have selected, and we find that the most number of annotations belong to the conversational mechanics category, followed by intention and engagement.

Annotator Consistency: To ensure consistent annotations, we recruited and trained three undergraduate research assistants, and these annotators produced the full set of more than 10k annotation. They were provided a set of definitions and annotation guidelines as shown in Appendix C. The annotators met multiple times to discuss edge scenarios and ambiguous segments. After annotating independently, annotators cross-validated each other's work. Some scenarios that were too subjective to reach full consensus yielded some annotations in the Disagree category of 8.7%.

IV. Tasks & Experiments

We developed eight tasks to measure the social reasoning capability of foundational models. Each task measures a different social reasoning capability of the model, spanning from identifying social errors and competencies, to reasoning about the errors and offering correct actions. Below, we describe four main research questions which motivated the design of the social reasoning tasks and the corresponding tasks which serve to address these questions.

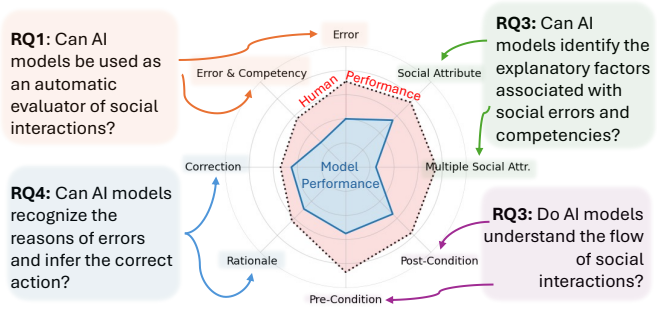

Fig. 3. Gemini-2.5-Pro performance across the eight social reasoning tasks defined in Sec. IV, grouped by the four research questions motivating our benchmark. The radar plot compares model and human performance across error/competence detection, social attribute reasoning, pre-/post-condition reasoning, rationale generation, and correction reasoning. All model results are provided in Fig. 5.

- **RQ1: Can AI models be used as an automatic evaluator of social interactions?**
  (1) Errors and Competence Detection, (2) Error Detection (Sec. IV-A, V-A)
- **RQ2: Can AI models identify the explanatory factors associated with social errors and competencies?**
  (3) Social Attribute Identification, (4) Multiple Social Attribute Presence (Sec. IV-B, V-B)
- **RQ3: Do AI models understand the sequential contingencies or the "flow" of social interactions?**
  If-Then Reasoning: (5) Pre-Condition and (6) Post-Condition (Sec. IV-C, V-C)
- **RQ4: Can AI models recognize the reasons of errors and infer the correct action?**
  (7) Rationale and (8) Correction Reasoning (Sec. IV-D, V-D)

We utilize LLMs & VLMs and formulate the tasks in the following manner. We treat a foundational model as π. We

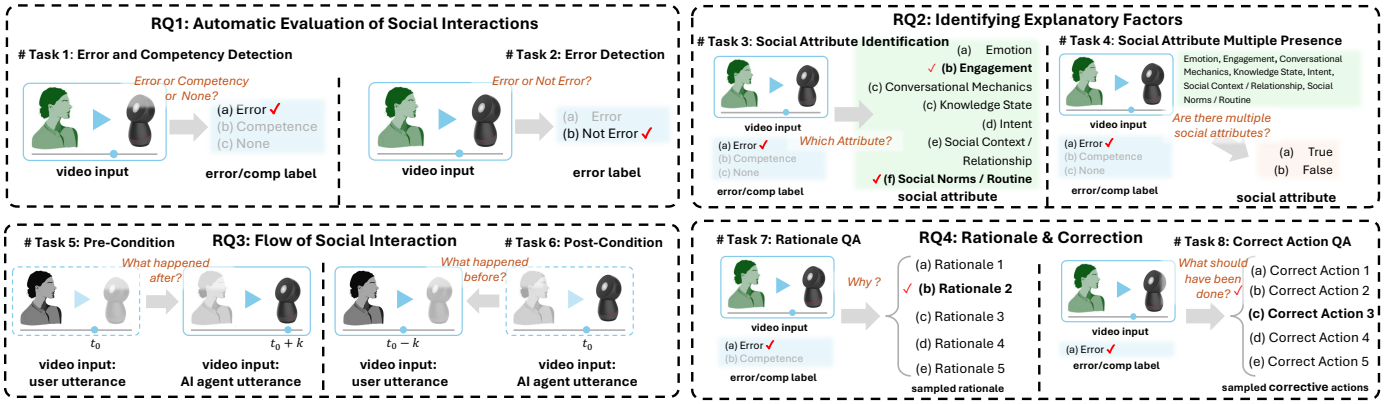

Fig. 4. Our benchmark offers eight tasks dedicated to probing four core facets of AI model's social reasoning: (1) detecting social errors and competencies, (2) identifying social attributes, (3) understanding the flow of social interactions, and (4) rationalization and correction of social errors.

define relevant contextual information (images and transcript) and the task-specific question as query Q. Formally, we define $Q = \{q_1, q_2, \ldots, q_n\}$ as the set of tokens representing the question and the transcript (including video or image frames if multimodal LLM). For the image-based models, we provide 15 uniformly sampled frames from the video segment. For Tasks 5 and 6, this window is restricted to the robot-only or user-only segment; for the other tasks, it is the full annotated segment. We used uniform sampling as a standardized baseline. For video-based models, we feed in the raw video as input and rely on model-specific preprocessing steps.

The output of the model is $\pi(Q) = O$, where O is the set of tokens representing the output of the LLM. We denote the ground-truth answer for each task as A, which can take the form of a multiple-choice option (A, B, C, D, E), a Boolean value (True/False). Since O is often a free-form string, we apply a post-processing step using Pydantic [8], where another LLM coerces O into the discrete answer space of the task (e.g., mapping "The correct choice is option C" $\rightarrow$ "C"). We then define correctness as $\text{Correctness}(A, O) = 1$ if $\text{Pydantic}(O) = A$, 0 otherwise. This definition ensures that correctness measures exact agreement with the ground truth rather than string overlap. **Human Comparison:** For fairness, we asked human annotators to perform the same tasks as the LLMs/VLMs (given identical prompts as instructions). Unlike the original annotations, where annotators watched videos and marked salient error/competency regions, here the annotator viewed the pre-segmented data and completed the same task as the models. We then compared their responses to the original ground-truth labels, yielding a estimate of human-level performance under same constraints.

A. Error and Competence

We use *Error Detection and Competence Detection* as a proxy to evaluate whether AI models can effectively serve as automatic evaluators of social interactions [38, 67, 33, 7, 66]. The model is provided with a prompt which includes an interaction segment between a user and a social robot. The model then predicts whether the sequence contains a social *Error* (moments in which the robot violates social norms in ways that degrade perceived social competence or the relationship with the user [52]), *Competence* (moments in which the robot is perceived as successfully conducting social interaction) [22]. In addition, we include *None* labels, which refers to segments where the annotator has neither labeled the video as 'competence' or 'error'. Social competence is widely understood as a multidimensional construct, and the levels of competence can vary on a spectrum [29]. Accordingly, we view segments labeled None are also meaningful, as they may reflect socially neutral, ambiguous, or weakly valenced behavior rather than a complete absence of social signal.

> **Social Error, Competence, None Detection (Error/-Comp./None):** We provide the interaction between social agent and a user: {*Interaction Transcript*} Does the agent exhibit (A) Social Competence or (B) Social Error or (C) None?
> **Social Error Detection (Error):** We provide the interaction between social agent and a user: {*Interaction Transcript*} Does the agent exhibit (A) Social Error or (B) No Social Error?

For the *Error Detection* task, the model is only required to determine whether a given instance constitutes an error or not. This binary task isolates the foundational ability to detect social error, which is the first decision point for safety and repair. This separation mirrors how reward models are used in RLHF [38], making the binary task a meaningful intermediate milestone. For these tasks, we evaluate with accuracy and macro-F1 (unweighted average of F1) scores.

B. Social Attribute

Additionally, if a sequence is identified as either a social error or competence, we further determine the explanatory factor, i.e. which specific *social attribute* it was associated with. This helps assess whether AI models can provide more detailed evaluations by identifying specific types of social attributes.

Social Attribute Identification (Attr.): We provide the interaction between social agent and a user: {*Interaction Transcript*}. This segment corresponds to an {*Error or Competence*} in social behavior. Which of the following categories is this segment related to? (A) Emotions, (B) Engagement, (C) Conversational Mechanics, (D) Knowledge State of Others and Self, (E) Understanding Intention of the User, (F) Social Context and Relationships, (G) Social Norms and Routines.

Furthermore, fine-grained feedback in the form of attribute-specific reward signals allows us to disentangle different dimensions of social behavior, such as emotional response, conversational mechanics, or knowledge state, thereby enabling more interpretable and targeted model improvements [57]. By aligning rewards with these distinct social attributes, we can not only evaluate whether a model exhibits socially competent behavior but also pinpoint which aspect it succeeded or failed in, facilitating modular training and fine-tuning strategies. Using the same context as Section IV-A, Models are further provided with the label indicating whether the instance is a social error or competence, and the model predicts the relevant attribute(s) as defined in Section III. This task can be viewed as a multi-label classification problem, as multiple social attributes may co-occur in a single instance. For example, a sequence might be annotated with both conversational mechanics (e.g., a delayed response) and knowledge state (e.g., the robot forgetting the user's name). To evaluate whether models can detect instances associated with more than one social attribute, we introduce the multiple social attribute detection task.

Multiple Social Attribute Presence (Multi. Attr.): We provide a transcript of an interaction between the social agent and a user: {*Interaction Transcript*}. The agent's behavior in this interaction corresponds to {*Error or Competence*}. Consider the following seven social attributes: [Same as Above]. Based on the transcript, determine whether the agent's behavior involves multiple social attributes. Respond with "True" if the behavior demonstrates more than one social attribute. Respond with "False" otherwise.

In our setup, the model predicts all applicable social attributes jointly in a single inference, rather than making a separate prediction for each attribute. For the task of social attribute identification, we evaluate with accuracy and macro-F1 (F1) scores. As there can be more than a single attribute label associated with a sample, we further report Partial Match (PM) to evaluate the proportion of instances where the model correctly predicts at least one of the true labels. For multiple social attribute detection, we evaluate with accuracy and macro-F1.

C. If-Then Reasoning

We test whether or not AI models can understand sequential contingencies or the flow of social interactions, by testing if they can predict probable pre-and-post conditions of a given competent social interaction, otherwise known as *if-then reasoning*. We view this capability as central to social reasoning, since social competence requires understanding how interactions unfold over time and how one social state gives rise to another. A model that can do this is less likely to rely purely on superficial correlations and more likely to have learned a transferable representation of interaction dynamics, which has been shown widely in NLP literature [10, 31, 44]. We formulate two tasks which checks for the pre-conditons and post-conditions.

Interaction Flow (Pre-Condition\Post-Condition): We provide the agent\user's behavior: {*Agent\User Transcript*}. From the following choices of the user\agent's behaviors: (1) {*User\Agent Transcript 1*}, (2) {*User\Agent Transcript 2*}, (3) {*User\Agent Transcript 3*}, (4) {*User\Agent Transcript 4*}, (5) {*User\Agent Transcript 5*}, select which user\agent's behavior was the appropriate response before\after the agent\user's action.

For *pre-condition*, given the agent's response, the model must identify the plausible pre-condition, i.e., the user's behavior prior to the agent's utterance. Vice versa, for *post-condition*, given the user's response, the model must predict the plausible post-conditions, i.e., the agent's actions after the event. These tasks are set-up as a multiple choice Q&A set up, where they are asked to predict the correct choice of post or pre-condition transcript. To acquire incorrect answer choices, first, we remove any samples that share the same transcript as the correct answer, if such samples exist. Next, we randomly select four other samples and extract the relevant transcript (either the user's or the agent's utterance) which serve as incorrect options. Finally, we shuffle the correct answer among the options to ensure its position is randomized. For these tasks, we use accuracy scores for evaluation.

D. Rationale and Correction

We formulate two tasks, *rationale*, which require models to provide rationale on why the segment contains a social error and *correction*, requiring models to suggest what would have been the correct action. These tasks probes the model's ability to not just diagnose an interaction segment but also to explain why the detected behavior is incorrect or inappropriate and identify a correct alternative action. Hence, they are directly tied to evaluating the social reasoning capabilities of AI models. Furthermore, it is well-known that models and robots that can provide rationales and correct actions enhance their robustness [43, 34] and trustworthiness [30, 25, 15].

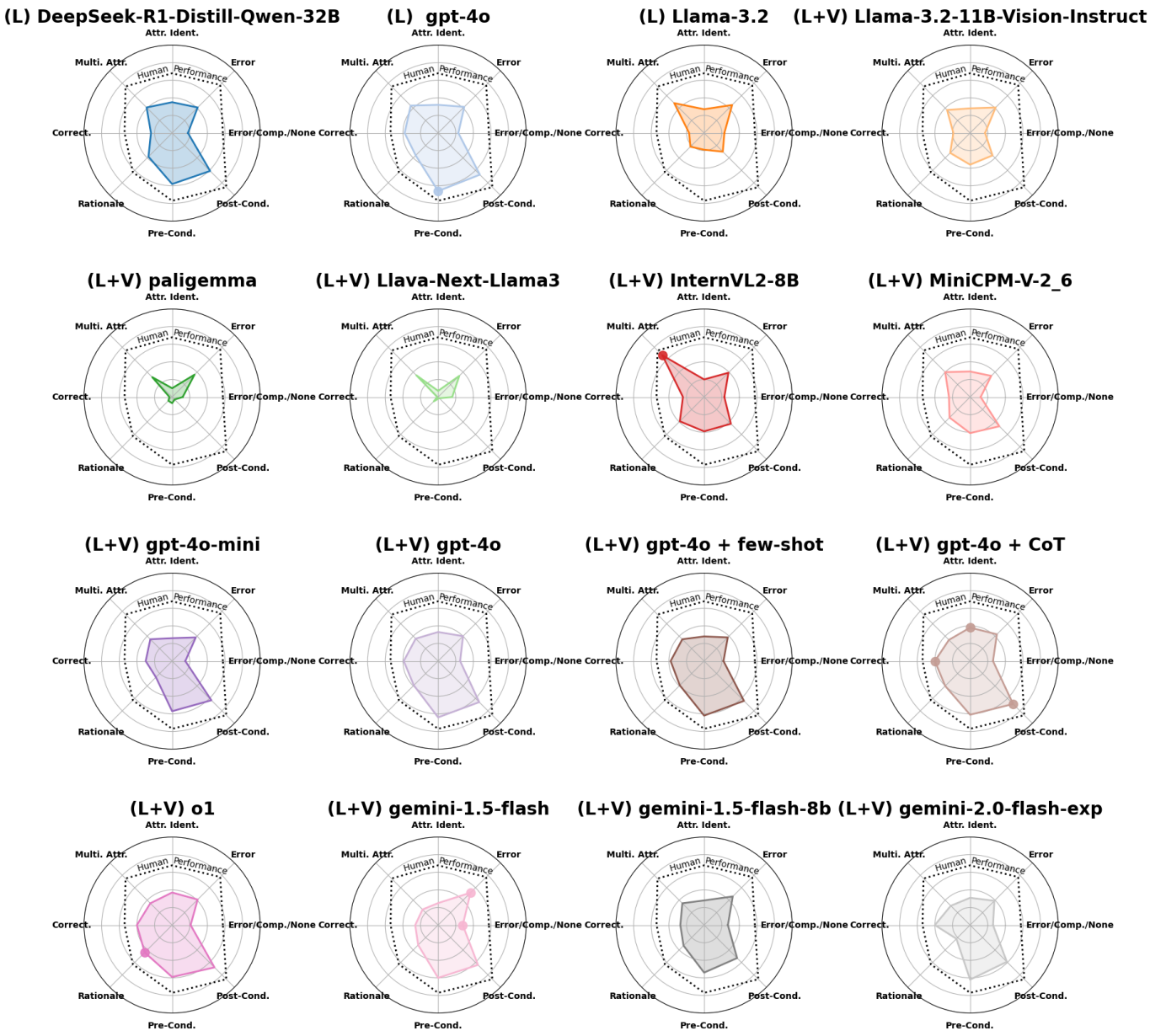

Fig. 5. Results per model across all 8 tasks. Human performance is marked in dashed lines. (L): language-only inputs, (L+V): language and visual inputs. ***Models from the same family tend to show similar shapes on the radar plot, reflecting consistent reasoning patterns across capabilities, which supports the benchmark's sensitivity to measure underlying social reasoning abilities.***

Rationale: We provide the interaction between social agent and a user, which corresponds to an error in social behavior: {*Interaction Transcript*}. Select which is the correct rationale behind the error. (1) {*rationale 1*} (2) {*rationale 2*} (3) {*rationale 3*} (4) {*rationale 4*} (5) {*rationale 5*}

Correction: We provide the interaction between social agent and a user, which corresponds to an error in social behavior: {*Interaction Transcript*}. From the following choices select which behavior the social agent should have done instead. (1) {*Correction 1*} (2) {*Correction 2*} (3) {*Correction 3*} (4) {*Correction 4*} (5) {*Correction 5*}

For the *rationale* and *correction* task, models are given interaction segments corresponding to social error. Then, for the *rationale* task, the model is asked to predict the correct reason for the error. For the *correction* task, the model must choose the alternate correct action. Both tasks are set-up as multiple choice Q&As, we provide five answer choices: one ground-truth option and four incorrect choices sampled from other instances. To ensure the incorrect choices are distinct, we only select samples with different social attribute annotations

Model Name	Errors and Competence				Social Attribute				
	Error/Comp./None		Error		Attr. Identification			Multi. Attr. Presence	
	Acc.	F1	Acc.	F1	Acc.	F1	Partial	Acc.	F1
(L) DeepSeek-R1-Distill-Qwen-32B	0.22±0.02	0.18±0.02	0.51±0.02	0.41±0.01	0.02±0.01	0.35±0.01	0.67±0.02	0.52±0.01	0.41±0.03
(L) gpt-4o-mini	0.26±0.02	0.21±0.02	0.52±0.02	0.46±0.01	0.01±0.01	0.17±0.01	0.41±0.04	0.46±0.02	0.40±0.03
(L) gpt-4o	0.25±0.02	0.23±0.02	0.50±0.02	0.42±0.02	0.02±0.01	0.20±0.01	0.49±0.03	0.52±0.04	0.44±0.06
(L) Llama-3.2	0.26±0.01	0.23±0.00	0.50±0.02	0.45±0.02	0.02±0.02	0.18±0.04	0.43±0.07	0.55±0.04	0.48±0.04
(L) Llama-3.2-11B-Vision-Instruct	0.22±0.04	0.17±0.03	0.50±0.05	0.41±0.04	0.01±0.01	0.24±0.01	0.56±0.03	0.50±0.02	0.37±0.03
(L+V) paligemma	0.19±0.02	0.12±0.01	0.49±0.02	0.36±0.02	0.00±0.00	0.11±0.02	0.27±0.04	0.53±0.03	0.32±0.02
(L+V) Llava-Next-Llama3	0.19±0.03	0.16±0.03	0.49±0.03	0.34±0.02	0.01±0.01	0.09±0.02	0.25±0.03	0.51±0.03	0.35±0.02
(L+V) InternVL2-8B	0.24±0.04	0.23±0.04	0.49±0.04	0.39±0.04	0.01±0.01	0.21±0.02	0.51±0.06	**0.67±0.02**	**0.67±0.02**
(L+V) MiniCPM-V-2_6	0.18±0.02	0.12±0.01	0.48±0.03	0.34±0.02	0.00±0.00	0.27±0.02	0.58±0.04	0.51±0.01	0.40±0.08
(L+V) gpt-4o-mini	0.20±0.03	0.15±0.03	0.49±0.02	0.38±0.01	0.03±0.03	0.17±0.01	0.42±0.02	0.50±0.01	0.35±0.01
(L+V) gpt-4o	0.26±0.03	0.25±0.03	0.50±0.02	0.40±0.02	0.01±0.02	0.21±0.02	0.51±0.04	0.49±0.01	0.36±0.01
(L+V) gpt-4o + few-shot	0.29±0.04	**0.28±0.03**	0.40±0.03	0.35±0.03	0.04±0.01	**0.45±0.01**	**0.82±0.02**	0.49±0.01	0.34±0.01
(L+V) gpt-4o + cot	0.27±0.03	0.26±0.03	0.50±0.02	0.43±0.02	0.04±0.01	0.33±0.02	0.64±0.06	0.49±0.01	0.34±0.01
(L+V) o1	0.24±0.04	0.21±0.04	0.50±0.02	0.41±0.02	0.02±0.02	0.32±0.02	0.67±0.02	0.49±0.01	0.35±0.02
(L+V) gemini-1.5-flash	**0.32±0.01**	**0.28±0.02**	**0.55±0.02**	**0.52±0.01**	0.01±0.02	0.17±0.02	0.43±0.05	0.45±0.03	0.25±0.02
(L+V) gemini-1.5-flash-8b	0.28±0.01	0.27±0.02	0.51±0.02	0.46±0.02	0.03±0.02	0.24±0.02	0.72±0.04	0.52±0.02	0.35±0.02
(L+V) gemini-2.0-flash-exp	0.27±0.05	0.26±0.04	0.49±0.03	0.39±0.02	0.01±0.00	0.28±0.01	0.62±0.03	0.47±0.03	0.31±0.02
(L+V) gemini-2.5-pro	0.29±0.02	0.27±0.02	0.40±0.01	0.33±0.02	0.01±0.01	0.22±0.03	0.55±0.02	0.48±0.01	0.25±0.02
Human	0.57	0.45	0.77	0.71	0.24	0.67	0.76	0.57	0.75

TABLE I
RESULTS FOR **ERRORS AND COMPETENCE** DETECTION TASKS AND **SOCIAL ATTRIBUTE** IDENTIFICATION TASKS, WITH ± INDICATING ONE STANDARD DEVIATION. (L): LANGUAGE-ONLY, (L+V): LANGUAGE AND VISION.

than that of the answer sample. These tasks are evaluated with accuracy.

V. RESULTS & DISCUSSION

We evaluate the performance of 18 language and vision-language models (VLMs), including `DeepSeek-R1-Qwen-32B` [21], `gpt-4o-mini`, `gpt-4o` [23], `Llama-3.2`, `Llama-3.2-Vision-Instruct` [20], `paligemma` [4], `Llava-Next-Llama3` [65], `InternVL2-8B` [6], `MiniCPM` [59], `o1` [24], `gemini-1.5` [50] and `gemini-2.0` [41] and their variants on our benchmark tasks. Fig. 5 summarizes the evaluation results. No single model excels across all social reasoning tasks, underscoring the need for advancements in training FMs for social reasoning.

A. RQ1: Can AI models be used as automatic evaluators of social interactions? [Tab. I-Left]

We refer readers to the left side of Table I, which presents results for the *Error/Competence/None Detection* task. The best-performing models are `gemini-1.5-flash` using both language and video inputs, achieving 0.32 accuracy and 0.28 F1. For the binary *Error Detection* task, the same model achieve higher performance—0.55 exact match and 0.52 F1—suggesting that models are more effective at detecting errors alone than distinguishing among all three categories. Nonetheless, the overall performance remains modest, reflecting the task's complexity and the models' difficulty in capturing the nuances of social error and competence. Appendix C further analyzes model performance by social attribute, revealing that some models excel in detecting specific types of errors. These findings underscore the gap between current AI capabilities and human-level social reasoning, pointing to the need for continued research in this area.

B. RQ2: Can AI models identify the explanatory factors associated with social errors and social competencies? [Tab. I-Right]

In Table I, under the column *Attr. Identification*, we evaluate the ability of foundational models to identify explanatory factors in the form of social attributes. We report accuracy, F1, and partial accuracy (i.e., predicting at least one attribute correctly). Notably, most models perform well on partial accuracy, demonstrating the ability to identify at least one relevant attribute. In particular, `gpt-4o-fewshot` with image input achieve partial accuracy scores above 0.80, highlighting their relative strength in capturing aspects of social attribute prediction. However, all models struggle to predict the full and correct set of attributes, as reflected in low accuracy and modest F1 scores, which indicates FMs remain limited in handling the complex, multi-label, and co-occurring nature of social attribute classification. Human evaluation further reveals that, despite a clear performance gap, this remains a challenging task even for humans. To further probe these limitations, we refer readers to the right-most columns of Table I, where we assess whether models can detect the presence of multiple attributes within a segment. Interestingly, `InternVL2` achieves the highest accuracy and F1 in this setting. However, the majority of models perform at or below 0.5, indicating persistent difficulty in recognizing multi-attribute cases—a key failure mode in this task. Additional analysis is provided in Appendix C, which examines the role of subjectivity and co-occurrence, as social attributes often co-occur in a single segment and are subject to perceiver-dependent interpretations of social constructs [46, 35]. Furthermore, in Appendix C, Fig. 9, we carry out further analysis to identify which attributes are easier for error detection.

Model Name	If-Then Reasoning		Rationale & Correct.	
	Pre-Condition	Post-Condition	Rationale	Correction
	Acc.	Acc.	Acc.	Acc.
(L) DeepSeek-R1-Distill-Qwen-32B	0.58±0.03	0.61±0.02	0.38±0.05	0.24±0.01
(L) gpt-4o-mini	0.58±0.02	0.62±0.03	0.29±0.04	0.31±0.08
(L) gpt-4o	**0.66±0.00**	0.67±0.02	0.36±0.01	0.38±0.04
(L) Llama-3.2	0.19±0.08	0.30±0.01	0.22±0.02	0.17±0.02
(L) Llama-3.2-11B-Vision-Instruct	0.36±0.05	0.36±0.05	0.32±0.05	0.19±0.02
(L+V) paligemma	0.07±0.02	0.04±0.03	0.06±0.05	0.03±0.01
(L+V) Llava-Next-Llama3	0.02±0.01	0.02±0.01	0.07±0.02	0.01±0.01
(L+V) InternVL2-8B	0.39±0.09	0.43±0.02	0.39±0.05	0.24±0.03
(L+V) MiniCPM-V-2_6	0.41±0.07	0.47±0.02	0.33±0.05	0.24±0.05
(L+V) gpt-4o-mini	0.57±0.05	0.63±0.04	0.26±0.04	0.30±0.06
(L+V) gpt-4o	0.64±0.05	0.66±0.05	0.39±0.08	0.39±0.08
(L+V) gpt-4o + few-shot	0.57±0.06	0.67±0.01	**0.55±0.06**	**0.48±0.07**
(L+V) gpt-4o + cot	0.61±0.05	**0.69±0.05**	0.40±0.03	0.40±0.05
(L+V) o1	0.59±0.08	0.68±0.02	0.44±0.10	0.40±0.03
(L+V) gemini-1.5-flash	0.60±0.10	0.64±0.07	0.32±0.04	0.26±0.07
(L+V) gemini-1.5-flash-8b	0.54±0.09	0.53±0.06	0.33±0.06	0.27±0.04
(L+V) gemini-2.0-flash-exp	0.62±0.07	0.59±0.09	0.22±0.04	0.40±0.05
(L+V) gemini-2.5-pro	0.55±0.09	0.55±0.06	0.49±0.06	0.45±0.08
Human	0.77	0.87	0.63	0.54

TABLE II
IF-THEN REASONING & RATIONALE & CORRECTION RESULTS; ± INDICATES ONE STD. DEV. (L): LANGUAGE-ONLY, (L+V): LANGUAGE & VISION.

C. RQ3: Do AI models understand the sequential contingencies or the "flow" of social interactions? [Tab. II-Left]

Many large language models, including early versions of BERT [13] and ALBERT [32], were trained with next sentence prediction (NSP) and sentence ordering objectives, which aim to model discourse coherence and temporal continuity by predicting whether one sentence logically follows another. This objective aligns with the structure of *if-then reasoning* tasks, partially explaining the relatively strong performance of language-only models—particularly the `gpt-4o` variants, which achieve 0.66 on *pre-condition* and 0.69 on *post-condition* inference. However, their performance still lags behind humans. Despite progress in temporal reasoning [44], we observe a significant gap between nascent open-source vision-language models (VLMs) and even text-only LMs, suggesting that visual input alone does not guarantee better reasoning about interactional contingencies. Recent studies have shown that adding vision to a strong LLM can degrade its original text-only reasoning abilities, particularly on out-of-distribution benchmarks [9, 64].

D. RQ4: Can AI models recognize the reasons of errors and infer the correct action? [Tab. II-Right]

The *rationale* and *correction* tasks evaluate a model's ability to interpret interaction context, infer the cause of social errors or competencies, and suggest appropriate alternative correct actions, making them key indicators of social reasoning. These are among the most challenging tasks in the dataset, as reflected by human-level difficulty. For the **rationale** task, the `gpt-4o-fewshot` model performs best with a score of 0.55. For the **correction** task, `gpt-4o-fewshot` again perform well, each achieving a score of 0.48. Few-shot prompting may help by making the dataset's normative decision boundaries more explicit: the examples illustrate which interaction cues and conversational moves annotators tend to treat as competent (or as violations) for a given error type, providing exemplars can better align outputs with the dataset's social error taxonomy.

E. Limitations and Future Directions

SHREC is a benchmark for socially situated reasoning grounded in real human–robot conversations, but it inherits limitations from both the underlying deployments and our benchmark abstraction. First, social error/competence labels reflect human normative judgments: while we mitigate idiosyncrasy through overlapping annotations and agreement checks, conversational norms can vary across cultures, contexts, and neurodivergent communication styles, and models trained or evaluated on SHREC may learn these decision boundaries rather than universally valid social rules.

Second, our interactions are drawn from a tabletop social robot setting and a limited set of long-term studies; generalization to other embodiments (e.g., mobile manipulators), languages, and interaction contexts remains an open question.

Third, our released benchmark targets the operating regime of current VLM evaluation (1 Hz frame sampling), which can under-represent fine-grained timing cues (e.g., micro-pauses, overlap, gaze shifts) that are important for conversational mechanics and turn-taking. Finally, our most diagnostic tasks—multi-attribute cases and rationale/correction—remain challenging, suggesting that progress will require better multimodal grounding, stronger counterfactual action modeling, and more explicit alignment to social preferences. We view SHREC as a step toward standardized evaluation and training signals for socially competent embodied agents, and an invitation to broaden coverage via more diverse deployments and richer temporal/behavioral sensing.

Fourth, while we focus on predicting attributes all at once, task formulation focusing on individual attributes, such as detecting engagement only, would be highly valuable. SHREC already supports this type of analysis: i.e. the dataset contains more than 2,640 engagement instances alone as shown in Fig. 2, providing sufficient coverage for future work to benchmark construct-specific recognition. More generally, the underlying annotations make it possible to decompose the problem either jointly across attributes or separately by individual construct, including sequential yes/no formulations. In this paper, our aim was to establish an initial broad benchmark and baseline for social reasoning across multiple dimensions, rather than to fully enumerate all possible task decompositions or variants.

VI. CONCLUSION

We present the Social Human Robot Embodied Conversation (SHREC) Dataset, consisting of 400+ real-world interaction videos, 10,000+ annotations and eight new benchmark tasks spanning error detection, attribute reasoning, interaction flow, and rationale/correction inference, SHREC offers a critical resource to address unique socio-cognitive limitations of embodied agents. Through systematic evaluation of state-of-the-art LLMs and VLMs, we find that their overall performance falls far short from human-level, which underscores the limitations of current models in social reasoning. We envision SHREC as a foundation for advances in social reasoning for embodied social AI agents.

ACKNOWLEDGMENTS

This research was supported by the National Research Council of Science & Technology (NST) grant funded by the Korea government (MSIT) (No. GTL25041-000). LPM is partially supported by the National Institutes of Health under awards R01MH125740, R01MH132225, U01MH136535, and R21MH130767. DWL is partially supported by the Amazon AI Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred.

REFERENCES

[1] Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. Talking turns: Benchmarking audio foundation models on turn-taking dynamics. *arXiv preprint arXiv:2503.01174*, 2025.

[2] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596, 2008.

[3] Simon Baron-Cohen, Michelle O'Riordan, Rosie Jones, Valerie Stone, and Kate Plaisted. A new test of social sensitivity: Detection of faux pas in normal children and children with asperger syndrome. *Journal of Autism and Developmental Disorders*, 29(5):407–418, 1999.

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[5] Sree Bhattacharyya and James Z. Wang. Evaluating vision-language models for emotion recognition. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1798–1820, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.97/.

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

[7] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

[8] Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, Alex Hall, and Victorien Plot. Pydantic, 1 2025. URL https://docs.pydantic.dev/latest/.

[9] Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, et al. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*, 2025.

[10] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

[11] Mark H Davis. Interpersonal reactivity index. *Journal of Personality and Social Psychology*, 1980.

[12] Katherine Deng, Arijit Ray, Reuben Tan, Saadia Gabriel, Bryan A Plummer, and Kate Saenko. Socratis: Are large multimodal models emotionally aware? *arXiv preprint arXiv:2308.16741*, 2023.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[14] Isabel Dziobek, Stefan Fleck, Elke Kalbe, Kimberley Rogers, Jason Hassenstab, Matthias Brand, Josef Kessler, Jan K Woike, Oliver T Wolf, and Antonio Convit. Introducing masc: a movie for the assessment of social cognition. *Journal of autism and developmental disorders*, 36:623–636, 2006.

[15] Connor Esterwood and Lionel P Robert. Repairing trust in robots?: A meta-analysis of hri trust repair studies with a no-repair condition. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 410–419. IEEE, 2025.

[16] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2022.

[17] Riccardo Fusaroli and Kristian Tylén. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive science*, 40(1):145–171, 2016.

[18] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in psychology*, 6:931, 2015.

[19] Ofer Golan, Simon Baron-Cohen, Jacqueline J Hill, and Yael Golan. The "reading the mind in films" task: complex emotion recognition in adults with and without autism spectrum conditions. *Social Neuroscience,*, 1(2):111–123, 2006.

[20] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

[22] Amy G Halberstadt, Susanne A Denham, and Julie C Dunsmore. Affective social competence. *Social development*, 10(1):79–119, 2001.

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

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

[25] Misbah Javaid and Vladimir Estivill-Castro. Explanations from a robotic partner build trust on the robot's decisions for collaborative human-humanoid interaction. *Robotics*, 10(1):51, 2021.

[26] Sooyeon Jeong, Sharifa Alghowinem, Laura Aymerich-Franch, Kika Arias, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. A robotic positive psychology coach to improve college students' wellbeing. In *2020 29th IEEE international conference on robot and human interactive communication (RO-MAN)*, pages 187–194. IEEE, 2020.

[27] Sooyeon Jeong, Laura Aymerich-Franch, Sharifa Alghowinem, Rosalind W Picard, Cynthia L Breazeal, and Hae Won Park. A robotic companion for psychological well-being: A long-term investigation of companionship and therapeutic alliance. In *Proceedings of the 2023 ACM/IEEE international conference on human-robot interaction*, pages 485–494, 2023.

[28] Sooyeon Jeong, Laura Aymerich-Franch, Kika Arias, Sharifa Alghowinem, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. Deploying a robotic positive psychology coach to improve college students' psychological well-being. *User Modeling and User-Adapted Interaction*, 33(2):571–615, 2023.

[29] Caroline Junge, Patti M Valkenburg, Maja Deković, and Susan Branje. The building blocks of social competence: Contributions of the consortium of individual development. *Developmental cognitive neuroscience*, 45:100861, 2020.

[30] Esther S Kox, José H Kerstholt, Tom F Hueting, and Peter W de Vries. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous agents and multi-agent systems*, 35(2):30, 2021.

[31] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40: e253, 2017.

[32] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[33] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

[34] Josh Magnus Ludan, Yixuan Meng, Tai Nguyen, Saurabh Shah, Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Explanation-based finetuning makes models more robust to spurious cues. *arXiv preprint arXiv:2305.04990*, 2023.

[35] Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. Advancing social intelligence in ai agents: Technical challenges and open questions. *arXiv preprint arXiv:2404.11023*, 2024.

[36] Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. *arXiv preprint arXiv:2502.15109*, 2025.

[37] Anastasia K Ostrowski, Cynthia Breazeal, and Hae Won Park. Mixed-method long-term robot usage: Older adults' lived experience of social robots. In *2022 17th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 33–42. IEEE, 2022.

[38] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[39] Hae Won Park, Rinat Rosenberg-Kima, Maor Rosenberg, Goren Gordon, and Cynthia Breazeal. Growing growth mindset with a social robot peer. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pages 137–145, 2017.

[40] Hae Won Park, Cynthia Breazeal, Sharifa Alghowinem, Anastasia K Ostrowski, Jon Ferguson, Xiajie Zhang, and Dong Won Lee. Jibo community social robot research platform@ scale. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1346–1348, 2024.

[41] Sundar Pichai, D Hassabis, and K Kavukcuoglu. Introducing gemini 2.0: our new ai model for the agentic era, 2024.

[42] 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 Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

[43] Alexis Ross, Matthew E Peters, and Ana Marasović. Does self-rationalization improve robustness to spurious

correlations? *arXiv preprint arXiv:2210.13575*, 2022.

[44] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035, 2019.

[45] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[46] John R Searle. Social ontology and the philosophy of society. *Analyse & Kritik*, 20(2):143–158, 1998.

[47] Natalie Shapira, Guy Zwirn, and Yoav Goldberg. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, 2023.

[48] Guangyao Shen et al. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4937–4945, 2020.

[49] Jocelyn Shen, Yubin Kim, Mohit Hulse, Wazeer Zulfikar, Sharifa Alghowinem, Cynthia Breazeal, and Hae Won Park. Empathicstories++: A multimodal dataset for empathy towards personal experiences. *arXiv preprint arXiv:2405.15708*, 2024.

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

[51] Peggy A Thoits. Emotion norms, emotion work, and social order. In *Feelings and emotions: The Amsterdam symposium*, pages 359–378. Cambridge University Press Cambridge, UK, 2004.

[52] Leimin Tian and Sharon Oviatt. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(2):1–32, 2021.

[53] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

[54] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958, 2023.

[55] Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. *Social-iq 2.0 challenge: Benchmarking multimodal social understanding*, 2023.

[56] Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, and Minlie Huang. Coke: A cognitive knowledge graph for machine theory of mind. *arXiv preprint arXiv:2305.05390*, 2023.

[57] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36: 59008–59033, 2023.

[58] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8703–8712, 2024.

[59] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[60] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.

[61] 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 (CVPR)*, pages 8807–8817, 2019.

[62] AmirAli Bagher Zadeh, Paul Pu Liang, Sahisnu Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[63] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[64] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

[65] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

[66] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[67] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence

in language agents, 2024. URL https://arxiv.org/abs/2310.11667.