

Toward Socially-Aware LLMs: A Survey of Multimodal Approaches to Human Behavior Understanding

ZIHAN LIU, University of Illinois Urbana-Champaign, USA

PARISA RABBANI, University of Illinois Urbana-Champaign, USA

VEDA DUDDU, University of Illinois Urbana-Champaign, USA

KYLE FAN, University of Illinois Urbana-Champaign, USA

MADISON LEE, University of Illinois Urbana-Champaign, USA

YUN HUANG, University of Illinois Urbana-Champaign, USA

LLM-powered multimodal systems are increasingly used to interpret human social behavior, yet how researchers apply the models' "social competence" remains poorly understood. This paper presents a systematic literature review of 176 publications across different application domains (e.g., healthcare, education, and entertainment). Using a four-dimensional coding framework (application, technical, evaluative, and ethical), we find (1) frequent use of pattern recognition and information extraction from multimodal sources, but limited support for adaptive, interactive reasoning; (2) a dominant "modality-to-text" pipeline that privileges language over rich audiovisual cues, stripping away nuanced social cues; (3) evaluation practices reliant on static benchmarks, with socially grounded, human-centered assessments rare; and (4) Ethical discussions focused mainly on legal and rights-related risks (e.g., privacy), leaving societal risks (e.g., deception) overlooked—or at best acknowledged but left unaddressed. We outline a research agenda for evaluating socially competent, ethically informed, and interaction-aware multi-modal systems.

CCS Concepts: • ;

1 Introduction

Human social interaction is inherently multimodal, unfolding through complex combinations of verbal utterances, facial expressions, gestures, gaze patterns, and vocal prosody [74]. Understanding these rich social signals is not just a perceptual challenge; instead, it lies at the heart of what makes humans socially intelligent. Social intelligence, broadly defined as the ability to perceive, interpret, and act appropriately in social contexts [143, 145], is a foundational dimension of human intelligence and a critical capability for AI systems that aim to interact meaningfully with people [43, 74, 159].

The emergence of AI systems that integrate Large Language Models (LLMs) with multimodal perception capabilities marks a significant shift in how machines can understand human behavior. Unlike traditional approaches that rely on isolated modality-specific model [81], LLM-powered multimodal systems leverage the rich semantic understanding and reasoning abilities that LLMs have developed through large-scale language training. Drawing an analogy to human perception, modality modules such as visual or audio encoders act like eyes and ears, capturing and pre-processing visual and acoustic signals, while the LLM serves as the brain, integrating these inputs and reasoning over their social meaning [167]. For example, LLM-powered multimodal systems are being used to analyze student engagement patterns in educational settings [169], assess social development in children with autism spectrum disorders [29], and summarize collaborative strategies in team sports [73].

Authors' Contact Information: Zihan Liu, University of Illinois Urbana-Champaign, USA; Parisa Rabbani, University of Illinois Urbana-Champaign, USA; Veda Duddu, University of Illinois Urbana-Champaign, USA; Kyle Fan, University of Illinois Urbana-Champaign, USA; Madison Lee, University of Illinois Urbana-Champaign, USA; Yun Huang, University of Illinois Urbana-Champaign, USA.

However, little is known about how researchers have applied the models’ social competence in designing these systems. Several challenges limit our understandings of the research landscape. First, multiple research traditions approach this challenge from diverging perspectives: psychologists emphasize multidimensional constructs [25, 155], AI researchers build technical systems around isolated cues [109], and evaluation remains dominated by static benchmarks [4]. Because of this fragmentation, there is no coherent framework for AI system designers, developers, and evaluators to understand what “social intelligence” means for AI in multimodal systems, how it is technically realized, and how it should be assessed. Moreover, a disconnection persists between technical and applied communities: AI research emphasizes benchmark progress, while fields such as HCI, education, and healthcare lack clear frameworks for interpreting what these systems can and cannot do in socially meaningful contexts. These gaps underscore the need for a systematic review that bridges disciplinary boundaries and provides a structured account of how LLM-powered multimodal systems engage with human social behavior.

This paper addresses the gap through a systematic literature review of LLM-powered multimodal systems that are designed to understand human social behavior. Our analysis synthesizes research across technical venues in AI and applied communities such as HCI, education, and healthcare. Specifically, we examine the following four perspectives:

- **RQ1** (application): *What social intelligence of LLM-powered multimodal systems is applied to human social behavior analysis across different contexts?*
- **RQ2** (technical): *How is LLM-powered multimodal systems social intelligence technically operationalized?*
- **RQ3** (evaluative): *How is the performance of socially intelligent LLM-powered multimodal systems evaluated?*
- **RQ4** (ethical): What are the primary ethical challenges and risks associated with socially intelligent LLM-powered multimodal systems?

This paper makes a timely contribution to the HCI community by presenting a thorough review of LLM-enabled social intelligence applications. We synthesize, systematize, and critically assess this fractured landscape, yielding important implications for design, technical, and applied communities:

- **A Systematic, Interdisciplinary Synthesis of 176 Studies.** We provide the first large-scale literature review of LLM-powered multimodal systems for human social behavior understanding, bridging fragmented work across AI, HCI, education, and healthcare to reveal patterns, gaps, and opportunities.
- **A Four-Dimensional Coding Framework for Social Intelligence:** We introduce and apply a four-dimensional coding framework—covering *application*, *technical*, *evaluative*, and *ethical* perspectives—to systematically characterize how social intelligence is operationalized in LLM-powered multimodal systems. This framework offers researchers a structured vocabulary to guide future system design, benchmarking, and governance.
- **Empirical Insights into the Current Landscape:** Our analysis reveals several key takeaways: (1) Nearly all these research focus on social perception (100%) and reasoning (95%), while social interaction (41.8%) and creativity (18%) are rarely implemented; (2) a *modality-to-text bottleneck* that compresses rich multimodal signals into transcripts or keyframes before reasoning, losing rich multimodal cues; (3) evaluation practices that remain *machine-centric*, with relatively few longitudinal, interactive, or human-centered studies; and (4) 45% of reviewed papers do not mention ethics and risks.
- **Research Agendas for Socially Competent and Ethical LLM-powered Multimodal Systems:** Building on these findings, we chart a research agenda emphasizing (1) richer multimodal integration that preserves more modality information—prosody and gaze, etc; (2) development of socially grounded, human-centered, and longitudinal evaluation protocols; (3) diversification of model families and adoption of hybrid operationalization

strategies; and (4) systematic mitigation of fairness, bias, and misuse risks beyond privacy, enabling accountable and socially situated AI behavior. We also consolidate a directory of benchmarks and datasets (Table 1), mapping human goals to computational tasks and evaluation resources, which can be used to evaluate LLM-powered multimodal systems comprehensively.

2 Related Work

2.1 Social Intelligence and Social AI

The concept of social intelligence was first introduced by Thorndike (1920) as the ability to “understand and manage men and women, boys and girls, and to act wisely in human relations” [143]. Vernon defined social intelligence as “knowledge of social matters and insight into the moods or personality traits of strangers” and as the ability to “get along with others and ease in society” [145]. Importantly, both of these two definitions highlighted a dual nature: a cognitive facet (understanding others) and a behavioral facet (acting effectively in social situations). Building on this, subsequent definitions alternated between the two aspects [58, 155]. For instance, Wedeck [153] conceptualizes social intelligence from a cognitive perspective, defining it as “the capacity to judge correctly the feelings, moods, and motivations of individuals.”

Building on these psychological foundations, researchers have increasingly asked how such social competencies might be implemented in computational systems, giving rise to the field of social AI or artificial social intelligence (ASI). Social AI is often defined as the pursuit of “socially-intelligent AI agents” [74], while ASI is framed as the capability to help these agents calibrate the outputs of their internal models to be understood by humans [163]. Scholars identify three key capabilities necessary for effective social understanding: the ability to interpret multimodal social cues, to reason about multi-party dynamics, and to infer beliefs or mental states [74]. Mathur et al. (2024) further argues that building socially intelligent AI requires agents that can not only sense and reason about human affect, behavior, and cognition, but also adapt dynamically across diverse social contexts [100]. Efforts to engineer ASI have thus focused on how social information is embodied in behavior. Researchers stress that an agent must grasp that intentions, emotions, and personalities are expressed through verbal and non-verbal cues, and therefore require integrated processing pipelines that link perception with social reasoning [43, 158, 159]. These discussions connect directly to broader questions of alignment and governance: just as individuals with high social intelligence can manage conflicts between personal and group objectives and avoid toxic behaviors, socially intelligent AI is envisioned as a path toward systems that are both norm-sensitive and collaborative [6, 72].

Despite these advances, the notion of social intelligence in AI remains fragmented and inconsistently defined. Psychology has long emphasized social intelligence as a multidimensional construct encompassing perception, memory, reasoning, and behavior [25]. However, AI research often narrows this scope to isolated elements such as theory of mind [158] or social cue recognition [74], without integrating them into a coherent framework. This conceptual gap motivates our review: to clarify how social intelligence has been framed in AI research and to prepare the ground for analyzing how systems have evolved from unimodal to multimodal approaches.

2.2 Social AI: from Uni-modal to Multimodal

Early efforts in social AI were largely unimodal [144] and primarily focused on descriptive recognition of isolated social cues [146]. Natural language processing systems classified utterances as polite or impolite, empathetic or neutral [127]. Computer vision models identified emotions from facial expressions or categorized gestures. Speech analysis

models mapped vocal prosody to affective states [40]. These systems demonstrated that isolated social cues could be computationally detected, but reduced social intelligence to recognition tasks, offering little capacity to interpret meaning or context.

The development of LLMs created a significant shift in AI social capabilities bringing explicit reasoning and interactive abilities [13]. Unlike unimodal classifiers that simply detected the presence of signals, LLM-based systems could leverage prior knowledge and existing instructions to draw inferences about moral reasoning, intent detection, and social commonsense [11, 108, 138]. From this perspective, LLMs enabled a move from answering the question "what signal is present?" to asking "why does this social signal matter in this context?", marking an important step toward artificial social intelligence [81]. At the same time, their reasoning remains fundamentally constrained: because these models operate only over textual descriptions, they may fail to capture the implicit cues, sarcasm, and non-verbal signals that are central to human social understanding [18, 20].

Overcoming the limitations of both classic unimodal models and text-only LLMs, recent research has turned toward multimodal AI systems that incorporate LLMs, aiming to integrate textual, visual, and auditory channels for more socially grounded understanding [60, 74]. These systems utilize LLMs as a reasoning brain and encode other data types, allowing the models to perceive and reason about the world while capturing the same information that humans do through senses [167]. As Li et al. observe, this reflects a broader paradigm shift in multimodal reasoning: from perception-driven modular systems, where reasoning was implicit within task-specific classifiers, to language-centric reasoning frameworks, where inference is articulated through structured prompts and extended chains of thought [81].

The trajectory from unimodal recognition to LLM-enabled reasoning to multimodal integration illustrates the rapid evolution of social AI. Yet the technical landscape remains fragmented: systems are often built as loosely coupled pipelines [81], with language reasoning dominating [100], and integration across modalities limited in depth [75]. These constraints highlight a technical gap between the richness of human social interaction and the limited architectures used to approximate it, setting the stage for our subsequent discussion of evaluation.

2.3 Evaluating Social Intelligence in AI Systems

The rise of multimodal AI systems that incorporate LLMs has been accompanied by numerous surveys cataloging their architectures, training techniques, and performance on general-purpose benchmarks [75, 82, 168]. However, existing surveys are often general or technically focused, lacking attention to social intelligence. When included in such surveys, social intelligence is often not properly evaluated to capture the dynamic, interactive, and context-dependent nature of real social situations. Without robust, socially grounded evaluation frameworks, we cannot truly measure the capabilities or limitations of these systems in understanding human social dynamics [47]. We risk developing models that perform well on simplified benchmarks but fail in real-world social interactions.

Current evaluation practices in the AI community have largely followed a technical benchmark tradition, breaking social intelligence down into focused, task-specific capabilities. Tasks such as emotion recognition or intent prediction are typically framed as classification problems, with performance reported as a single accuracy score [19, 77]. Early benchmarks like Social-IQ [170] provided models with video clips and multiple-choice questions, enabling quantitative comparisons but also allowing models to exploit dataset biases without engaging in genuine social understanding. More recent efforts have tried to address these shortcomings. SIV-Bench [71], for example, decomposes the evaluation into three components: Social Scene Understanding (SSU), Social State Reasoning (SSR), and Social Dynamics Prediction (SDP). SocialMaze [165] goes further by introducing multi-turn, game-like interaction scenarios such as hidden role deduction, aiming to capture more strategic forms of social reasoning. While these benchmarks represent progress, they

remain limited to structured task outcomes and continue to overlook the reciprocal, adaptive, and situated qualities that characterize real social intelligence [15, 172].

As Mathur et al. (2024) emphasize, evaluating artificial social intelligence involves challenges that go beyond technical benchmarking: the ambiguity of defining social constructs, the subtlety of multimodal signals, the need to represent multiple perspectives, and the requirement of agency and adaptation across contexts [100]. Current approaches fall short in each of these dimensions. This uneven landscape motivates our review: to synthesize how existing studies evaluate social intelligence in multimodal AI systems, to identify where evaluations succeed, and to highlight where socially grounded approaches are still lacking.

3 Method

This literature review aims to investigate the emerging capabilities, limitations, and evaluation practices associated with social intelligence in LLM-powered multimodal systems. We specifically focused on synthesizing research that explores how these systems perceive, process, and interpret human social interactions through externally observable, non-verbal social cues (e.g. visual and auditory signals).

Based on this focus, we defined the following inclusion criteria. (a) *Model Architecture*: The paper must describe a system or method that integrates a LLM component as part of its architecture. (b) *Multimodal social cues*: This system must be designed to process or analyze non-verbal social cues (e.g., gaze, facial expression, prosody, body posture) from vision or audio modalities in the context of understanding or engaging in human social interactions. (c) *Contribution Type*: We include papers that make a relevant contribution to the field of socially intelligent multimodal models.

We excluded papers that met any of the following categories: (a) *Text-Only Focus*: Works that relied solely on textual input without addressing nonverbal cues were excluded. (b) *Biosensor Data Emphasis*: We excluded studies that primarily rely on biosensor data—such as heart rate or skin conductance—captured via wearable devices (e.g., [128]). While such data is relevant in affective computing, it is not directly observable in typical face-to-face human interaction. Our review focuses on externally perceivable social cues (e.g., facial expressions, gestures, vocal tone) that can be interpreted through visual and auditory modalities in natural social settings.

Eventually, the eligible contributions encompass a range of types, including: the introduction of a new model or algorithm; the development of a system, application, or toolkit; the proposal of a conceptual framework, architecture, or processing pipeline; or the execution of an empirical study that evaluates or applies LLM-powered multimodal AI system in social contexts. Our methodology follows a three-stage process—Search, Selection, and Analysis—which is visually summarized in Figure 1. The subsequent subsections detail each of these stages.

3.1 Search

Our search strategy was structured around three main categories of keywords: (1) technical model terms, (2) social behavior terms, and (3) media modality terms. For the technical terms, we included variations such as: "*large language model*", "*vision language model*", "*visual language model*", and "*large multimodal model*". To capture work related to social intelligence, we used keywords describing social behavior, including: "*social interaction*", "*social behavior*", "*social cue*", "*social signal*", "*human interaction*", and "*human behavior*". Finally, we added media modality terms to narrow the scope to video-based analysis, using phrases such as: "*video understanding*", "*audio understanding*", "*audio analysis*", "*voice analysis*", "*voice understanding*", "*speech analysis*", and "*speech understanding*". These terms were combined using Boolean logic to construct flexible search queries, which were tailored to the syntax and capabilities of each database.

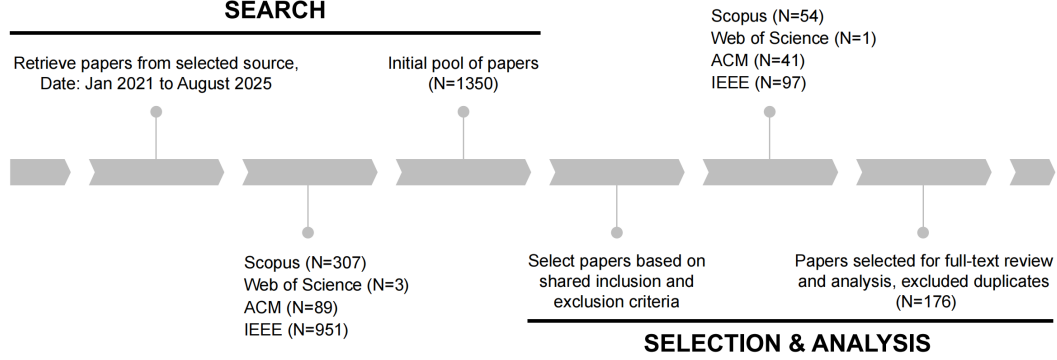


Fig. 1. Workflow of the Literature Search and Selection Process. This flowchart illustrates the four-stage methodology used to identify relevant papers. The process began with an initial pool of 1,350 papers from four databases, which was narrowed down based on inclusion and exclusion criteria, ultimately resulting in a final corpus of 176 papers for analysis.

We searched four databases to ensure comprehensive coverage: the Association for Computing Machinery Digital Library (ACM DL), Scopus, Web of Science, and IEEE. The initial search covered papers published from January 2021 to August 2025 and resulted in 89 papers from ACM DL, 307 from Scopus, 3 from Web of Science, and 951 from IEEE. After removing duplicates, there were a total of 1350 papers in the initial pool.

3.2 Screening and Selection

To identify relevant papers, we conducted a full-text manual review of each candidate based on the inclusion and exclusion criteria outlined earlier. Rather than relying solely on titles and abstracts, we carefully examined the full content of each paper to determine whether it involved the use of a LLM-powered multimodal system to interpret or respond to externally observable nonverbal social signals. Notably, many papers addressed aspects of social interaction without explicitly framing their work as achieving "social intelligence." These were still considered if the system demonstrated relevant competencies. This screening process resulted in a final selection of 176 papers for analysis.

3.3 Coding Procedure

To scale our qualitative analysis while preserving rigor, we adopted a human-in-the-loop, LLM-assisted coding workflow, drawing on recent best practices for AI-supported qualitative research (e.g., [28, 102, 161]). Five authors collaboratively coded an initial set of 15 papers to develop a structured codebook across four focal dimensions: application, technical, evaluative, and ethical. This human-coded set served as the ground truth for subsequent AI-assisted analysis.

We iteratively designed few-shot prompting templates and used the Gemini 2.5 Pro model (temperature = 0) to code the remaining papers. The LLM was instructed to extract supporting textual quotes and providing justifications before synthesizing final codes. This two-step protocol was designed to reduce hallucinations and enhance traceability [102]. All LLM-generated outputs were manually reviewed and corrected by human coders by reading justifications and quotes.

To evaluate the reliability of our pipeline, we conducted an inter-rater reliability (IRR) analysis on a random sample of 30 papers per coding category. LLM-assigned codes were compared against independent human annotations. Results showed substantial agreement (e.g. in *Ethical Risks*, the micro-averaged Cohen's kappa was 0.717), supporting the

validity and credibility of our approach. Additional details on the full pipeline, including prompting strategies, model configuration, and full per-category IRR results, are provided in Appendix.

Our systematic review and analysis of 176 selected papers reveal the current landscape of socially intelligent LLM-powered multimodal systems. The findings are structured around our four research questions, covering the scope of applications (RQ1), the technical operationalization (RQ2), evaluation practices (RQ3), and the primary challenges and risks (RQ4).

4 RQ1: Scope and Nature of Social Intelligence in LLM-powered Multimodal Systems

What social intelligence of LLM-powered multimodal systems is applied to human social behavior analysis across different contexts?

Our coding scheme for RQ1 was developed through a top-down process, beginning with theory-driven constructs derived from existing literature [100], and iteratively refined through collaborative discussions among the research team. All coding categories for RQ1 are summarized in Figure 2.

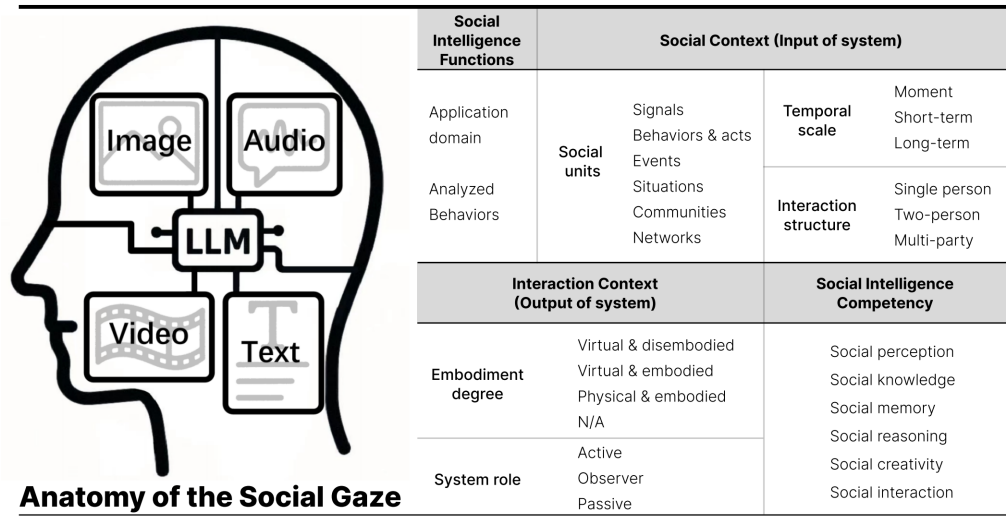


Fig. 2. A Conceptual Framework for Analyzing Social Intelligence (RQ1). This diagram outlines the multi-dimensional coding framework used to analyze how social intelligence is applied in LLM-powered systems. It deconstructs the analysis into four key areas: the system’s core Social Intelligence Functions, the Social Context it takes as input, the Interaction Context of its output, and the specific Social Intelligence Competencies it demonstrates.

4.1 Social Intelligence Functions

To understand how LLM-powered multimodal systems operationalize social intelligence, we analyzed both the application domains these systems target, and the types of human behaviors they are designed to recognize and interpret.

In analyzing the application contexts of our reviewed systems, we found that existing research divides into two distinct but complementary threads: application-oriented research, which targets specific real-world application contexts, and research-oriented work, which focuses on advancing foundational knowledge or core technical capabilities without a direct application in mind.

Among the application-oriented papers, the most prominent domain is Healthcare and Well-being, accounting for nearly a quarter of the corpus (23.73%). These studies primarily target clinical or therapeutic settings, such as creating AI systems that reduce anxiety for autistic adolescents [162]. This is followed by applications in key societal sectors like Media, Entertainment, & Content Creation (12.43%) and Education & Training (11.86%), which aim to enhance human creativity and learning. For instance, by generating personalized video comments [86] or by analyzing classroom dialogues to improve teaching quality [68]. More specialized domains like Security & Surveillance (5.65%) and Autonomous Driving (4.52%) also leverage these technologies for critical tasks, including detecting abnormal events in security footage [116] and predicting pedestrian intentions [110].

In contrast to these domain-specific efforts, a substantial portion of the literature pursues epistemic or infrastructural goals. Notably, Human-Robot/Agent Interaction appears as the single largest category (27.68%). This line of work investigates the broader principles of human-machine interaction in everyday settings such as homes [14], public spaces [66], or service environments [162]. These studies, rooted in human-computer interaction (HCI), emphasize the design and deployment of socially responsive agents and interactive systems, treating interaction itself as the primary object of inquiry. Similarly, 14.12% of papers focus on advancing core social intelligence capabilities rather than specific applications. These researches develop generalizable models for basic social perception tasks, such as zero-shot action recognition [26] and human-object interaction detection [46]. These foundational contributions provide the essential tools that researchers in the application-driven domains can then adapt and deploy.

We identified four major categories of behavior cues from our corpus (Figure. 3), reflecting the diverse ways systems interpret human activity. The most common are verbal and language cues (61.93%), which include analyzing written or spoken dialogue, generating captions, and responding to voice commands. This is followed by body and motion cues (48.29%), such as gesture recognition, action classification, and motion tracking. These two categories dominate the field, indicating a strong focus on linguistically and physically expressive forms of behavior. Other less commonly used behavior types include vocal and auditory cues (23.29%), which involve speech prosody, speaker identification, or emotion detection through audio, as well as facial and gaze cues (22.16%), including recognition of expressions or eye contact. Notably, these nuanced non-verbal cues, often essential in human social perception, remain relatively underutilized.

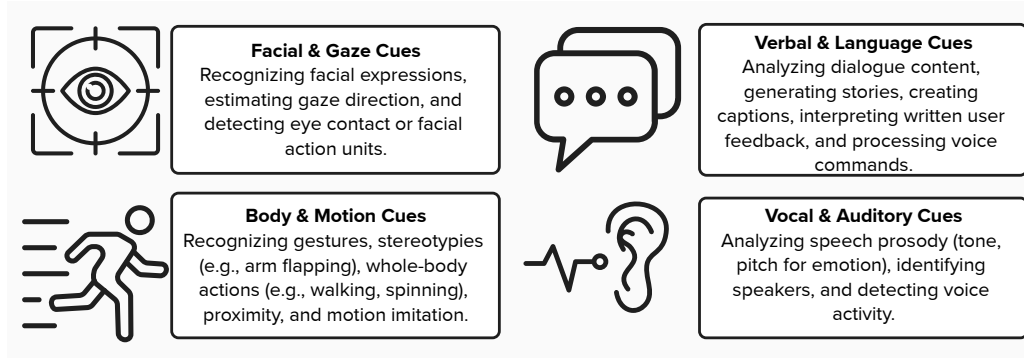


Fig. 3. Four Major Categories of Behavioral Cues Analyzed in Reviewed Systems. The figure displays the primary types of human behavioral cues that the surveyed systems are designed to interpret. These include verbal and language cues, body and motion cues, vocal and auditory cues, and facial and gaze cues.

Our analysis also reveals clear differences in cue usage across application contexts (Figure. 4). For example, healthcare and well-being systems tend to integrate a broad spectrum of cues—including facial expressions, prosody, and motion—to support emotionally responsive interactions, such as in cognitive therapy. In contrast, education and training applications emphasize verbal and language cues, aligning with their focus on dialogue analysis and instruction. Security and surveillance systems skew heavily toward motion and action recognition, while domains like autonomous driving prioritize both complex goal-directed actions and spatial body movement cues.

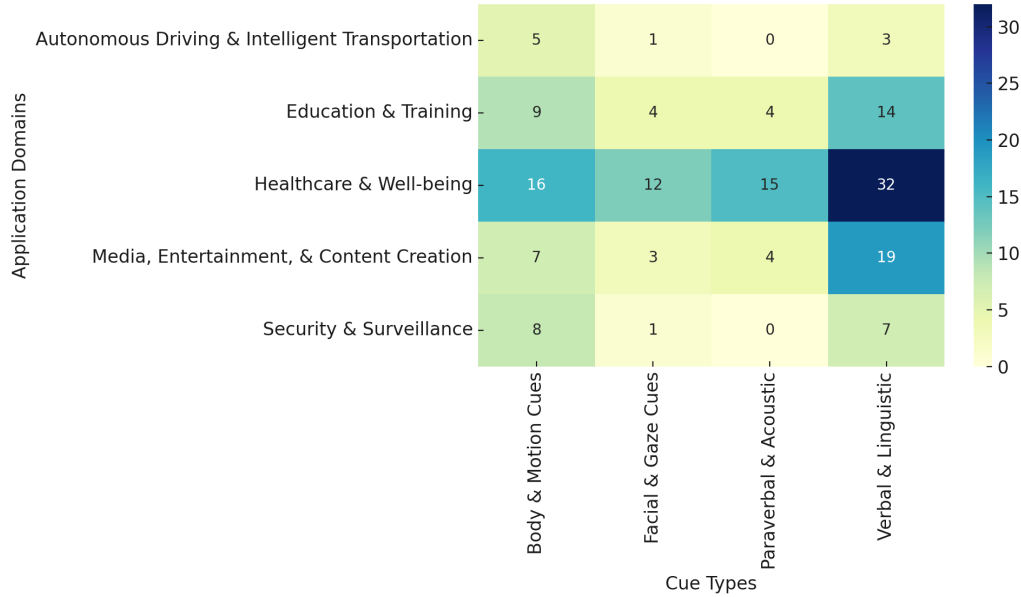


Fig. 4. Distribution of Behavioral Cue Types Across Application Domains. This heatmap illustrates the frequency of different behavioral cues analyzed within key application areas. It reveals distinct patterns of focus; for instance, "Healthcare & Well-being" utilizes a wide range of cues, whereas "Education & Training" predominantly focuses on verbal and linguistic cues.

4.2 Behavior Characteristics

To characterize the nature of social behaviors analyzed in LLM-powered multimodal systems, we examined three key dimensions: interaction structure (i.e., the number of interacting participants), temporal scale (i.e., the time span of behavior under analysis), and social units (i.e., the building blocks of social analysis). Together, these dimensions offer a structured view of the complexity of human behavior addressed by current systems (Figure. 5).

A clear pattern emerges across the corpus: most studies center on micro-level, short-term, and individually framed behavior. The temporal scope of analysis is heavily skewed toward the immediate (94.3%) and short-term (82.9%), with only a small minority (8.5%) addressing long-term behavior. That is, most systems are designed to interpret actions and interactions within a single session or isolated moment, rather than track evolving dynamics over days or weeks.

Examination of social units further confirms this micro-level orientation: the dominant social units analyzed are behaviors and acts (98.3%) and signals (95.4%)—that is, discrete gestures, expressions, utterances, or cues such as gaze and tone. In contrast, more complex structures such as communities (43.2%) or networks (2.8%) are comparatively rare. Even in cases where papers touch on group-level constructs (e.g., "classroom community" or "team collaboration"),

the analytical unit is often reduced to individual actions or attributes within that context. For instance, studies like [2] dissect speaker behaviors in group conversations based on fine-grained individual cues rather than modeling the collective dynamic. [35] and [79] offer notable accounts of network structures: the former analyzes collaboration in medical teams through sociograms and communication networks, while the latter constructs character networks in films through knowledge graphs.

This tendency toward “parallel individual analysis in group contexts” is further reflected in the interaction structures observed. Single-person analysis dominates (63.6%), even when conducted within multi-party environments. Studies rarely analyze contingent interactions between multiple participants (32.9% for two-person, 26.7% for multi-party), and even fewer model relational dynamics such as turn-taking or coordination patterns.

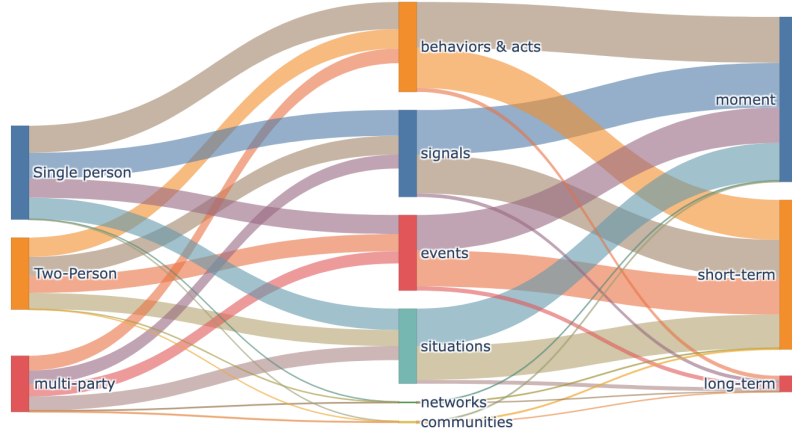


Fig. 5. Analysis of Social Behavior Along Three Key Dimensions. This Sankey diagram visualizes the relationships between the interaction structure, social units, and temporal scale in the reviewed literature. The flow highlights a dominant research focus on analyzing individual behaviors (“single person”) at a micro-level (“signals,” “behaviors & acts”) over brief timescales (“moment,” “short-term”).

4.3 System Characteristics

To complement our analysis of behavior characteristics, we next examine the context in which systems are deployed and interact, focusing on two key dimensions: embodiment degree and system role. Together with the behavior layer, these form what we refer to as a two-layered social context for socially intelligent systems. While behavior characteristics capture the nature of social inputs that systems interpret, system characteristics reflect how the system itself is situated and engaged within interactional settings.

The majority of reviewed papers (61.93%) fall into the N/A category for embodiment degree, indicating a strong emphasis on non-agentic systems, that is, systems which analyze or generate behavior but do not take on an interactive social role. Among embodied systems, physically embodied agents (21.02%) are more prevalent than virtual embodiments (8.52% each for both embodied and disembodied forms), suggesting a notable interest in deploying social intelligence in physical, real-world environments (e.g., service robots, therapeutic companions [84, 162]), rather than purely digital interfaces.

System role refers to the system’s functional position in the interaction, such as acting as a passive responder, an active participant, or an external observer. We observe that most systems (57.39%) function as observers, reflecting the dominance of a post hoc analytical paradigm in which LLMs are used to interpret recorded social behavior. A smaller number of systems (22.73%) operate as passive responders, such as in Q&A or dialogue completion tasks. Notably, only 35 papers (19.89%) describe systems that act as active participants, meaning systems capable of steering interaction, providing instructions, or setting goals (e.g. [65, 135, 139]).

Taken together, these findings reinforce a key asymmetry in current research: while many systems are designed to analyze complex social behaviors, relatively few are positioned as socially capable agents within those interactions. Bridging this gap—between observing social complexity and participating in it—remains a central challenge for the development of socially intelligent multimodal systems.

4.4 Operationalized Social Intelligence Competencies

Prior research in psychology has conceptualized social intelligence as comprising both cognitive and behavioral facets: the former involves interpreting others’ mental states and social norms, while the latter refers to acting appropriately in social situations [155]. However, operationalizing these facets for AI systems remains an open challenge. Drawing on performance-based models in psychology [155] and recent frameworks proposed for AI agents [100], we adapted a six-part taxonomy of competencies as the foundation for our coding:

- **Social perception.** The system’s capacity to perceive and discriminate socially relevant signals from multimodal inputs (e.g., emotion recognition, gaze tracking).
- **Social knowledge.** The ability to apply contextual or normative information to interpret social situations (e.g., social norms, roles, expectations).
- **Social memory.** The ability to retain and retrieve socially relevant information across time (e.g., remembering faces or past interactions).
- **Social reasoning.** The capacity to infer hidden states, intentions, or causal relationships from observed cues (e.g., intent or moral reasoning).
- **Social creativity.** The ability to produce contextually novel and socially effective outputs rather than replicating patterns..
- **Social interaction.** The ability to engage in contingent, co-regulated exchanges with humans or agents in real time, adapting dynamically to the interactional flow.

To examine how LLM-powered multimodal systems instantiate these competencies, we coded each system for the presence of the six capabilities. Our results reveal clear trends: social perception (100%), social knowledge (97.74%), and social Reasoning (95.48%) are by far the most commonly operationalized competencies. This aligns with our earlier findings that most systems emphasize fine-grained recognition of expressive behavior (e.g., facial expressions, gestures, verbal cues), interpretive functions rooted in norms and context (e.g., classroom roles, clinical routines), and inference of hidden states (e.g., emotion, intent, collaboration dynamics). This foundational stack of Perception, Knowledge, and Reasoning forms the cognitive engine of nearly every system we reviewed.

Complementing this cognitive foundation is social memory (69.49%), which supports more context-sensitive analysis by allowing systems to store and recall prior information, such as user preferences or conversational history. Far less common, however, are systems that implement social interaction (41.81%) or social creativity (18.08%). This gap underscores a broader asymmetry across our corpus: while systems are skilled at analyzing and interpreting human

social behavior, far fewer are designed to participate in or co-construct social exchanges in real time. This is consistent with our earlier observation that most systems function as observers rather than actors, and that behavior modeling remains primarily individual and immediate rather than collective or longitudinal.

This asymmetry is also domain-dependent. As shown in Figure 6, domains such as Healthcare and Well-being and Media & Entertainment more frequently engage creative or interactional competencies—e.g., by generating empathetic responses or dynamic social content—while Security & Surveillance systems are heavily concentrated around perception and reasoning.

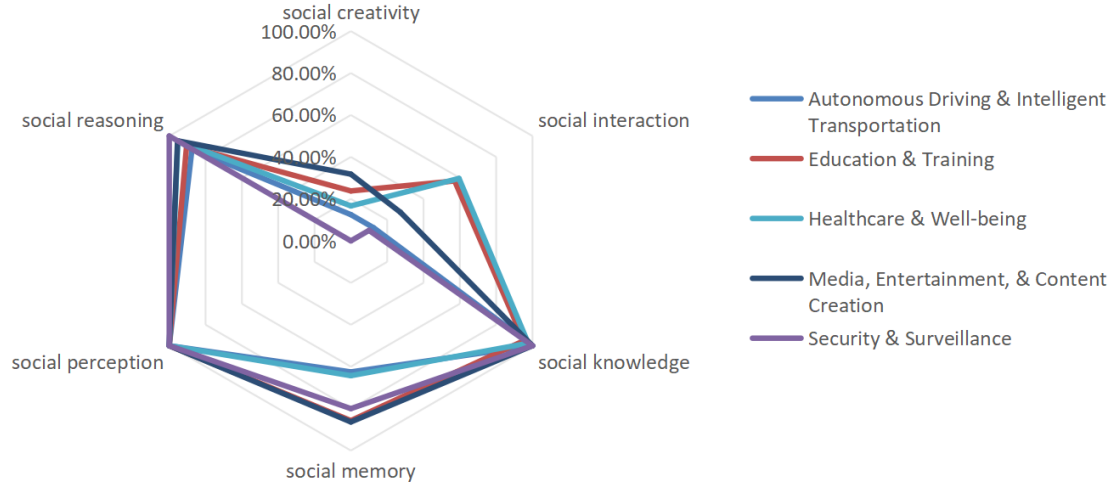


Fig. 6. Distribution of Social Intelligence Competencies Across Application Domains. This radar chart compares the implementation of six key social intelligence competencies across five major application domains. The chart shows that competencies like social perception and reasoning are nearly universal, while social creativity and social interaction are far less common, particularly in domains like Security & Surveillance.

Major Takeaways—Scope and Nature of Social Intelligence (RQ1)

- **Dominance of Perception & Recognition:** Nearly all of the 176 works focus on social perception (100%) and reasoning (95%), while social interaction (41.8%) and creativity (18%) are rarely implemented.
- **Short-Term, Individual-Centric Focus:** 94% of studies analyze immediate or short-term behavior and 63% focus on single-person analysis, with very limited modeling of long-term trajectories or multi-party dynamics.
- **Uneven Cue Usage:** Verbal/language cues dominate, while nuanced non-verbal signals like gaze, facial expressions, and prosody are underutilized despite their importance for social intelligence.
- **Observer Role Prevails:** Most systems act as post-hoc observers rather than active participants, highlighting a gap between analyzing and co-constructing social interaction.

5 RQ2: Technical Operationalization of Social Intelligence in Multimodal LLM Systems

How is LLM-powered multimodal systems social intelligence technically operationalized?

To address our second research question, we examined how social intelligence is technically implemented across LLM-powered multimodal systems. Our analysis focused on three key dimensions: (1) how input modalities are handled and transformed, (2) what modalities are processed by the core LLM, and (3) what operationalization strategies are employed—i.e., whether through architectural changes, prompting, or model fine-tuning.

5.1 Modality Reflection

Across the reviewed papers, a staged and text-centric processing architecture emerges as the dominant design pattern. Systems typically begin with rich multimodal inputs—video, audio, and text—capturing the complexity of human social behavior. As shown in Figure 7 (left), video (109 papers), audio (89 papers), and text (96 papers) are the most common input sources. The most frequent combination is video + audio (48 papers), reflecting a growing interest in audiovisual social interaction, such as analyzing online conversations [148].

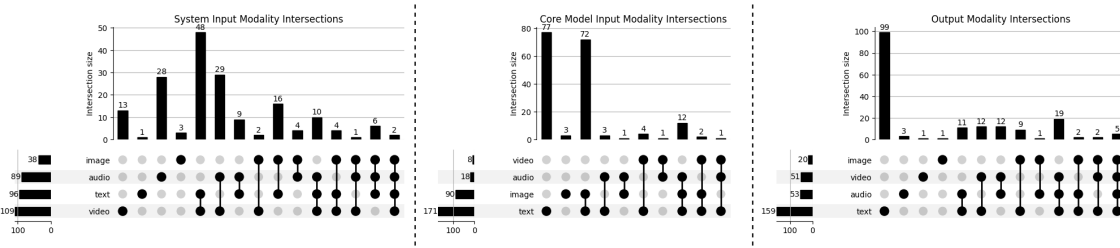


Fig. 7. The Modality-to-Text Bottleneck: A Visual Journey from System Input to Final Output. This series of charts visualizes the transformation of data modalities as they are processed by the reviewed systems, revealing a distinct "modality-to-text" pipeline. **System Input Modality Intersections (Left):** This chart shows that systems are designed to capture rich, multimodal data. The most frequent combination involves video, audio, and text together (48 papers), indicating an ambition to analyze complex, real-world social interactions. **Core Model Input Modality Intersections (Center):** This chart illustrates a critical shift in the pipeline. Rich inputs like video and audio are rarely processed directly by the core LLM. Instead, they are overwhelmingly converted into text (used alone in 77 papers) or a combination of text and static images (72 papers), simplifying dynamic signals for language-centric reasoning. **Output Modality Intersections (Right):** This chart shows the final output of the systems. The process culminates in a predominantly unimodal, text-only result (99 papers), reinforcing the role of these systems as analytical engines that distill a complex multimodal reality into a linguistic summary.

However, a significant discrepancy appears when comparing system input modalities with the actual inputs passed into the core LLM (Figure 7 (center)). While system-level pipelines often begin with raw video or audio data, these signals are rarely processed directly by the LLM. Instead, they are transformed into text (e.g., transcripts, captions) or static images through upstream modules like speech recognition, visual labeling, or temporal sampling. Text is overwhelmingly the dominant core input modality, appearing in 171 out of 176 papers.

These conversions highlight a design bias toward simplifying temporally rich and socially expressive modalities into more LLM-compatible formats. While this alignment enables efficient reasoning through LLMs, it may also introduce limitations—potentially compressing nuanced social signals such as gaze dynamics, conversational timing, or vocal affect into overly coarse representations.

Finally, this technical pipeline culminates in a predominantly uni-modal, text-based output. As shown in Figure 7 (right), text is the primary form of system output, with text-only output appearing in 99 papers, far exceeding any other modality or combination. This reinforces our earlier finding of systems as "analytical engines": the complex processing

of a rich, multimodal social reality is ultimately distilled into a linguistic output, such as generating stories for cognitive therapy [14] or providing writing assistance to users [17].

5.2 Operationalization Technique

To further analyze how social intelligence is implemented, we categorized each system based on three technical approaches:

- **Architecture-centric:** Changes to system-level structures or perception pipelines.
- **Inference-centric:** Use of prompting or other runtime configurations.
- **Model-centric:** Fine-tuning or modifying the LLM weights.

Architecture-centric techniques dominated the field, appearing in 164 papers. These typically involved custom pipelines for multimodal perception or modules to scaffold interaction. Inference-centric approaches, used in 137 papers, relied on prompt engineering or conversational role-playing to adapt behavior at runtime. Model-centric techniques, such as fine-tuning or adapter training, were the least common (82 papers), likely due to higher technical complexity.

Hybrid strategies were prevalent. Over 40% of the papers used both architecture- and inference-level strategies (e.g. [16]), and 52 papers employed all three methods simultaneously (e.g. [19]). These full-stack approaches often paired prompting with tailored pipelines and lightweight model adaptation.

Across all strategies, system design consistently centered around a narrow set of large language models. Over half of the papers used GPT-family models, including GPT-3.5, GPT-4, GPT-4V, and GPT-4o, as well as related variants. While other LLMs such as Claude, LLaVA, Vicuna, Mistral, and Gemini were present, their use was comparatively sparse. Notably, few papers used Gemini models despite their multimodal capabilities [97], and some papers did not report their base LLM at all [27]. This reliance on GPT-based architectures suggests a lack of diversity in model experimentation.

Major Takeaways—Technical Operationalization (RQ2)

- **Modality-to-Text Bottleneck:** Multimodal inputs (video/audio) are typically converted to text transcripts or keyframes before reaching the LLM, stripping away timing, prosody, and subtle context cues.
- **Text-Dominant Outputs:** Over half of the systems output text-only results, reinforcing their role as analytical engines rather than multimodal interactive agents.
- **Architecture-Centric Designs Dominate:** 93% rely on custom pipelines for preprocessing; model fine-tuning and inference-level adaptation are less common, treating LLMs as mostly fixed reasoning cores.
- **Model Homogeneity:** GPT-family models are used in the majority of studies; few explore alternatives like Gemini, LLaVA, or open-source models, limiting insights on model-specific performance trade-offs.

6 RQ3: Evaluation of Social Intelligence in Multimodal LLM Systems

How is the performance of socially intelligent LLM-powered multimodal systems evaluated?

6.1 Evaluating Social Intelligence: a divided field

To address our third research question, we analyzed the methods and metrics used in the literature to evaluate socially intelligent MLLMs. The analysis reveals a significant divergence in evaluation practices, characterized by a dominant, machine-centric paradigm focused on technical correctness and a less common yet more diverse, human-centric paradigm focused on human impact.

Out of 176 papers, 10 (5.6%) did not conduct any form of evaluation. Among those that did, 89 (50.5%) conducted only system evaluations using custom or established benchmarks and common ML metrics (e.g., accuracy); 4 (2.2%) conducted only human evaluations (e.g., interviews/surveys) including qualitative feedback metrics such as enjoyment, comprehension; and 73 (41.4%) conducted both system and human evaluations.

A common pattern is to report *Classification/Recognition Accuracy* as the premier metric for system evaluation, i.e., success measured by correspondence to ground truth (e.g., mAP in [95]). Authors also report *Computational Efficiency* (e.g., FPS in [16]) and *Language Generation Quality* (e.g., BLEU in [174]) to capture purely technical performance. This aligns with our earlier finding that many systems are designed as “analytical engines”.

Human evaluations adopt a more holistic approach. The most common methods are *Surveys* (34.6%) and *Interviews* (19.3%). Reported human-centered metrics span three aspects:

- (1) Usability & User Experience: assessing if the system is “good to use” through measures like the System Usability Scale (SUS) [84];
- (2) Task Performance & Behavioral Change: measuring if the system is “effective” through metrics like learning gain (pre-post test scores [91]);
- (3) Perception of AI Qualities: evaluating if the system is “likeable” or “trustworthy” by assessing factors such as perceived robot empathy (RoPE scale) [162].

This indicates that when researchers adopt a human-centric perspective, they emphasize a technology’s value and impact in authentic human contexts.

6.2 System Evaluation Mechanism

To bridge the gap between HCI goals and AI tasks in evaluating socially intelligent LLM systems, we further investigated how these systems were evaluated in terms of *System Evaluation* and compiled Table 1. The table provides a comprehensive list of benchmarks and the associated papers in which they were used to evaluate socially intelligent multimodal LLM systems. It is organized by Application Area (Human Goal) and Computational Task (Model Framing) to guide future researchers in selecting appropriate benchmarks and prior evaluation techniques for their analysis goals.

For papers with system evaluation, 101 papers used benchmarks (55 established, 17 mixed, 29 new). These benchmarks cover four **Application Areas** related to human goals when analyzing social multimodal data: (1) *Action & Activity Understanding*, (2) *Social & Affective Intelligence*, (3) *Video & Language Comprehension*, and (4) *Domain-Specific Applications*. These goals map to common **Computational Tasks** in machine learning (e.g., 1.1 Classification & Localization, 3.1 VideoQA & Reasoning) and to more nuanced domain specific tasks (e.g., 2.1 Social Reasoning & Theory of Mind (ToM), 4.1 Healthcare & Accessibility) that are benchmark-evaluable.

Tasks are then mapped to **Evaluation Resources**—benchmarks/datasets the authors either (a) used as established (e.g., RefCOCO [46], GazeFollow [52]), (b) derived/built upon (e.g. Playlogue[70], OSCaR[112]), or (c) created as custom resources for nuanced human behavior analysis (e.g., SMILE [63], Autism Restricted and Repetitive Behavior Dataset (ARRBD)[151]).

Major Takeaways–Evaluation Practices (RQ3)

- **Machine-Centric Benchmarking Dominates:** Across 176 papers, 10 (5.6%) had no evaluation; of the rest, 89 (50.5%) were system-only, 4 (2.2%) human-only, and 73 (41.4%) both; human methods centered on Surveys (34.6%) and Interviews (19.3%).
- **Benchmark Uses and Updates:** Researchers both rely on traditional benchmarks and increasingly adapt or introduce new ones to evaluate nuanced multimodal social tasks: among 162 papers with system evaluation, 101 (62.3%) used benchmarks—55 (54.5%) established, 17 (16.8%) adapted/derived, and 29 (28.7%) newly created.
- **Dataset Directory–Mapping Goals→Tasks→Resources:** (see Table 1). The table compiles benchmarks and their associated papers, organized by Application Area (Human Goals) and Computational Tasks (Model Framing), providing researchers a practical guide to select appropriate benchmarks and system evaluation techniques for their specific analysis goals.

Table 1. Bridging HCI goals and AI tasks in evaluating socially intelligent LLM systems. Entries are organized by Application Area (Human Goal), Computational Task (Model Framing), and Evaluation Resource (Benchmark / Dataset). Dataset names include citations from papers they’ve been used in.

Application Area (Human Goal)	Computational Task (Model Framing)	Evaluation Resource (Benchmark / Dataset)
Action & Activity Understanding	Anomaly & Safety Monitoring	AN-Workout[116]; UCF-Crime[116]; XD-Violence[116].
	Anticipation & Forecasting	Gaze: AVA[52]; ChildPlay[52]; EMS[95]; ENCAA[95]; GazeFollow[52]; MIP-GAF[[95]; MS[95]; NCAA[95].
	Classification & Localization	A/V: AVS benchmark[46]; COCO-20i[46]; PASCAL-5i[46]; RefCOCO(+/g)[46] Ego: CharadesEgo[26]; Charades[85]; EGTEA[26]; EPIC-KITCHENS-100[26, 114]; OSCaR(derived from EPIC-KITCHENS)[112] Ego-Exo4D[129]; Ego4D [57, 90, 105, 112, 114, 124, 150]; EgoSchema[105] HMDB51[23, 85, 154]; K600[85]; Kinetics[23, 154]; MiT[85]; MiniSSv2[85]; NTU RGB+D(60, 120)[23, 24]; PKU-MMD[24]; UAG-(FunQA, SSBD, OOPS[1]) ; UAV[85]; UCF101[23, 85, 154]; “Chaotic World” benchmark[22]; Charades-STA[1]; Subsets of Charades(ToM)[125].
	Interaction, Pose & Motion	3DPW[38]; AGD20K[149]; HICO-(DET[46, 122],IIF[149]); Human3.6M[38]; HumanML3D(ToM)[173]; PoseFix[38]; PosePart[38]; PoseScript[38]; Reasoning-based Pose Estimation(RPE)[42]; SWIG-HOI[46]; Speculative Pose Generation(SPG)[42]; V-COCO[122];

Table 1 – continued from previous page

Application Area (Human Goal)	Computational Task (Model Framing)	Evaluation Resource (Benchmark / Dataset)
	Procedures & State Change	50 Salads dataset[147]; Breakfast dataset[147]; COIN[115]; CrossTask[115]; Annotated version of the GTEA dataset(gaze, ToM)[101]; KIT Bimanual Action[174]; MS-COCO[174]; Multiple Object States Transition (MOST) dataset[140]; VirtualHome simulator[151]; Youcook2[174]; Annotated dataset derived from YouCook2(ToM)[59];
	Retrieval & Long-Form Understanding	ACM MM 2023 Grand Challenge[79]; NIST TRECVID 2022 DVU[79]; RSL dataset[31]; UVSD dataset[31]; Waldo[7] (retrieval); Wenda[7] (retrieval); subset of imSitu[7] (retrieval)
Social & Affective Intelligence	Dialogue & Interaction	Playlogue[70]
	Emotion & Affect	AffectNet[120]; CASME2[93]; CelebA-(base, dialog)[94]; DEEP corpus[109]; DISFA[93]; ECF2.0(sourced from 'Friends')[148]; EMOTIC[37, 141](Subset of EMOTIC(ToM)[166]); E ³ dataset[41] GoEmotions[120]; Hume-Vidmimic2[118]; IEMOCAP[19, 40]; MELD[40]; MSP-(Improv, Podcast)[40]; RAF-DB[93, 94]; Real-Life Trial dataset[93]; SCB-dataset3[141]; iMiGUE[93]
	Social Reasoning & Theory of Mind (ToM)	CH-SIMS[89]; CK+[89]; CMU-MOSI[89]; Friends[92]; Interpersonal Relation Dataset (IPR)[5]; JAFFE[89]; KDEP[89]; LMU Munich Executive Leadership Perception (LMU-ELP[98]; MHD[92]; MMLSCU dataset[103]; MMTOM-QA(ToM)[69]; MUSTARD[92]; Memento10k[99]; Modified version of Learning to Listen (L2L) dataset[111]; MuMA-ToM[133]; Multimodal Tamil Hate (MATH) dataset[106]; People in Social Context (PISC)[5]; SMILE[63]; Social-IQ 2.0[4]; Subset of 'Automobile Cabin Voice Interaction Data' dataset(ToM)[104]; TBBT[92]; UR-Funny[92]; USC's Split-Steal corpus (newly annotated)[55]; VCC2018(ToM)[12]; VVALUES[147]; VideoConviction[45]; VCC2018(ToM)[12]
Video & Language Comprehension	Captioning & Narration	Activity-Net[33]; DiDemo[33]; Flickr8K[142]; LSMDC[33, 56]; MAD-eval[56]; MSR-VTT[3, 33]; StreetAware[3]; Tokyo MODI[3]; VATEX[33]; VQA-v2[3]; Flickr8Knew(ToM)[142]

Table 1 – continued from previous page

Application Area (Human Goal)	Computational Task (Model Framing)	Evaluation Resource (Benchmark / Dataset)
	Retrieval & Grounding	ActivityNet-Captions[164]; Charades-STA[164]; QVHighlights[164]; QueryYD dataset[113]
	Semantic Graphs & Story	MovieGraphs+[32, 117]; PerVidCom[86]; ViSR+[117]
	VideoQA & Reasoning	AGQA 2.0[152]; ActivityNet-QA[30]; EgoSchema[152]; IntentQA dataset[77]; Kinetics[30]; MSRVTT-QA[30]; MSVD-QA[30]; NExT-(GQA, QA)[152]; QA-Ego4D[132]; Something-Something V2[30]; Video-MME[30]
Domain-Specific Applications	Driving & Scene Understanding	Driving Actions: BDD100K(ToM)[160]; CIFAR-10/100(ToM)[160]; DHPR (Driving Hazard Prediction)[21]; HighD[96]; JAAD[96, 110]; Derived from KITTI[67]; NuScenes[67]; PIE[110]; PSI[96]; PedPrompt[110]; Street Scene Understanding: StreetAware[9, 10]; Tokyo MODI[10]
	Education	Kyoto University’s lectures[65]; Coursera educational videos[68]; YouTube educational videos[68];
	Healthcare & Accessibility	ASD & Autism: Audio-Visual Autism Spectrum Dataset (AV-ASD)[29]; Autism Restricted[151]; Repetitive Behavior Dataset (ARRBD)[151]; Accessibility: Derived from VideoA11y-40K dataset[76] Sign Language Recognition: AUTSL[87]; Isolated Sign Language Recognition Corpus(ToM)[83]; MSASL[87]; WLASL100[87]; General: MEDIQA-QS[156]; MeQSum[156]; Reddit posts[156]
	Human-Robot Interaction	DataHRC[80]; Disfluent Navigational Instruction Audio Dataset(DNIA)[137]; HHIRChat[130]; JRDB-Social[66]; RoboCup@Home GPSR task benchmark[134]; RoboTHOR[137];

7 RQ4: Ethical Challenges and Risks of Socially Intelligent Multimodal LLM Systems

What are the primary ethical challenges and risks associated with socially intelligent LLM-powered multimodal systems?

7.1 Setup and Coding Approach

We distinguish two outcomes for each ethics issue reported in a paper: *Addressed Ethical Risks*—risks explicitly acknowledged *and* accompanied by a concrete control—and *Remaining Ethical Risks*—risks explicitly raised as unresolved, emerging, or out-of-scope. We coded all ethics-relevant statements with the AIR’24 framework [171], which contains 314 categories in a four-tier taxonomy; at Tier 1 the families are *System & Operational*, *Content Safety*, *Societal*, and *Legal & Rights*. In parallel, we coded papers for *Implemented Mitigations* (actions or design choices actually used in the reported system/study) versus *Proposed Mitigations* (unimplemented strategies suggested for future work or deployment).

7.2 Coverage of Ethical Considerations (Paper-Level)

Figure 8 summarizes paper-level coverage as a funnel: of 176 papers, 97 (55%) mention ethics in any form, while 79 (45%) do not. Among the ethics-aware set, 75/97 (77%) report at least one *implemented* mitigation, whereas 22/97 (23%) *only* mention risks without action. In absolute terms, this means 43% of all papers (75/176) include any concrete mitigation, and 12.5% (22/176) stop at mention-only.

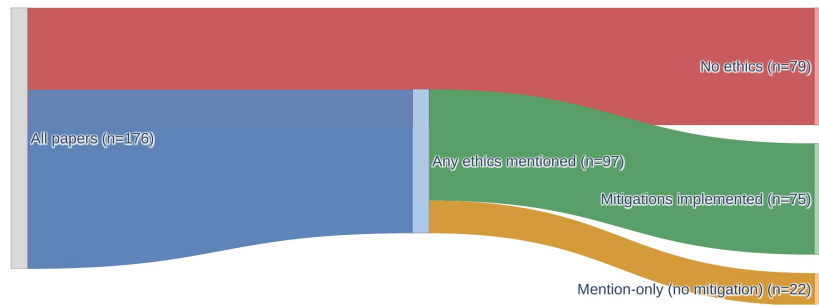


Fig. 8. Paper-Level Coverage of Ethical Considerations. This Sankey diagram shows the breakdown of ethical engagement across the 176 reviewed papers. It reveals that 45% of papers do not mention ethics. Of the 55% that do, 77% report implementing at least one mitigation, while 23% only mention risks without describing a concrete action.

This analysis reveals that ethical practice remains unevenly integrated into LLM-powered multimodal work. Even among ethics-aware papers, *about a quarter* stop at acknowledgement without intervention, and *nearly half* of the literature omits ethics entirely. In the following subsections, we examine *which* categories receive mitigation and *how* they are addressed, proceeding through the AIR’24 taxonomy: **Tier 1** (families), **Tier 2** (subfamilies), and **Tier 3** (subcategories). This structure surfaces where interventions are reported and where gaps remain without presupposing outcomes.

7.3 Risk Families: What Gets Addressed vs. What Remains Open

Figure 9 (Tier 1 and 2) and our full coding (Tier 3 below) point to a consistent pattern: Legal & Rights dominates what gets *addressed* (97 mentions), while Societal concerns are more visible among what *remains* (23 vs. 12). System & Operational is comparatively stable across buckets (20 vs. 19), and Content Safety stays smaller overall (13 vs. 7).

Within Legal & Rights, the emphasis flips: *addressed* items center on *Privacy* (76) with far fewer *Discrimination/Bias* mentions (14), whereas *remaining* items are led by *Discrimination/Bias* (48) over *Privacy* (28). System & Operational splits similarly in both buckets (*Security Risks* 10 and *Operational Misuses* 10 addressed; 10 and 9 remaining). For Societal,

Deception grows sharply among remaining (17 vs. 4 addressed), while *Manipulation* appears in both (5 addressed; 3 remaining). Content Safety is broader in the addressed set (*Hate/Toxicity* 9 plus *Violence & Extremism*, *Child Harm*, *Self-harm*), but thinner among remaining (mostly *Hate/Toxicity* 6 plus *Violence & Extremism* 1).

Privacy-focused subcategories dominate addressed mentions (*Unauthorized Privacy Violations* 46; *Types of Sensitive Data* 30) but are much smaller among remaining (19 and 9). By contrast, fairness-related items cluster in the remaining set (*Discriminatory Activities* 26; *Protected Characteristics* 22 vs. 8 and 6 addressed). Security is symmetric (*Confidentiality* 6; *Integrity* 4 in both). Operational Misuses diverge: *Autonomous Unsafe Operation* (6 addressed vs. 2 remaining) and *Advice in Regulated Domains* (4 addressed; 5 remaining), with *Automated Decision-Making* appearing only among remaining (2). Societal harms show a similar skew: *Mis/disinformation* is higher among remaining (13 vs. 4 addressed), while *Misrepresentation* appears in both (5 addressed; 3 remaining). Addressed Content Safety spans more severe categories (*Self-harm* 2; *Child Harm* 1) than the remaining set, which concentrates on toxicity variants (e.g., *Hate Speech*, *Offensive Language*, *Harassment*) at lower counts.

The field reports concrete controls where pathways are established—especially *privacy/security*—but fairness and socially grounded harms surface more strongly as *remaining* issues. Figure 9 (Tier 1 and 2) outlines this shift, and the Tier 3 details above show where, specifically, privacy is operationalized (e.g., unauthorized disclosure; sensitive data types) while fairness (*discriminatory activities*, *protected characteristics*) and deception (*mis/disinformation*) persist. Complete results table for these 3 tier coding is provided in Appendix.

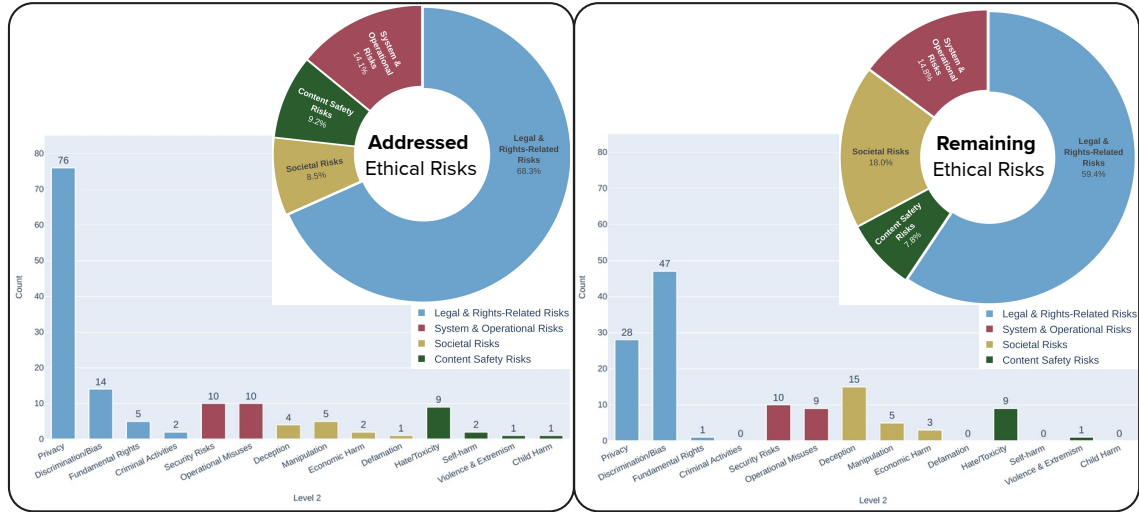


Fig. 9. Comparison of Addressed vs. Remaining Ethical Risk Mentions. This figure compares the types of ethical risks that are actively addressed with mitigations versus those that are mentioned as remaining or open challenges. The charts highlight a clear shift: addressed risks are dominated by concerns over Legal & Rights, particularly Privacy. In contrast, remaining risks show a significant increase in concerns related to Discrimination/Bias and broader Societal harms like deception.

7.4 Mitigations in Practice vs. On the Horizon

Implemented mitigations cluster around what papers can control immediately in deployment: (1) *privacy-by-design* (on-device/edge or local models; de-identification, masking/blurring, restricted releases; secure storage/configuration

and no-retention settings; IRB/consent, opt-out flows, pseudonyms, and controlled lab studies); (2) *human oversight and gating* (therapists/caregivers or instructors in the loop; researcher moderation; explicit audience transparency about virtual agents; post-hoc identity disclosure in studies; professor control over classroom robots); (3) *guardrails on generation and execution* (content filters and guideline-constrained prompting; verbosity caps; pre-generated outputs to avoid non-determinism; RAG/ReAct/template scaffolds; sandboxed interpreters, input validation, and “predefined functions only” for code execution); (4) *embodied/interactive safety engineering* (motion-planning constraints, simulation-before-execution, scene-graph anomaly detection, intent-aware planning; model-side explainability and recovery affordances such as inner speech, tiered summaries, or KG+fuzzy-rule rationales); (5) *data/process governance* (use of pre-approved or de-identified datasets; licensing/GDPR-aware collection; restricted sharing; cloud configurations disabling retention/training); and (6) *user controls and protective UX* (moment-selection to filter captured content; color-blind modes; supportive default content to displace toxic inputs; expert-editable behavior codebooks).

Proposed (unimplemented) directions extend beyond immediate controls toward harder open problems: (1) *fairness and generalization* (dataset diversification, bias audits, culturally sensitive fine-tuning; replication across tasks/populations; annotator diversity, informed codebooks, stereotype-aware modeling); (2) *technical privacy and data control* (differential privacy, homomorphic encryption, modular disabling of PII processing, generative PII replacement; silent-speech input; bystander indicators/recording lights and local face-blurring; machine unlearning; on-device-by-default for future deployments); (3) *transparency, verification, and evaluation* (interpretable outputs with citations; trust cues and confidence scores; standardized evaluation APIs; human verification/deferral for sensitive judgments; user-driven fact checking; deployment guidelines, audits, and accountability channels; offline fallbacks); (4) *robust autonomy under uncertainty* (uncertainty handling/denoising; voice-activity detection for turn-taking; temporal reasoning; multi-sensor/video expansion; improved trajectory planning, feedback-based replanning, memory and intention modeling); (5) *person-centered UX* (customization/personalization of agents and haptics/visual parameters; tapered intervention to reduce over-reliance; accessibility features; designs that amplify existing human bonds); and (6) *computational pragmatics and safety-of-content* (model distillation and energy-aware operation; stronger filters with human-in-the-loop moderation; clearer annotation/wording to avoid refusals). In short, what is *implemented* centers on privacy, consent, and controllability in today’s pipelines; what is *proposed* targets fairness, reliable reasoning under uncertainty, and accountable, socially situated behavior.

Major Takeaways—Ethical Challenges and Risks (RQ4)

- **Ethics is Unevenly Addressed:** 45% of reviewed papers do not mention ethics, and 23% of those that do stop at acknowledgment without implementing mitigation. This gap is visible in the funnel (Figure 8).
- **What gets fixed vs. what lingers.** Reported mitigations cluster in *privacy/security* (e.g., data handling, consent), while *fairness* and broader *social harms* (mis/disinformation, manipulation) surface more often as remaining issues. (Figure 9).
- **Near-term controls vs. hard problems.** Implemented actions emphasize controllability—on-device processing, de-identification, human oversight, sandboxed execution, and prompt/RAG scaffolds—whereas proposed strategies target the harder frontiers: fairness/generalization, transparent verification, and uncertainty-aware autonomy in socially situated settings.

8 Discussion

In this section, we explore the research gaps identified across four key perspectives: the application of LLM-based social intelligence, technological operations, system evaluation, and ethical considerations. Building on these observations, we propose a forward-looking research agenda for each area individually, as well as for the broader field as a whole.

8.1 Beyond Cognitive Recognition: Toward Behavioral and Interactive Social Intelligence (RQ1)

Our coding reveals a consistent imbalance across competencies: almost all papers target social perception, social knowledge, and social reasoning, with many also using social memory, while social interaction and social creativity remain far less common. Systems are built to detect, recall, and reason about social cues, yet far fewer attempt to act on them in contextually appropriate, adaptive ways. In short, the systems we reviewed largely optimize the cognitive facet of social intelligence but underserve the behavioral facet, precisely the duality emphasized in psychology and in recent AI-oriented syntheses [49, 100].

Several patterns in system design help explain this imbalance. Most systems adopt an observer role, analyzing social signals post-hoc rather than participating actively in interaction. They also tend to focus on immediate, short-term behavior (94% of studies) and largely consider individuals in isolation, leaving multi-party and longitudinal dynamics underexplored (63%). Furthermore, verbal and language cues dominate system inputs, while nuanced non-verbal signals such as gaze, facial expressions, and prosody are rarely utilized, limiting the richness and adaptability of social behaviors these systems can generate.

8.1.1 Social Knowledge remains generic, not norm-sensitive. In psychology, social knowledge is understood as a complex competency encompassing not only declarative facts about the social world but also procedural knowledge of social norms and etiquette that guide behavior in specific contexts [25, 155]. Crucially, this knowledge is not universal; it is highly dependent on cultural values and situational specifics.

Our review, however, finds that in the LLM-based systems we surveyed, this competency is frequently operationalized as generic, text-based commonsense rather than as situated, norm-sensitive knowledge. For example, IntentQA [77] leverages an LLM to provide commonsense context for disambiguating intent, such as knowing that "putting a spoon in one's mouth" likely means "eat food." While useful for reasoning about individual actions, this approach overlooks the procedural knowledge needed to navigate norms in larger communities [100].

A notable exception is ClassMeta [91], where a virtual agent embodies the social knowledge specific to a classroom "community." The agent's behaviors, such as knowing when to break the silence or how to correct off-topic chatter, are forms of procedural knowledge specific to that learning environment. The scarcity of such systems highlights a significant gap, and the emergence of benchmarks like EgoNormia [123], designed to evaluate understanding of physical social norms, signals a growing recognition that AI must move beyond generic commonsense toward more nuanced, norm-sensitive social knowledge.

8.1.2 Cognitive Modeling Choices Shape the Behavioral and Interactive Gap. The way social knowledge and creativity are modeled in current systems amplifies the behavioral gap [100]. Many systems are designed to converge on a single "correct" output rather than generate multiple possibilities or flexibly adapt [107]. This convergent problem-solving approach is misaligned with the divergent nature of social creativity and interaction, which often require exploring multiple behavioral options and adapting to dynamic social contexts [121]. Combined with still-developing

capabilities, systems seemed to follow a reactive pattern: they map recognized inputs to predefined responses rather than co-constructing interactions in real time.

The short-term, individual focus of most research further limits opportunities to model multi-party coordination or longitudinal interaction trajectories. Meanwhile, the heavy reliance on verbal cues, with non-verbal signals largely underutilized, restricts sensitivity to the subtle information that humans use to coordinate social behavior. Together, these factors foster a reactive, observer-style design: systems can respond to detected inputs but are not structured to engage adaptively or collaboratively.

Advancing the interactive dimension of social intelligence therefore requires a deliberate shift. Future systems must move beyond perception and reasoning to support agents that can act creatively, adaptively, and socially in complex, real-world contexts. Only by integrating norm-sensitive social knowledge, multi-party and longitudinal modeling, and richer multimodal cue processing can AI systems begin to bridge the cognitive-behavioral gap and participate meaningfully in social interactions.

8.2 Modality Reflection: Building Social AI Beyond Textual Bottlenecks (RQ2)

Across the corpus, “multimodal” pipelines frequently compress rich signals into simpler, language-ready surrogates: video becomes frames or captions; audio becomes transcripts. This choice appears to be a pragmatic adaptation to the task at hand, for many of the specific goals outlined in these papers, a complete, native understanding of the original multimodal signal is not required. The simplified data is often sufficient for their descriptive task. However, this compression inevitably leads to a loss of nuance, creating a critical gap between what the system can process and how humans communicate.

This methodological constraint leads to systemic limitations. For example, preprocessing pipelines that convert video to keyframes and transcripts for content coding are effective at identifying concrete elements but demonstrate reduced accuracy for abstract concepts like emotion or communication style, as they exclude the very modalities through which that content is conveyed [88]. Similarly, rendering personal media searchable by converting inputs into text-based annotations and embeddings inherently discards non-textual information. This conversion process results in the loss of small-scale visual details (e.g., brand logos) and eliminates the contextual associations between discrete events (e.g., notebook imagery and meeting contexts) [78].

These examples illustrate a fundamental architectural constraint: current multimodal systems prioritize computational tractability over communicative fidelity. The resulting frameworks capture explicit content (“what was said” or “what appears in a frame”) while operating independently of the social, temporal, and contextual frameworks through which human communication derives meaning.

We also observe a pattern of visual and textual dominance and an under-use of the audio channel in our reviewed papers. While vision encoders are widely used, audio is often reduced to its transcript or, in more sophisticated cases, represented by simplified surrogates. This approach is exemplified by Hyun et al. [63], who converted prosody (e.g., pitch, intensity) into numerical features within a text prompt to help LLMs reason about laughter. Conversely, Ng et al. [111] took this to its logical extreme by excluding audio entirely, successfully generating listener reactions from transcripts alone. This text-only success suggests why audio may be de-prioritized: transcripts appear “good enough”—preserving semantic and temporal cues while avoiding the complexity of audio processing.

Yet, the limitations of this text-only approach are defined by the authors themselves. Ng et al. [111] report that the model is limited in cases where socially diagnostic cues are audible but not in the text, such as sarcastic jokes where vocal tone contradicts the literal words. This underscores the importance of the audio channel, a point powerfully

reinforced by the findings of Hyun et al. [63] The fact that their simplified numerical surrogates for prosody significantly improved performance over text-alone is compelling evidence of the information density within the original audio signal: if even the "ghost" of the signal is this impactful, the signal itself must be vital.

These design choices reflect a common architectural constrain: multi-modal inputs are converted to textual representations to enable LLM processing. This pattern suggests an underlying assumption that symbolic linguistic representations can adequately capture multimodal social communication. However, recent empirical work challenges this "logocentric" assumption [157]. Xu et al. [163] document only a "low correlation" between linguistic and social intelligence in LLMs, observing "superficially friendly" styles without contextual grounding. This aligns with fundamental questions in social AI about whether language can serve as a sufficient intermediate representation for social signals that may not be "effectively described in language" [100]. This critique parallels broader calls to "align perception with language models" [62], reflected in their assertion that "Language Is Not All You Need."

8.3 Optimized for Benchmarks, Misaligned with Human Context: A Call for Ethically-Informed Social Evaluation (RQ3-RQ4)

Our review reveals a critical disconnect between how socially intelligent multimodal systems are developed and the real-world contexts they are meant to inhabit. The predominant evaluation paradigm skews heavily toward machine-centric benchmarking over human-centered assessment. Our analysis of 176 papers shows that 50.5% report system-only evaluations, while 41.4% evaluate on both, and a mere 2.2% rely on human evaluation alone (see §6.1). This focus risks optimizing for performance on narrow, quantitative benchmarks that often fail to predict success in the messy, dynamic reality of human social interaction. Such systems may learn to exploit dataset-specific heuristics—a phenomenon known as "shortcut learning"—rather than developing robust, generalizable social understanding [48, 53]. Even as benchmarks evolve from simple recognition to more complex reasoning tasks, they continue to privilege structured, static outcomes, overlooking the reciprocal, adaptive qualities that define genuine social intelligence [36, 64]. Consequently, benchmark progress becomes a poor proxy for meaningful social impact [36].

This evaluation gap is compounded by an ethics gap. Our findings show that ethical considerations are addressed unevenly: while a slight majority of papers (55%) mention ethics, fewer than half (43%) implement any concrete mitigation (see §7.2). Furthermore, existing mitigations tend to cluster around technically tractable issues like privacy and data security. Systemic risks related to fairness, discrimination, and broader societal harms—the very issues that manifest most acutely upon deployment—remain largely unaddressed, relegated to the category of "future work."

Taken together, these patterns create a significant deployment gap. Current research primarily reports on proxy indicators of success, such as benchmark scores and technical efficiency, rather than the human outcomes that matter in practice. As foundational HCI scholarship reminds us, true impact is measured by tracing the path from system inputs and outputs to real-world outcomes, accounting for the complex context in which they occur [50]. Without socially-grounded evaluation and concrete ethical controls, the development of multimodal LLMs remains optimized for publication, not for people.

Ignoring this twofold gap creates a direct pathway from benchmark success to real-world harm. Models that appear socially competent in controlled tests can be brittle in practice. This is supported by empirical work revealing a low correlation between the linguistic polish of LLMs and their actual social intelligence, often resulting in a "superficially friendly" style that is ungrounded in context [127]. This brittleness is amplified by evaluation practices that reward success on simplified tasks or biased datasets, as was seen with early benchmarks like Social-IQ, where models could exploit statistical patterns rather than engaging in genuine social reasoning [51]. When deployed in sensitive domains,

the failure modes are predictable and severe: 1) In mental healthcare and related public service contexts, misreading non-verbal cues by emotion recognition systems can lead to dangerous, inappropriate clinical responses and exacerbate existing inequities or harms [126]. 2) In education or counseling, where an embodied agent’s synthetic rapport—an empathetic tone, a nodding avatar—lends unearned credibility to incorrect guidance [136]. 3) In hiring and lending, where biased analysis of a candidate’s facial expressions or speech patterns from video interviews can perpetuate and scale discrimination [119]. In short, benchmark-optimized and ethically under-engineered systems do not just underperform; they externalize social and legal risks onto the very users and communities they are intended to serve. These are precisely the risks—fairness, bias, and societal harm—that our analysis shows are persistently under-mitigated in current multimodal systems.

To bridge this deployment gap, we advocate for a pragmatic re-orientation that tightly couples evaluation with ethics. We propose that researchers *evaluate for the behaviors they intend to deploy and co-measure success with risk exposure in the same study*. This involves several concrete steps:

1) Integrate Metrics: Pair quantitative system metrics with qualitative, human-in-the-loop evaluations that assess interaction quality, including reciprocity, responsiveness, transparency, and trust [8, 34]. **2) Benchmark for Doing, Not Naming:** Design evaluation tasks that require wise or appropriate action in realistic scenarios (e.g., de-escalating a conflict, providing an empathetic response), not just the labeling of static data. **3) Treat Ethics as Implementation:** Move beyond acknowledgement. Make at least one concrete mitigation a standard for any ethics-aware study—whether through privacy-by-design, human-in-the-loop governance, or robust content guardrails—with focused attention on the fairness and societal harms where mitigation currently lags. **4) Report for Impact, Not Proxies:** Frame results in terms of the full input → output → outcome chain, interpreting performance alongside measured risk.

Adopting this ethically-informed approach will help ensure that gains on benchmarks translate into credible evidence of socially competent, interaction-aware, and deployment-ready multimodal systems.

8.4 Social Creativity: a Blue Ocean Opportunity for Designing LLM-Supported Social Intelligence

Our review reveals a landscape shaped by striking imbalances in how multimodal, LLM-powered systems are designed, deployed, and evaluated for social intelligence. While systems excel at perception and recognition, they fall short in enabling interaction and remain especially underdeveloped in fostering creativity. These systemic gaps do more than mark limitations—they point to a blue ocean opportunity for designing LLM-based multimodal solutions that not only interpret social behavior but also support and promote social creativity.

Social creativity emerges as the most underdeveloped competency in our review, reflecting a fundamental mismatch between its divergent nature and the convergent design paradigm of current AI systems. In psychology, social creativity is defined as the capacity to generate multiple, diverse interpretations or solutions for a given social situation, grounded in counterfactual reasoning and Theory of Mind (ToM) [25, 100, 155]. Yet most systems are optimized to reduce ambiguity and converge on a single, most probable output. For example, intent recognition tasks that resolve uncertainty into one correct inference [77]. This pursuit of ground truth is inherently at odds with the exploratory and generative character of creativity.

The root of this limitation lies in the weak foundation of what has been termed Neural Theory-of-Mind (N-ToM) in Large Language Models [127]. Although robust ToM is considered essential for modeling others’ mental states [158], empirical studies show that LLMs rely on shallow heuristics and spurious correlations rather than genuine reasoning [127, 131]. Their apparent ToM quickly collapses under adversarial testing. The implication is clear: if systems cannot

reliably perform convergent ToM tasks, they are even less equipped to engage in the divergent reasoning needed for creativity.

These limitations are evident in real-world applications. Some systems show early signs of creativity, such as generating novel gestures for humanoid robots [61], but outputs typically collapse into a single “optimal” behavior. By contrast, human-centered contexts—from creative tasks seeking alignment with a subjective “vibe” [54] to home-based speech therapy requiring adaptable strategies [27]—demand repertoires of possibilities rather than singular solutions. Without the capacity for divergent, flexible responses, AI systems remain analytical observers rather than co-creative partners.

Beyond competency imbalances, the pursuit of social creativity also reveals structural gaps in modality, evaluation, and ethics as we identified in our earlier discussion. Divergent generation depends on nuanced multimodal signals (prosody, gaze, timing, embodied gestures) that are often stripped away when inputs are reduced to text or static frames [81, 100], raising doubts about whether textual abstraction alone can approximate human social nuance [39]. Evaluation practices further constrain progress: prevailing benchmarks reward convergence on a single correct answer, privileging recognition accuracy over generative flexibility [44]. Metrics for “naming” thus fail to capture success in “doing,” where the goal is to generate multiple plausible pathways toward social outcomes [100]. Finally, creativity introduces distinct ethical stakes. While divergent generation can expand user agency, it also risks producing misleading or manipulative behaviors and overwhelming users with unbounded options [100]. Ethical considerations must therefore shift from post-hoc safeguards to design preconditions to ensure outputs enhance trust and agency.

Looking forward, these gaps point toward a research agenda centered on social creativity, which we conceptualize not as artistic innovation, but as flexibility in social interpretation. This requires systems that can leverage ToM and counterfactual reasoning to generate diverse interpretations for a single social situation (e.g., a student’s silence) rather than converging on one label. The implications of achieving this are profound: it fosters a more human-like, contextually-sensitive AI that moves beyond labeling “engagement low” to suggesting possibilities (“student may be thinking, or hesitant”). Most importantly, this capability enables a paradigm shift from AI as a judge to AI as a partner, facilitating “AI-human co-interpretation.” By providing an “interpretive space” instead of a singular answer, the system empowers users (e.g., therapists or teachers) to make more empathetic and ethically-grounded decisions. Social creativity thus represents the frontier for transforming AI from an analytical observer into a co-creative partner that helps humans explore and navigate complex social relations.

8.5 Limitations and Future Work

Our review provides a clear, rigorously gathered snapshot of multimodal LLM work on social signals while making our scope choices explicit. The corpus draws on major indexes within a defined time window (so some gray/venue-specific work may be absent); the systems we synthesize mainly process visual and auditory inputs—the primary channels for face-to-face social cues; and we focus on input multimodality (how models interpret rich human behaviors) rather than output generation, which may underplay biosensor-centric or generative strands; our keyword emphasis could favor audiovisual pipelines; and an LLM-assisted coding workflow can misclassify. We mitigated these by searching multiple databases, tailoring Boolean queries, full-text screening, documenting criteria, anchoring on a manually coded seed set, human-verifying all machine labels, and checking inter-rater reliability. Field-wide tendencies toward short, benchmark-centric evaluations bound generalizability; we treat these as directions, not deficits. Looking ahead, we will apply the proposed framework in more diverse contexts—and within these transparent boundaries, our synthesis remains trustworthy and actionable.

References

- [1] Hasnat Md Abdullah, Tian Liu, Kangda Wei, Shu Kong, and Ruihong Huang. 2025. Ual-bench: The first comprehensive unusual activity localization benchmark. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 5801–5811.
- [2] Kaori Abe, Changqin Quan, Sheng Cao, and Zhiwei Luo. 2025. Classification of Properties in Human-like Dialogue Systems Using Generative AI to Adapt to Individual Preferences. *Applied Sciences* 15, 7 (2025), 3466.
- [3] Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024. iRAG: Advancing RAG for Videos with an Incremental Approach. *arXiv e-prints* (2024), arXiv–2404.
- [4] Aviral Agrawal, Carlos Mateo Samudio Lezcano, Iqui Balam Heredia-Marin, and Prabhdeep Singh Sethi. 2024. Listen then see: Video alignment with speaker attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018–2027.
- [5] Simge Akay, Duygu Cakir, and Nafiz Arica. 2025. Interpersonal Relationship Detection Using Multi-Head Graph Attention Networks with Multi-Feature Fusion. *IEEE Access* (2025).
- [6] Karl Albrecht. 2009. *Social intelligence: The new science of success*. John Wiley & Sons.
- [7] Morris Alper and Hadar Averbuch-Elor. 2023. Learning human-human interactions in images from weak textual supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2887–2899.
- [8] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (2014), 105–120.
- [9] Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2024. TrafficLens: Multi-Camera Traffic Video Analysis Using LLMs. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 3974–3981.
- [10] Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024. Vita: An efficient video-to-text algorithm using vlm for rag-based video analysis system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2266–2274.
- [11] Gaurav Arora, Shreya Jain, and Srjana Merugu. 2024. Intent Detection in the Age of LLMs. arXiv:2410.01627 [cs.CL] <https://arxiv.org/abs/2410.01627>
- [12] Abdul Basit and Muhammad Shafique. 2024. tinyDigiClones: A Multi-Modal LLM-Based Framework for Edge-optimized Personalized Avatars. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–9.
- [13] Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. 2025. Emergent Abilities in Large Language Models: A Survey. arXiv:2503.05788 [cs.LG] <https://arxiv.org/abs/2503.05788>
- [14] Antonio Blanco, Gerardo Pérez, Alicia Condón, Trinidad Rodríguez, and Pedro Núñez. 2024. AI-enhanced social robots for older adults care: evaluating the efficacy of ChatGPT-powered storytelling in the EBO platform. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2109–2116.
- [15] Borhane Bili-Hamelin, Leif Hancox-Li, and Andrew Smart. 2024. Unsocial Intelligence: An Investigation of the Assumptions of AGI Discourse. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 141–155.
- [16] Josep Bravo, Jonathan Cacace, Devis Dal Moro, Daniel Serrano, and Magi Dalmau-Moreno. 2024. Improving Robot Social Perception in Human-Robot-Interaction Using Multi-modal Cues. In *2024 7th Iberian Robotics Conference (ROBOT)*. IEEE, 1–6.
- [17] Runze Cai, Nuwan Janaka, Yang Chen, Lucia Wang, Shengdong Zhao, and Can Liu. 2024. PANDALens: Towards AI-assisted in-context writing on OHMD during travels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.
- [18] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815* (2019).
- [19] Adil Chakhtouna, Sara Sekkate, and Abdellah Adib. 2024. Multi-features learning via attention-blstm for speech emotion recognition. In *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*. IEEE, 1–6.
- [20] Zongyu Chang, Feihong Lu, Ziqin Zhu, Qian Li, Cheng Ji, Zhuo Chen, Hao Peng, Yang Liu, Ruifeng Xu, Yangqiu Song, Shangguang Wang, and Jianxin Li. 2025. Bridging the Gap Between LLMs and Human Intentions: Progresses and Challenges in Instruction Understanding, Intention Reasoning, and Reliable Generation. arXiv:2502.09101 [cs.HC] <https://arxiv.org/abs/2502.09101>
- [21] Korawat Charoenpitaks, Van-Quang Nguyen, Masanori Suganuma, Masahiro Takahashi, Ryoma Niihara, and Takayuki Okatani. 2024. Exploring the potential of multi-modal ai for driving hazard prediction. *IEEE Transactions on Intelligent Vehicles* (2024).
- [22] Keke Chen, T Zhewei, and Xiangbo Shu. 2024. Leveraging Multimodal Knowledge for Spatio-Temporal Action Localization. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 1–5.
- [23] Xiaoyu Chen, Wanru Xu, Shichao Kan, Linna Zhang, Yi Jin, Yigang Cen, and Yidong Li. 2025. Vision-semantics-label: A new two-step paradigm for action recognition with large language model. *IEEE Transactions on Circuits and Systems for Video Technology* (2025).
- [24] Yang Chen, Tian He, Junfeng Fu, Ling Wang, Jingcai Guo, Ting Hu, and Hong Cheng. 2024. Vision-language meets the skeleton: Progressively distillation with cross-modal knowledge for 3d action representation learning. *IEEE Transactions on Multimedia* (2024).
- [25] Kristin Conzelmann, Susanne Weis, and Heinz-Martin Süß. 2013. New findings about social intelligence. *Journal of Individual Differences* (2013).
- [26] Guangzhao Dai, Xiangbo Shu, Wenhao Wu, Rui Yan, and Jiachao Zhang. 2024. GPT4Ego: unleashing the potential of pre-trained models for zero-shot egocentric action recognition. *IEEE Transactions on Multimedia* (2024).
- [27] Aayushi Dangol, Aaleya Lewis, Hyewon Suh, Xuesi Hong, Hedda Meadan, James Fogarty, and Julie A Kientz. 2025. “I Want to Think Like an SLP”: A Design Exploration of AI-Supported Home Practice in Speech Therapy. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

- [28] Fernando M Delgado-Chaves, Matthew J Jennings, Antonio Atalaia, Justus Wolff, Rita Horvath, Zeinab M Mamdouh, Jan Baumbach, and Linda Baumbach. 2025. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences* 122, 2 (2025), e2411962122.
- [29] Shijian Deng, Erin E Kosloski, Siddhi Patel, Zeke A Barnett, Yiyang Nan, Alexander Kaplan, Sisira Aarukapalli, William T Doan, Matthew Wang, Harsh Singh, et al. 2024. Hear me, see me, understand me: Audio-visual autism behavior recognition. *IEEE Transactions on Multimedia* (2024).
- [30] Xi Ding and Lei Wang. 2025. Do language models understand time?. In *Companion Proceedings of the ACM on Web Conference 2025*. 1855–1868.
- [31] Yang Ding, Yi Dai, Xin Wang, Ling Feng, Lei Cao, and Huijun Zhang. 2024. Integrating Content-Semantics-World Knowledge to Detect Stress from Videos. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 10373–10381.
- [32] Wenlong Dong, Qing Zhu, and Qirong Mao. 2025. Key Clues Guided Video Character Social Relationship Recognition Enhanced by LLM. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [33] Xingning Dong, Zipeng Feng, Chunlun Zhou, Xuzheng Yu, Ming Yang, and Qingpei Guo. 2024. M2-RAAP: A multi-modal recipe for advancing adaptation-based pre-training towards effective and efficient zero-shot video-text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2156–2166.
- [34] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 278–288.
- [35] Vanessa Echeverria, Linxuan Zhao, Riordan Alfredo, Mikaela E Milesi, Yueqiao Jin, Sophie Abel, Jie Xiang Fan, Lixiang Yan, Samantha Dix, Rosie Wotherspoon, et al. 2025. TeamVision: An AI-powered Learning Analytics System for Supporting Reflection in Team-based Healthcare Simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [36] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. *arXiv preprint arXiv:2502.06559* (2025).
- [37] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2024. Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4769–4776.
- [38] Dong Feng, Ping Guo, Encheng Peng, Mingmin Zhu, Wenhao Yu, and Peng Wang. 2025. PoseLLaVA: Pose Centric Multimodal LLM for Fine-Grained 3D Pose Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 2951–2959.
- [39] Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. 2024. How far are we from agi: Are llms all we need? *arXiv preprint arXiv:2405.10313* (2024).
- [40] Tiantian Feng and Shrikanth Narayanan. 2024. Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 12116–12120.
- [41] Yueying Feng, WenKang Han, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, Jingyuan Chen, et al. 2024. E3: Exploring Embodied Emotion Through A Large-Scale Egocentric Video Dataset. *Advances in Neural Information Processing Systems* 37 (2024), 118182–118197.
- [42] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. 2024. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2093–2103.
- [43] Stephen M Fiore, Travis J Wiltshire, Emilio JC Lobato, Florian G Jentsch, Wesley H Huang, and Benjamin Axelrod. 2013. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology* 4 (2013), 859.
- [44] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296* (2024).
- [45] Michael Galarnyk, Veer Kejriwal, Agam Shah, Yash Bhardwaj, Nicholas Watney Meyer, Anand Krishnan, and Sudheer Chava. 2025. VideoConviction: A Multimodal Benchmark for Human Conviction and Stock Market Recommendations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 5447–5458.
- [46] Jianjun Gao, Kim-Hui Yap, Kejun Wu, Duc Tri Phan, Kratika Garg, and Boon Siew Han. 2024. Contextual human object interaction understanding from pre-trained large language model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 13436–13440.
- [47] Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Song Dingjie, Xidong Wang, Anningzhe Gao, Zhang Zhiyi, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria. *arXiv:2311.13951* [cs.CL] <https://arxiv.org/abs/2311.13951>
- [48] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2 (2020), 665–673.
- [49] Daniel Goleman. 2006. *Social intelligence: The new science of human relationships*. Bantam.
- [50] Elizabeth Goodman, Mike Kuniavsky, and Andrea Moed. 2012. *Observing the User Experience: A Practitioner’s Guide to User Research* (2 ed.). Morgan Kaufmann.
- [51] Xiao-Yu Guo, Yuan-Fang Li, and Gholamreza Haffari. 2023. DeSIQ: Towards an Unbiased, Challenging Benchmark for Social Intelligence Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 3169–3180.
- [52] Anshul Gupta, Pierre Vuillecard, Arya Farkhondeh, and Jean-Marc Odobez. 2024. Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 615–624.
- [53] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*:

Human Language Technologies, Volume 2 (Short Papers). 107–112.

- [54] Noor Hammad, C Ailie Fraser, Erik Harpstead, Jessica Hammer, and Mira Dontcheva. 2025. “It’s more of a vibe I’m going for”: Designing Text-to-Music Generation Interfaces for Video Creators. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 2738–2754.
- [55] Bin Han, Cleo Yau, Su Lei, and Jonathan Gratch. 2024. Knowledge-based emotion recognition using large language models. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–9.
- [56] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. Autoad: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18930–18940.
- [57] Tianhao He, Andrija Stanković, Evangelos Niforatos, and Gerd Kortuem. 2025. DesignMinds: Enhancing Video-Based Design Ideation with a Vision-Language Model and a Context-Injected Large Language Model. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. 1–15.
- [58] Moana Hendricks et al. 1969. Measuring Creative Social Intelligence. Final Report. (1969).
- [59] Chiori Hori, Motonari Kambara, Komei Sugiura, Kei Ota, Sameer Khurana, Siddarth Jain, Radu Corcodel, Devesh Jha, Diego Romeres, and Jonathan Le Roux. 2025. Interactive robot action replanning using multimodal llm trained from human demonstration videos. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [60] Jiaxing Huang and Jingyi Zhang. 2024. A Survey on Evaluation of Multimodal Large Language Models. arXiv:2408.15769 [cs.CV] <https://arxiv.org/abs/2408.15769>
- [61] Peide Huang, Yuhan Hu, Nataliya Nechyporenko, Daehwa Kim, Walter Talbott, and Jian Zhang. 2025. Emotion: Expressive motion sequence generation for humanoid robots with in-context learning. *IEEE Robotics and Automation Letters* (2025).
- [62] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems* 36 (2023), 72096–72109.
- [63] Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2023. Smile: Multimodal dataset for understanding laughter in video with language models. *arXiv preprint arXiv:2312.09818* (2023).
- [64] Lujain Ibrahim, Saffron Huang, Umang Bhatt, Lama Ahmad, and Markus Anderljung. 2024. Towards interactive evaluations for interaction harms in human-AI systems. *arXiv preprint arXiv:2405.10632* (2024).
- [65] Shunichiro Ito, Kanae Kochigami, and Takayuki Kanda. 2025. A Robot Dynamically Asking Questions in University Classes. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 839–848.
- [66] Simindokht Jahangard, Zhixi Cai, Shiki Wen, and Hamid Reza Tofighi. 2024. Jrdb-social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22087–22097.
- [67] Sandesh Jain, Surendrabikram Thapa, Kuan-Ting Chen, A Lynn Abbott, and Abhijit Sarkar. 2024. Semantic understanding of traffic scenes with large vision language models. In *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1580–1587.
- [68] Zipeng Ji, Pengcheng An, and Jian Zhao. 2025. ClassComet: Exploring and Designing AI-generated Danmaku in Educational Videos to Enhance Online Learning. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 552–575.
- [69] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743* (2024).
- [70] Manasa Kalanadhabhatta, Mohammad Mehdi Rastikerdar, Tauhidur Rahman, Adam S Grabel, and Deepak Ganesan. 2024. Playlogue: Dataset and benchmarks for analyzing adult-child conversations during play. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–34.
- [71] Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. 2025. SIV-Bench: A Video Benchmark for Social Interaction Understanding and Reasoning. *arXiv preprint arXiv:2506.05425* (2025).
- [72] Anton Korinek and Avital Balwit. 2022. *Aligned with whom? Direct and social goals for AI systems*. Technical Report. National Bureau of Economic Research.
- [73] Chunggi Lee, Tica Lin, Hanspeter Pfister, and Chen Zhu-Tian. 2024. Sportify: question answering with embedded visualizations and personified narratives for sports video. *IEEE Transactions on Visualization and Computer Graphics* (2024).
- [74] Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, et al. 2024. Towards social ai: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316* (2024).
- [75] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. 2024. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision* 16, 1-2 (2024), 1–214.
- [76] Chaoyu Li, Sid Padmanabhuni, Maryam S Cheema, Hasti Seifi, and Pooyan Fazli. 2025. Videoa11y: Method and dataset for accessible video description. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–29.
- [77] 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.
- [78] Jiahao Nick Li, Zhuohao Zhang, and Jiaju Ma. 2025. OmniQuery: Contextually Augmenting Captured Multimodal Memories to Enable Personal Question Answering. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

- [79] Ruizhe Li, Jiahao Guo, Mingxi Li, Zhengqian Wu, and Chao Liang. 2023. A hierarchical deep video understanding method with shot-based instance search and large language model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9425–9429.
- [80] Xin Li, Bin He, Zhipeng Wang, Yanmin Zhou, Gang Li, and Xiang Li. 2024. Toward cognitive digital twin system of human–robot collaboration manipulation. *IEEE Transactions on Automation Science and Engineering* (2024).
- [81] Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. 2025. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921* (2025).
- [82] Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. 2024. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284* (2024).
- [83] JongYoon Lim, Inkyu Sa, Bruce A MacDonald, and Ho Seok Ahn. 2024. Enhancing human-robot interaction: Integrating ASL recognition and LLM-driven co-speech gestures in Pepper robot with a compact neural network. In *2024 21st International Conference on Ubiquitous Robots (UR)*. IEEE, 663–668.
- [84] Maria R Lima, Amy O’Connell, Feiyang Zhou, Alethea Nagahara, Avni Hulyalkar, Anura Deshpande, Jesse Thomason, Ravi Vaidyanathan, and Maja Matorić. 2025. Promoting Cognitive Health in Elder Care with Large Language Model-Powered Socially Assistive Robots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [85] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. 2023. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2851–2862.
- [86] Xudong Lin, Ali Zare, Shiyuan Huang, Ming-Hsuan Yang, Shih-Fu Chang, and Li Zhang. 2024. Personalized Video Comment Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 16806–16820.
- [87] Edmond Liu, Jong Yoon Lim, Vineeth Johnson, Bruce MacDonald, and Ho Seok Ahn. 2025. SignPepper: Multimodal Social Robot for Sign Language Teaching. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1791–1793.
- [88] Jiaying Liu, Yunlong Wang, Yao Lyu, Yiheng Su, Shuo Niu, Xuhai" Orson" Xu, and Yan Zhang. 2024. Harnessing llms for automated video content analysis: An exploratory workflow of short videos on depression. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. 190–196.
- [89] Xiaofeng Liu, Qincheng Lv, Jie Li, Siyang Song, and Angelo Cangelosi. 2024. Multimodal emotion fusion mechanism and empathetic responses in companion robots. *IEEE Transactions on Cognitive and Developmental Systems* (2024).
- [90] Xingyu Bruce Liu, Jiahao Nick Li, David Kim, Xiang’Anthony’ Chen, and Ruofei Du. 2024. Human i/o: Towards a unified approach to detecting situational impairments. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [91] Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie A Pepler, and Karthik Ramani. 2024. Classmeta: Designing interactive virtual classmate to promote vr classroom participation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [92] Zhi-Song Liu, Robin Courant, and Vicky Kalogeiton. 2024. Funnynet-w: Multimodal learning of funny moments in videos in the wild. *International Journal of Computer Vision* 132, 8 (2024), 2885–2906.
- [93] Hao Lu, Xuesong Niu, Jiyao Wang, Yin Wang, Qingyong Hu, Jiaqi Tang, Yuting Zhang, Kaishen Yuan, Bin Huang, Zitong Yu, et al. 2024. Gpt as psychologist? preliminary evaluations for gpt-4v on visual affective computing. 2024 IEEE. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR) workshop*, Vol. 3.
- [94] Jin Lu. 2024. A face retrieval technique combining large models and artificial neural networks. *Concurrency and Computation: Practice and Experience* 36, 15 (2024), e8094.
- [95] Surbhi Madan, Shreya Ghosh, Lownish Rai Sookha, MA Ganaie, Ramanathan Subramanian, Abhinav Dhall, and Tom Gedeon. 2025. MIP-GAF: A MLLM-annotated Benchmark for Most Important Person Localization and Group Context Understanding. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1467–1476.
- [96] Mohamed Manzour Hussien, Angie Nataly Melo, Augusto Luis Ballardini, Carlota Salinas Maldonado, Rubén Izquierdo, and Miguel Ángel Sotelo. 2024. RAG-based Explainable Prediction of Road Users Behaviors for Automated Driving using Knowledge Graphs and Large Language Models. *arXiv e-prints* (2024), arXiv–2405.
- [97] Diako Mardanbegi, Nicholas Hylands, Neil Sarkar, and Nino Zahirovic. 2025. GazeLog: Optimizing Eye-Tracking with Fixation Keyframes & LLM Insights. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*. 1–8.
- [98] Iván Martín-Fernández, Sergio Esteban-Romero, Jaime Bellver-Soler, Fernando Fernández-Martínez, and Manuel Gil-Martín. 2024. Larger Encoders, Smaller Regressors: Exploring Label Dimensionality Reduction and Multimodal Large Language Models as Feature Extractors for Predicting Social Perception. In *Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor*. 20–27.
- [99] Iván Martín-Fernández, Sergio Esteban-Romero, Fernando Fernández-Martínez, and Manuel Gil-Martín. 2025. Parameter-Efficient Adaptation of Large Vision–Language Models for Video Memorability Prediction. *Sensors (Basel, Switzerland)* 25, 6 (2025), 1661.
- [100] Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing social intelligence in ai agents: Technical challenges and open questions. *arXiv preprint arXiv:2404.11023* (2024).
- [101] Genta Matsukawa and Atsuo Yoshitaka. 2024. Data Augmentation with Diffusion Model for Hand Detection. In *2024 International Symposium on Multimedia (ISM)*. IEEE, 170–173.
- [102] Lachlan McGinness, Peter Baumgartner, Esther Onyango, and Zelalem Lema. 2025. Highlighting Case Studies in LLM Literature Review of Interdisciplinary System Science. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 29–43.

- [103] Zixiang Meng, Qiang Gao, Di Guo, Yunlong Li, Bobo Li, Hao Fei, Shengqiong Wu, Fei Li, Chong Teng, and Donghong Ji. 2024. Mmlscu: A dataset for multi-modal multi-domain live streaming comment understanding. In *Proceedings of the ACM Web Conference 2024*. 4395–4406.
- [104] Mobina Moeini, Rouhollah Ahmadian, and Mehdi Ghathe. 2024. Calibrated SVM for probabilistic classification of in-vehicle voices into vehicle commands via voice-to-text LLM transformation. In *2024 8th International Conference on Smart Cities, Internet of Things and Applications (SCIoT)*. IEEE, 180–188.
- [105] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video ReCap: Recursive Captioning of Hour-Long Videos. *arXiv e-prints* (2024), arXiv–2402.
- [106] Jayanth Mohan, Spandana Reddy Mekapati, Premjith B, Jyothish Lal G, and Bharathi Raja Chakravarthi. 2025. A multimodal approach for hate and offensive content detection in Tamil: From corpus creation to model development. *ACM Transactions on Asian and Low-Resource Language Information Processing* 24, 3 (2025), 1–24.
- [107] Erika Mori, Yue Qiu, Hirokatsu Kataoka, and Yoshimitsu Aoki. 2025. A Comprehensive Analysis of a Social Intelligence Dataset and Response Tendencies Between Large Language Models (LLMs) and Humans. *Sensors* 25, 2 (2025), 477.
- [108] Seyed Mahed Mousavi, Edoardo Cecchinato, Lucia Hornikova, and Giuseppe Riccardi. 2025. Garbage In, Reasoning Out? Why Benchmark Scores are Unreliable and What to Do About It. arXiv:2506.23864 [cs.CL] <https://arxiv.org/abs/2506.23864>
- [109] Philipp Müller, Alexander Heimerl, Sayed Muddashir Hossain, Lea Siegel, Jan Alexandersson, Patrick Gebhard, Elisabeth André, and Tanja Schöneberger. 2024. Recognizing emotion regulation strategies from human behavior with large language models. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 210–218.
- [110] Farzeen Munir, Shoaib Azam, Tsvetomila Mihaylova, Ville Kyrki, and Tomasz Piotr Kucner. 2025. Pedestrian vision language model for intentions prediction. *IEEE Open Journal of Intelligent Transportation Systems* (2025).
- [111] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. 2023. Can language models learn to listen?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10083–10093.
- [112] Nguyen Nguyen, Jing Bi, Ali Vosoughi, Yapeng Tian, Pooyan Fazli, and Chenliang Xu. 2024. Oscar: Object state captioning and state change representation. *arXiv preprint arXiv:2402.17128* (2024).
- [113] Zheng Ning, Brianna L Wimer, Kaiwen Jiang, Keyi Chen, Jerrick Ban, Yapeng Tian, Yuhang Zhao, and Toby Jia-Jun Li. 2024. SPICA: interactive video content exploration through augmented audio descriptions for blind or low-vision viewers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [114] Razvan-George Pasca, Alexey Gavryushin, Muhammad Hamza, Yen-Ling Kuo, Kaichun Mo, Luc Van Gool, Otmar Hilliges, and Xi Wang. 2023. Summarize the Past to Predict the Future: Natural Language Descriptions of Context Boost Multimodal Object Interaction Anticipation. *arXiv preprint arXiv:2301.09209* (2023).
- [115] Dhruv Patel, Hamid Eghbalzadeh, Nitin Kamra, Michael Louis Iuzzolino, Unnat Jain, and Ruta Desai. 2023. Pretrained language models as visual planners for human assistance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15302–15314.
- [116] Duc Tri Phan, Vu Hoang Minh Doan, Jaeyeop Choi, Byeongil Lee, and Junghwan Oh. 2025. AADC-Net: A multimodal deep learning framework for automatic anomaly detection in real-time surveillance. *IEEE Transactions on Instrumentation and Measurement* (2025).
- [117] Penggang Qin, Tong Xu, Chao Zhang, Heda Wang, Yao Hu, and Enhong Chen. 2025. Scenario-aware Multimodal Chain-of-Thought Prompting for Rationales of Video Social Relations. *IEEE Transactions on Circuits and Systems for Video Technology* (2025).
- [118] Feng Qiu, Wei Zhang, Chen Liu, Lincheng Li, Heming Du, Tianchen Guo, and Xin Yu. 2024. Language-guided Multi-modal Emotional Mimicry Intensity Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4742–4751.
- [119] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481.
- [120] Shaun George Rajesh, Smriti Vipin Madangarli, Gauri Santosh Pisharady, and Rolla Subrahmanyam. 2025. Enhancement of Virtual Assistants through MultiModal AI for Emotion Recognition. *IEEE Access* (2025).
- [121] Tuval Raz, Roni Reiter-Palmon, and Yoed N Kenett. 2024. Open and closed-ended problem solving in humans and AI: the influence of question asking complexity. *Thinking Skills and Creativity* 53 (2024), 101598.
- [122] Weihong Ren, Jinguo Luo, Weibo Jiang, Liangqiong Qu, Zhi Han, Jiandong Tian, and Honghai Liu. 2024. Learning self-and cross-triplet context clues for human-object interaction detection. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 10 (2024), 9760–9773.
- [123] MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, and Diyi Yang. 2025. Egonormia: Benchmarking physical social norm understanding. *arXiv preprint arXiv:2502.20490* (2025).
- [124] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. 2024. Action scene graphs for long-form understanding of egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18622–18632.
- [125] Javier Rodriguez-Juan, David Ortiz-Perez, Jose Garcia-Rodriguez, David Tomás, and Grzegorz J Nalepa. 2025. Integrating advanced vision-language models for context recognition in risks assessment. *Neurocomputing* 618 (2025), 129131.
- [126] Kat Roemmich, Shanley Corvite, Cassidy Pyle, Nadia Karizat, and Nazanin Andalibi. 2024. Emotion AI Use in U.S. Mental Healthcare: Potentially Unjust and Techno-Solutionist. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–46.
- [127] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312* (2022).

- [128] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 400–408.
- [129] Tatsuki Seino, Naoki Saito, Takahiro Ogawa, Satoshi Asamizu, and Miki Haseyama. 2025. Expert Comment Generation Considering Sports Skill Level Using a Large Multimodal Model with Video and Spatial-Temporal Motion Features. *Sensors* 25, 2 (2025), 447.
- [130] Lala Shakti Swarup Ray, Bo Zhou, Sungho Suh, and Paul Lukowicz. 2024. OV-HHIR: Open Vocabulary Human Interaction Recognition Using Cross-modal Integration of Large Language Models. *arXiv e-prints* (2024), arXiv–2501.
- [131] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763* (2023).
- [132] Junxiao Shen, John J Dudley, and Per Ola Kristensson. 2024. Encode-Store-Retrieve: Augmenting Human Memory through Language-Encoded Egocentric Perception. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 923–931.
- [133] Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 1510–1519.
- [134] Mimo Shirasaka, Tatsuya Matsushima, Soshi Tsunashima, Yuya Ikeda, Aoi Horo, So Ikoma, Chikaha Tsuji, Hikaru Wada, Tsunekazu Omija, Dai Komukai, et al. 2024. Self-recovery prompting: Promptable general purpose service robot system with foundation models and self-recovery. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 17395–17402.
- [135] Gabriel Skantze and Bahar Irfan. 2025. Applying general turn-taking models to conversational human-robot interaction. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 859–868.
- [136] Ian Steenstra and Timothy W. Bickmore. 2025. A Risk Taxonomy for Evaluating AI-Powered Psychotherapy Agents. *arXiv preprint arXiv:2505.15108* (2025).
- [137] Xingpeng Sun, Yiran Zhang, Xindi Tang, Amrit Singh Bedi, and Aniket Bera. 2024. Trustnavgpt: Modeling uncertainty to improve trustworthiness of audio-guided llm-based robot navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8794–8801.
- [138] Kazuhiro Takemoto. 2024. The moral machine experiment on large language models. *Royal Society open science* 11, 2 (2024), 231393.
- [139] Daniel Tanneberg, Felix Ocker, Stephan Hasler, Joerg Deigmoeller, Anna Belardinelli, Chao Wang, Heiko Wersing, Bernhard Sendhoff, and Michael Gienger. 2024. To help or not to help: Llm-based attentive support for human-robot group interactions. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9130–9137.
- [140] Masatoshi Tateno, Takuma Yagi, Ryosuke Furuta, and Yoichi Sato. 2025. Learning Multiple Object States from Actions via Large Language Models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 9555–9565.
- [141] Jayant Teotia, Xulang Zhang, Rui Mao, and Erik Cambria. 2024. Evaluating Vision Language Models in Detecting Learning Engagement. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 496–502.
- [142] Thuvaraka Thayaparan, Sanjay Jayakumar, BA I Fernando, Gobisan Ananthanadarajan, Lakmini Abeywardana, and Dharshana Kasthurirathna. 2023. Digital Assistant for Visually Impaired People. In *2023 5th International Conference on Advancements in Computing (ICAC)*. IEEE, 340–345.
- [143] Edward L Thorndike. 1920. Intelligence and its uses. *Harper's magazine* 140 (1920), 227–235.
- [144] Pragya Tomar, Kirti Mathur, and Ugrasen Suman. 2022. Unimodal approaches for emotion recognition: A systematic review. *Cognitive Systems Research* 77 (11 2022). doi:10.1016/j.cogsys.2022.10.012
- [145] Philip E Vernon. 1933. Some characteristics of the good judge of personality. *The Journal of Social Psychology* 4, 1 (1933), 42–57.
- [146] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. 2008. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM international conference on Multimedia*. 1061–1070.
- [147] Binglu Wang, Yao Tian, Shunzhou Wang, and Le Yang. 2025. Multimodal Large Models Are Effective Action Anticipators. *IEEE Transactions on Multimedia* (2025).
- [148] Fanfan Wang, Heqing Ma, Jianfei Yu, Rui Xia, and Erik Cambria. 2024. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. *arXiv preprint arXiv:2405.13049* (2024).
- [149] Shiyu Wang, Shanyi Zhang, Fengtao Sun, Wenbai Chen, and Peiliang Wu. 2025. AffordStruct: Weakly Supervised Affordance Grounding based on Spatial Interaction and Knowledge-Aware. *IEEE Transactions on Automation Science and Engineering* (2025).
- [150] Xinru Wang, Mengjie Yu, Hannah Nguyen, Michael Iuzzolino, Tianyi Wang, Peiqi Tang, Natasha Lynova, Co Tran, Ting Zhang, Naveen Sendhilnathan, et al. 2025. Less or More: Towards Glanceable Explanations for LLM Recommendations Using Ultra-Small Devices. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 938–951.
- [151] Yonggu Wang, Yifan Shao, Zengyi Yu, and Zihan Wang. 2025. MS-RRBR: A Multi-Model Synergetic Framework for Restricted and Repetitive Behavior Recognition in Children with Autism. *Applied Sciences* 15, 3 (2025), 1577.
- [152] Yuxuan Wang, Yueqian Wang, Pengfei Wu, Jianxin Liang, Dongyan Zhao, Yang Liu, and Zilong Zheng. 2024. Efficient temporal extrapolation of multimodal large language models with temporal grounding bridge. *arXiv preprint arXiv:2402.16050* (2024).
- [153] James Wedeck. 1947. The relationship between personality and psychological ability. *British Journal of Psychology* 37, 3 (1947), 133.
- [154] Ran Wei, Rui Yan, Hongyu Qu, Xing Li, Qiaolin Ye, and Liyong Fu. 2025. SVMFN-FSAR: Semantic-Guided Video Multimodal Fusion Network for Few-Shot Action Recognition. *Big Data Mining and Analytics* 8, 3 (2025), 534–550.
- [155] Susanne Weis and Heinz-Martin Süß. 2005. Social intelligence—A review and critical discussion of measurement concepts. *Emotional intelligence: An international handbook* (2005), 203–230.

- [156] Bo Wen, Raquel Norel, Julia Liu, Thaddeus Stappenbeck, Farhana Zulkernine, and Huamin Chen. 2024. Leveraging large language models for patient engagement: The power of conversational ai in digital health. In *2024 IEEE International Conference on Digital Health (ICDH)*. IEEE, 104–113.
- [157] Mark R Westmoreland. 2022. Multimodality: reshaping anthropology. *Annual Review of Anthropology* 51, 1 (2022), 173–194.
- [158] Jessica Williams, Stephen M Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence* 5 (2022), 750763.
- [159] Travis J Wiltshire, Emilio J Lobato, Jonathan Velez, Florian Jentsch, and Stephen M Fiore. 2014. An interdisciplinary taxonomy of social cues and signals in the service of engineering robotic social intelligence. In *Unmanned Systems Technology XVI*, Vol. 9084. SPIE, 124–138.
- [160] Kaijie Xiao, Yi Gao, Fu Li, Weifeng Xu, Pengzhi Chen, and Wei Dong. 2024. ChatCam: Embracing LLMs for Contextual Chatting-to-Camera with Interest-Oriented Video Summarization. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–34.
- [161] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces*. 75–78.
- [162] Baijun Xie and Chung Hyuk Park. 2024. An Empathetic Social Robot with Modular Anxiety Interventions for Autistic Adolescents. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 1148–1155.
- [163] Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and Yingfei Sun. 2024. Academically intelligent llms are not necessarily socially intelligent. *arXiv preprint arXiv:2403.06591* (2024).
- [164] Yifang Xu, Yunzhuo Sun, Zien Xie, Benxiang Zhai, and Sidan Du. 2024. Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. *Applied Sciences* 14, 5 (2024), 1894.
- [165] Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, et al. 2025. Socialmaze: A benchmark for evaluating social reasoning in large language models. *arXiv preprint arXiv:2505.23713* (2025).
- [166] Vera Yang, Archita Srivastava, Yasaman Etesam, Chuxuan Zhang, and Angelica Lim. 2023. Contextual emotion estimation from image captions. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.
- [167] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (Nov. 2024). doi:10.1093/nsr/nwae403
- [168] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (2024), nwae403.
- [169] Zefang Yu, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2024. From raw video to pedagogical insights: A unified framework for student behavior analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23241–23249.
- [170] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8807–8817.
- [171] Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. 2024. AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. (2024).
- [172] Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, et al. 2025. Socialeval: Evaluating social intelligence of large language models. *arXiv preprint arXiv:2506.00900* (2025).
- [173] Zixiang Zhou, Yu Wan, and Baoyuan Wang. 2024. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1357–1366.
- [174] Fatemeh Ziaetabar, Minija Tamosiunaite, and Florentin Wörgötter. 2024. A hierarchical graph-based approach for recognition and description generation of bimanual actions in videos. *IEEE Access* (2024).