Socially Intelligent Robots and Large Behavior Models: Challenges, Strategies, and Future Research Opportunities

Short Paper

Aaron Elkins
San Diego State University
San Diego, California, USA
aelkins@sdsu.edu

Sanchit Singh
San Diego State University
San Diego, California, USA
ssingh1949@sdsu.edu

Mary Pourebadi
San Diego State University
San Diego, California, USA
mpourebadi@sdsu.edu

Uyiosa Amadasun
San Diego State University
San Diego, California, USA
uamadasun@sdsu.edu

Kaveh Abhari
San Diego State University
San Diego, California, USA
kabhari@sdsu.edu

Abstract

Large Behavioral Models (LBMs) offer promise for developing socially intelligent robots capable of adaptive, multimodal interaction. Yet their progress is constrained by fragmented pipelines and insufficiently grounded inputs. We argue that LBMs must be trained through the same communicative channels humans use—verbal and non-verbal (gaze, gesture, posture, affect, timing). Using a clinical problematization approach, we examine upstream pipelines that distill raw signals into structured, socially meaningful cues. Foundational models—particularly LLMs and VLMs—can transduce speech, text, images, and video into compact social state variables, but their limitations remain underexplored. Across five diagnostic probes, we find limitations in temporal coherence and cross-modal fusion. By surfacing constraints, this study establishes diagnostic groundwork for architectural innovation and advances an agenda to bridge the gap between current model capacity and the demands of socially intelligent robotics.

Keywords: Socially Intelligent Robots, Large Behavioral Models (LBMs), Multimodal Foundation Models, Clinical Problematization, Human-Robot Interaction, AI Design

Introduction

Social robots—autonomous systems designed to engage humans in meaningful, affective, and goal-directed interactions—are undergoing a transformative shift propelled by recent advancements in multimodal foundation models (Chi et al., 2023; Isbister et al., 2022; Andersen et al., 2021). These systems operate as complex information systems, integrating multimodal inputs—from audio and vision to sensor streams—and transforming them into coordinated physical, verbal, and visual behaviors that are socially situated and oriented toward sustaining meaningful human interaction (Tuja et al., 2020). In doing so, they embody a complex sociotechnical problem: the challenge of synchronizing technical precision with the fluid,

emotionally charged norms of human social interaction (Gnewuch et al., 2024; You & Robert, 2018). Thus, social robots increasingly rely on large-scale AI architectures that integrate perception, language, and action to navigate the complexity of human environments (Shrestha et al., 2024). Central to this evolution are large language, vision, and video models, as well as emerging models for behavior and perception (Kim et al., 2024). Collectively, these models offer promising steps toward enabling robots to interpret instructions, perceive and analyze their surroundings, and learn from demonstrations (e.g., RT-1, RT-2) (Obrenovic et al., 2024). While these systems demonstrate how multimodal fusion could eventually enable more context-aware, fluent, and adaptable robot behaviors, they also highlight fundamental issues in accounting for the sociotechnical properties of human interactions (Techatassanasoontorn et al., 2023). To address this issue, developing foundational AI models is both essential and challenging. Among existing solutions, Large Behavior Models (LBMs) present a promising pathway. However, the application of LBMs is hindered by the complexity of feeding them behavioral data (Barreiros et al., 2025; Eliot, 2024; Salimpour et al., 2025). This study, therefore, asks: *How can a multimodal fusion pipeline—integrating vision, language, and video models—be designed to generate the sociobehavioral data required for LBMs?*

LBMs hold significant promise for enabling transformer-based systems that scale efficiently, especially in embodied AI settings (Wirtz & Stock-Homburg, 2025). These models are developed by aggregating behavioral data—motion trajectories, sensorimotor signals, and human interaction traces—into unified learning architectures (Bartalesi et al., 2024; Wirtz & Stock-Homburg, 2025). This shared-data paradigm, akin to transfer learning on a behavioral level, allows robots to inherit reusable skill priors from previously trained agents, significantly reducing the need for task-specific retraining or expensive real-world trials. Consequently, LBMs have the potential to accelerate deployment cycles, reduce development overhead, and make social robotics more accessible to a broader range of institutions and communities. Initiatives such as *BridgeData* and *RoboNet* exemplify this approach, providing large-scale, heterogeneous datasets that enhance behavioral generalization across diverse tasks and environments.

This study takes an initial step in problematizing the preparation of socially grounded inputs for LBMs, thereby shaping an emerging research agenda. Using clinical problematization, we conducted exploratory probes that exposed critical limitations and challenged prevailing assumptions about multimodal fusion, identity preservation, and behavioral fidelity. In doing so, we reframed and formalized the problem space for designing socially intelligent robots, emphasizing the need for upstream architectural innovation rather than downstream engineering fixes. Addressing these challenges is crucial for advancing socially intelligent robots and, in general, demonstrates how IS research can contribute to a new generation of socially aware, adaptive, and context-responsive technologies.

Background

Social robots are autonomous systems designed to interact with humans through diverse communication modalities (Huange et al, 2025). They employ multimodal interfaces—including speech, gestures, facial expressions, and proxemics—to support interactions that are intuitive and contextually appropriate (Hlee et al., 2023). Socially intelligent robots form a specialized subset, defined by their ability to interpret, respond to, and anticipate complex human behaviors and emotions. By integrating perceptual inputs with contextual cues, they generate behaviors that align with social norms and expectations. Designing such robots highlights the classic sociotechnical gap (Robert et al., 2024), necessitating a reconciliation between human expectations and technical capabilities (Gnewuch et al., 2024; Tuja et al., 2020; You & Robert, 2018). This challenge compels designers to address the social, emotional, and contextual dimensions of human-robot interaction (Mitchell & Jeon, 2025) through transformer-based models that enable robots to learn, recalibrate, and refine behaviors, thereby sustaining trust, coherence, and meaningful engagement over time (You & Robert, 2024). However, the relative immaturity of such models for social interaction tasks poses significant challenges (Salimpour et al., 2025). A central obstacle is the reliance on foundational models developed for general robotics, which often fail to account for the contextual nuances of social environments (Tang et al., 2025). A raised hand, for example, may signal a request for assistance in a hospital yet indicate a desire to participate in a classroom. Likewise, successful object handover hinges on spatio-temporal understanding: the giver initiates a reach, stabilizes the object, and pauses; the receiver aligns, grasps, and withdraws. Absent models of event order and timing, a robot may grasp too early or fail to yield. Research must therefore move beyond performance assessment to examine how robots interpret such cues. Such inquiry can lay the empirical groundwork for designing socially intelligent robots.

Behavior Models for Socially Intelligent Robots

Advancing socially intelligent robots requires flexible foundational models, as early rule-based or task-specific supervised methods are too rigid for dynamic social settings. Transformer-based architectures have transformed this landscape (Salimpour et al., 2025). For example, Large Language Models (LLMs) enable robots to understand user intent, hold conversations, and generate contextually appropriate responses and Vision-Language Models (VLMs) allow robots to identify people, objects, and actions, and describe or respond to what they "see" contextually. Pre-trained on large multimodal datasets, these models capture patterns across language, vision, and temporal sequences (Chen & Huang, 2024). When integrated into behavioral pipelines, these models can provide richer and more accurate inputs, enabling robots to respond more appropriately across various contexts. Recent research highlights the value of such integration for optimizing social interaction (Lin et al., 2024; Mahmud et al., 2025). By simulating human-like perception, these models allow robots to engage in fluid, multimodal, and context-aware exchanges that approximate natural communication (Izumi et al., 2024). Building on this foundation, we focus on the emerging paradigm of LBMs—defined as pre-trained models adapted to encode, predict, and generate coherent behavioral trajectories across multimodal social contexts by integrating language, perception, action, and affect over extended interactions (Salimpour et al., 2025; Sartor & Thompson, 2024).

Preparing Socially Grounded Inputs for LBMs

LBMs represent a significant advancement in modeling complex, multimodal human behavior. Pre-trained to encode, predict, and generate coherent behavioral trajectories across language, perception, action, and affect, they are inherently contextual (Barreiros et al., 2025; Salimpour et al., 2025; Sartor & Thompson, 2024). Nevertheless, these capacities do not automatically translate to real-world social settings. Social robots operate amid fluid, culturally inflected norms, shifting roles, and situational cues; without accurate social information at the input layer, LBMs misread context, fragment temporally, and fail to personalize. Accordingly, the first requirement for reliable, sociotechnical operation is socially grounded input: structured, context-rich signals—identity-aware, time-stamped, and multimodally aligned—that capture intent, affect, proxemics, and interaction history. LBMs must receive such inputs from upstream transformer models that can detect and encode these cues with sufficient fidelity (Figure 1); otherwise, their learned representations cannot align with the demands of embodied, real-time interaction. While the technical design of such architectures is beyond the scope of this short paper, our focus is on questioning their design from an IS perspective: the conditions or reasons that prevent robots from becoming trustworthy IS artifacts, or what hinders them. Without this groundwork, LBMs will continue to struggle with generalizing to the complexities of everyday human–robot interaction.

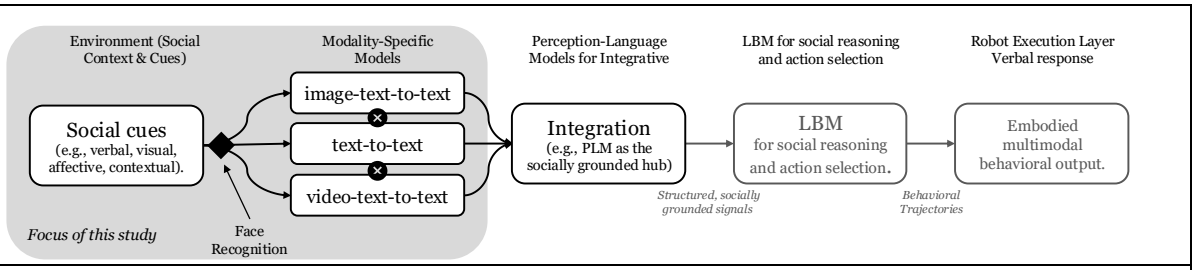

Figure 1. Pipeline for Preparing Socially Grounded Inputs for Social Robots via an LBM

Methodology

We adopt a clinical problematization methodology—a disciplined diagnostic approach for interrogating sociotechnical systems to expose their limits, latent assumptions, and internal contradictions. The goal is not to indict the system but to catalyze new lines of inquiry and guide responsible redesign. This empirically grounded method draws on Foucault's (1972) notion of problematization as a "critical history of the present," extends Alvesson and Sandberg's (2011) strategy for generating novel research questions, and is operationalized through Ulrich's (1983) critical systems heuristics, which emphasize boundary critique and reflective practice. Unlike generic forms of critique, clinical problematization is structured, interventionist, and forward-looking. It is particularly vital in the age of agentic artifacts and autonomous systems, where

problems and solutions co-evolve beyond any stable frame of reference. In this study, we apply the method to identify systemic limitations and surface new design opportunities. The process unfolds in three steps:

1. *Uncovering Assumptions*. The first step identifies what systems take for granted by surfacing hidden expectations, neglected signals, and implicit boundaries that define how they are presumed to operate. In our context, this means examining multimodal pipelines to articulate the kinds of inputs they privilege, the cues they overlook, and the interactional settings they assume as standard.
2. *Exposing Breakdowns*. The second step challenges these assumptions by deliberately testing them against empirical reality, revealing tensions, contradictions, and structural weaknesses. For our study, this involves probing where multimodal pipelines fail to uphold their own assumptions, reframing the problem space in terms that are concrete and analytically tractable.
3. *Formalizing Testable Problems*. The final step converts diagnostic insights into precise problem statements with explicit evaluation criteria. Within our context, this means articulating measurable targets—such as temporal consistency, identity continuity, or multimodal coherence—so that critique leads to testable interventions rather than remaining at the level of abstract diagnosis.

Through this approach, we expose the infrastructural fragilities and misalignments within current LBM pipelines and propose empirically grounded directions for their responsible design and evolution.

From Social Signals to LBM-Ready Inputs: An Exploratory Technical Diagnosis

Study Design. Instead of assessing LBM implementation, we developed a series of exploratory probes to test whether—and how—socially grounded signals can be captured, represented, and transmitted to LBMs as usable inputs. Our in-vitro protocol tested three stages: (1) capturing and curating social cues (such as identity anchors, affect, turn-taking, proxemics); (2) learning representations into summaries; and (3) creating handoff schemas that package these summaries as LBM-friendly trajectories, tags, or constraints. These prefigurative probes (Garud et al., 2008) revealed limits and design challenges early, helping to clarify input specifications for future LBMs. The pipeline integrated VLM modules for perceptual extraction and temporal modeling, with LLM summarization to generate coherent, time-aligned narratives suitable for LBM ingestion. Each probe examined one stage of the input pipeline. This sequence operationalized clinical problematization by revealing failure modes and identifying the conditions under which inputs can be made reliable, thereby reframing the design problem as one of specifying upstream inputs. The outcome provided evidence about the feasibility and format of socially grounded inputs, offering concrete specifications for future LBM training and control.

Data Preparation. We used egocentric and static-camera clips of common social gestures (e.g., waving, nodding, attentive stance) designed to simulate human–robot encounters. We recorded a series of short video clips, each demonstrating various social cues. Each video was preprocessed to enrich visual inputs with identity-aware annotations. Videos were sampled at a consistent rate of two frames per second (i.e., every 15th frame from a 30-*fps* recording) to reduce redundancy while preserving temporal continuity. Each extracted frame was passed through a face detection pipeline. This pipeline identified all visible faces in a frame and compared them against a dictionary of pre-encoded facial embeddings representing known individuals. Matching was determined using cosine similarity, with a threshold of 0.9 for confirmation. When a match was found, a bounding box was drawn around the face with the person's name as a label. Unrecognized faces were also annotated with a bounding box labeled "unknown." These annotated frames were used as inputs to the VLM, allowing it to produce behaviorally grounded descriptions such as "John is drinking" rather than generic outputs like "a person is drinking." This preprocessing step was crucial to ensure that the behavioral modeling pipeline could distinguish and track individual actions over time.

Diagnostic Exploration Settings and Evaluation Criteria. We implemented upstream input pipelines that integrate VLMs and LLMs to simulate the processing of socially relevant cues. We examined five configurations, including image-text-to-text model only, image-text-to-text model plus LLM, and video-text-to-text model only (Figure 1). Each configuration was qualitatively evaluated based on social-reasoning criteria (Lee et al., 2025). Specifically, model outputs were assessed based on: (1) *Identity consistency*: Tracking whether preprocessing labels persisted through model processing, (2) *Action accuracy*: Alignment between visual events and textual descriptions, (3) *Temporal coherence*: Preservation of event sequence and progression, (4) *Narrative Utility*: Practical relevance for downstream LBM consumption. These criteria were applied across experiments. These criteria served as diagnostic indicators rather than performance metrics, helping identify systemic failures in the pipeline.

Diagnostic Probe a. Static Behavior Interpretation from Individual Frames using an image-text-to-text model. This baseline probe evaluated the ability of VLMs to interpret human actions and context from a single frame. Using LLaVA-1.5-7B, we prompted the model to describe key elements such as objects, agents, and actions. Though behavior is inherently temporal, we evaluated whether static observations are sufficient to generate semantically meaningful behavioral cues.

Diagnostic Probe b. Incremental Behavioral Summarization with LLM Integration. In real-world scenarios, social robots must not only detect actions but also maintain an evolving understanding of behavioral context. To model this, we introduced a temporal summarization step using LLaMA-3.1-8B-Instruct (Meta, 2024). After obtaining each frame description from the VLM, we passed it along with the prior summary into the LLM to generate an updated behavioral narrative. This test modeled a simplified form of long-term behavior tracking, where the robot builds a growing understanding of a user's activity sequence.

Diagnostic Probe c. Time-Aware Behavior Modeling via Timestamp Conditioning. Sequential behavior modeling benefits from temporal awareness—recognizing not just what happened, but when. To encourage temporal coherence in the summary generation, we extended the prior setup (probe b) by incorporating timestamp metadata for both current and previous frames. These timestamps were used as additional conditioning input to the LLM, signaling the progression of events. The goal was to examine whether explicit temporal cues improved behavioral narrative continuity and sequencing.

Diagnostic Probe d. Direct Behavioral Inference from Raw Video using a video-text-to-text model. To examine the previous approaches, we conducted this experiment using the model LLaVA-NeXT-Video-34B-hf (LLaVA Team, 2025), which accepts raw video as input. Unlike the other probe, this model attempts to achieve end-to-end video understanding without requiring intermediate frame-level processing or summarization. Videos of varying lengths were tested, and responses were analyzed for accuracy.

Diagnostic Probe e. Short Clip Inference and Narrative Fusion. This experiment integrated all three models to generate cohesive behavioral narratives. Short clips (3–5s) were processed with the video-text-to-text model for high-level summaries, while keyframes were analyzed with a VLM for detailed static observations. These outputs with optional memory were fused with LLM prompted to produce spatiotemporally coherent narratives. The goal was to evaluate the effectiveness of multimodal fusion in understanding social cues.

Results

Our diagnostic probes uncovered fundamental architectural limitations that hinder current models from producing reliable inputs for LBMs. We identified four key failure modes across five pipeline configurations. These failure modes serve as a roadmap for future design and evaluations.

1. *Temporal Blindness Despite Sequential Processing*. Probes a and b exposed a lack of temporal sensitivity in foundation models. The VLM generated identical captions across 3–30 seconds of video (e.g., repeatedly describing "man holding cup and wearing glasses"), failing to detect behavioral progression. LLM augmentation likewise produced static accounts despite explicit action sequences. This shows transformer architectures lack implicit temporal reasoning, treating frames as isolated. For social robots, failing to perceive unfolding events weakens their responsiveness during dynamic interactions.
2. *Preprocessing Signal Loss Across Model Boundaries*. Probe c demonstrated that facial recognition preprocessing, although technically accurate, failed to propagate through the pipeline. The VLM either could not read or ignored these annotations, consistently outputting "Unknown" despite clear labels such as "Phillip" or "Sanchit." This indicates vision–language models treat preprocessing as noise rather than a usable signal. For social robots, this means identity recognition remains fragile, limiting continuity in personalized engagement.
3. *Error Cascade in Multi-Modal Fusion*. Probe e surprisingly showed that fusion amplifies rather than corrects model errors. Frame analysis generated massive redundancy, while LLM synthesis fabricated non-existent individuals and events. Instead of complementary strengths, model weaknesses are compounded, resulting in hallucinated narratives that are detached from reality. For social robots, this means multimodal integration risks creating distorted representations of social context, leading to incoherent or inappropriate behavior.
4. *Optimization for Plausibility Over Accuracy*. Probe d with end-to-end video models yielded generic summaries that omitted key actions. For example, when a subject removed their glasses,

the model only described peripheral activities, such as drinking coffee. This reveals an optimization toward statistically plausible narratives rather than faithful event detection. This means social signals may be overlooked, resulting in shallow or misleading interpretations of human behavior.

Together, these findings suggest that the gaps between current capabilities and LBM requirements originate from architectural constraints rather than integration engineering. Models lack temporal memory, fail to preserve preprocessing signals, and privilege narrative plausibility over behavioral fidelity—limitations that must be resolved for robots to engage socially with humans.

Discussion

This study employed clinical problematization to interrogate how socially grounded inputs are prepared for LBMs. By surfacing hidden assumptions and stress-testing them against empirical reality, we developed a diagnostic foundation for advancing the next stage of innovation. Our systematic probes of multimodal pipelines revealed that current foundation models fail to produce the temporally coherent, identity-consistent, and behaviorally faithful data streams that socially intelligent robots require. These findings extend beyond technical shortcomings: they unsettle three prevailing assumptions that have shaped ongoing development efforts and, in doing so, reframe the problem space that future research must address.

1. The assumption that multimodal fusion inherently enhances robustness is misplaced. Rather than compensating for individual weaknesses, fusion pipelines produced redundancy, contradictions, and hallucinated content. This suggests that integration alone does not ensure reliability; for social robots, naïve fusion risks distorting situational awareness and producing inappropriate responses.
2. The assumption that preprocessing enrichments naturally strengthen model performance proves untenable. Although facial recognition and labeling worked at the extraction stage, these signals failed to carry through the pipeline. The inability to preserve identity across representational boundaries points to a deeper incompatibility. For robots, this fragility undermines the continuity required for personalized and socially meaningful engagement.
3. The assumption that coherent narratives equate to behavioral accuracy is flawed. End-to-end video models generated summaries that were statistically plausible but behaviorally incomplete, overlooking micro-actions such as “removing glasses” in favor of generic accounts like “drinking coffee.” For social robots, this optimization toward plausibility over fidelity results in shallow situational understanding and misaligned embodied action.

Together, these challenges suggest that the central problem is not how to better integrate or fine-tune existing foundation models, but how to design new architectures and evaluation regimes that can sustain temporal state, preserve identity across transformations, and prioritize behavioral fidelity over narrative plausibility. The results also reveal that the gap between human expectations and robotic capabilities cannot be bridged through downstream engineering; it requires upstream architectural innovation. The assumption that scale and data alone will yield social intelligence obscures a fundamental sociotechnical reality: human interaction is temporally situated, multimodally entangled, and deeply interpersonal. Advancing socially intelligent robotics, therefore, demands a new generation of computational paradigms explicitly crafted to capture these dynamics, setting a research agenda that aligns technical development with the embodied, interactive, and relational nature of social life.

Limitations and Future Research Avenues

This study served as a diagnostic probe to empirically problematize the development of upstream pipelines for socially intelligent robots. As such, it carries several limitations that should inform future, more comprehensive experimentation. Its primary aim was to reframe the problem space and highlight the social dimensions of HRI, rather than to deliver a fully integrated technical solution. The probes were conducted in controlled laboratory conditions with scripted and brief scenarios, which likely underrepresent the variability and ambiguity of real-world social cues. Evaluation relied on a bespoke rubric—covering identity consistency, event-boundary agreement, temporal coherence, and handoff validity—rather than established community benchmarks, limiting cross-system comparability. Some observed failures may also reflect design choices, including preprocessing methods, model parameterization, and cascading integration errors, complicating causal attribution. Finally, only a subset of models and configurations was tested; alternative baselines or fine-tuned variants might produce different results. While these constraints narrow

immediate generalizability, they provide a structured foundation for LBM handoff. Building on this diagnostic orientation, we reframe the design space and identify five sociotechnical domains where foundational assumptions must be challenged (Table 1). Together, these domains identify structural obstacles in current pipelines and outline a research agenda for developing reliable architectures—an agenda crucial to the advancement of socially intelligent robots.

Domain	Technical HRI Implications	Social HRI Implications
Temporal Coherence	Improves spatiotemporal reasoning through attention and memory mechanisms, enabling recognition of evolving behavioral sequences.	Supports smoother turn-taking, joint attention, and trust by aligning robot actions with the temporal flow of social interaction.
Contextual Continuity	Integrates dynamic context windows and long-term memory to sustain shared state across interactions.	Enhances conversational flow, fosters rapport, and creates a sense of being "remembered," for relational depth.
Personalization Fidelity	Embeds identity-tagged multimodal traces and tracking modules for individualized adaptation.	Reinforces user identity, belonging, and trust through recognition and tailored interaction styles.
Communicative Novelty	Deploys entropy-based sampling, adaptive prompting, and feedback to reduce redundancy.	Increases engagement and authenticity by avoiding monotony and fostering adaptive, responsive dialogue.
Multimodal Integrity	Advances fusion architectures that integrate speech, gestures, facial expressions, and context.	Improves situational attunement, embodied interaction, and cultural sensitivity in complex social settings.

Table 1. Future Research Domains and Its Implications for Socially Intelligent Robots

Research Domain 1: Temporal Coherence, Building Fluid Awareness of Social Dynamics. Current pipelines assume that sequential frame processing is sufficient for capturing temporal dynamics. In practice, this reduces social action to static snapshots, erasing the flow of intent, transitions, and turn-taking. From a sociotechnical perspective, this reflects a mismatch between the linear logic of computational models and the fluid rhythms of human interaction. Research must explore temporal attention, hierarchical memory, and anchoring strategies to develop input streams that capture lived temporality in social encounters (e.g., Kim et al., 2024; Dillon et al., 2025).

Research Domain 2: Contextual Continuity, Sustaining Engagement Across Interactions. Designers often assume short-term memory windows can sustain engagement. Yet socially intelligent interaction depends on remembering histories, roles, and evolving goals over time. This limitation exposes the gap between the technical constraints of context windows and the sociotechnical demand for continuity. Future work should pursue pipelines with dynamic context windows and long-term memory modules, enabling robots to sustain conversations, build rapport, and support enduring social contracts (e.g., He et al., 2024).

Research Domain 3: Personalization Fidelity, Aligning Behavior with Identity and History. Prevailing designs treat personalization as an emergent property of scale, assuming large datasets can approximate individual differences. However, socially intelligent robots must account for identity, preference, and history to maintain trust and resonance. This limitation highlights a structural blind spot: the neglect of individual-level granularity in favor of generalized averages. Research should integrate identity-tagged multimodal traces and user-history tracking to produce pipelines that respect the situated, relational character of personalization in social contexts (e.g., Bian et al., 2024).

Research Domain 4, Communicative Novelty, Reducing Redundancy for Natural Interaction. Generative systems often equate plausibility with success, yielding repetitive, low-variance outputs that erode the authenticity of interaction. This reflects the assumption that smooth redundancy is preferable to occasional error. In sociotechnical terms, this privileges system stability over the dynamism of human dialogue, where novelty signals attentiveness and meaning. Research must explore entropy-based sampling, adaptive prompting, and reinforcement learning to reduce redundancy and foster more varied, socially engaging exchanges that sustain user trust.

Research Domain 5, Multimodal Integrity, Fusing Signals for Consistent Social Understanding. Current approaches presume that integrating modalities will naturally yield coherent representations. Instead, fusion often amplifies inconsistencies, producing mismatches across vision, language, and context. This limitation reveals the fallacy that technical aggregation alone creates shared meaning. From a sociotechnical lens, coherence requires architectures that respect the interdependence of signals—facial expressions, gestures, speech, proxemics—within a shared context. Future research should advance spatiotemporal graph networks and multimodal transformers to create pipelines that model the embodied, culturally situated nature of human communication (e.g., Song et al., 2021).

Conclusion

This study highlights the challenges of developing socially grounded inputs for LBMs. While foundational models in robotics have advanced rapidly, their social dimensions remain underdeveloped, particularly in parsing, contextualizing, and sustaining social cues. From an IS perspective, this gap reflects a persistent misalignment between technical progress and the sociotechnical requirements of HRI (Robert et al., 2024; Turja et al., 2020; You & Robert, 2018). By applying clinical problematization, we revealed structural barriers in input pipelines that challenge prevailing assumptions and reframe the design agenda. In doing so, we contribute to IS research on socially intelligent robots by identifying upstream architectural domains—temporal coherence, contextual continuity, personalization fidelity, communicative novelty, and multimodal integrity—that must be addressed to enable adaptive, context-sensitive, and coherent interaction (Chi et al., 2023; Gnewuch et al., 2024; Hlee et al., 2023).

References

Alvesson, M., & Sandberg, J. (2011). Generating research questions through problematization. *Academy of Management Review, 36*(2), 247–271. https://doi.org/10.5465/amr.2009.0188

Andersen, T. O., Nunes, F., Wilcox, L., Coiera, E., & Rogers, Y. (2023). Introduction to the special issue on human-centred AI in healthcare: Challenges appearing in the wild. *ACM Transactions on Computer-Human Interaction, 30*(2), Article 25. https://doi.org/10.1145/3589961

Barreiros, J., Beaulieu, A., Bhat, A., Cory, R., Cousineau, E., Dai, H., Fang, C. H., Hashimoto, K., Irshad, M. Z., Itkina, M., & Kuppuswamy, N. (2025). A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*

Bartalesi, V., Lenzi, E., & De Martino, C. (2024). Using large language models to create narrative events. *PeerJ Computer Science, 10*, e2242. https://doi.org/10.7717/peerj-cs.2242

Bian, T., Ma, Y., Chollet, M., Sanchez, V., & Guha, T. (2024). Interact with me: Joint egocentric forecasting of intent to interact, attitude and social actions. *arXiv preprint arXiv:2412.16698*

Chen, C. T., & Huang, H. H. (2024). Integrating LLM, VLM, and Text-to-Image Models for Enhanced Information Graphics: A Methodology for Accurate and Visually Engaging Visualizations. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (pp. 8627-8630).

Chi, O. H., Chi, C. G., Gursoy, D., & Nunkoo, R. (2023). Customers' acceptance of artificially intelligent service robots: The influence of trust and culture. *International Journal of Information Management, 70*, Article 102623. https://doi.org/10.1016/j.ijinfomgt.2023.102623

Dillon, F., Halvorsen, G., Tattershall, S., Rowntree, M., & Vanderpool, G. (2025). Contextual memory reweaving in large language models using layered latent state reconstruction. *arXiv:2502.02046*

Eliot, L. (2024). Large behavior models surpass large language models to create AI that walks and talks. *Forbes*.

Foucault, M. (1972). *The archaeology of knowledge*. New York: Pantheon Books.

Garud, R., Jain, S., & Tuertscher, P. (2008). Incomplete by design and designing for incompleteness. *Organization Studies, 29*(3), 351–371. https://doi.org/10.1177/0170840607088018

Gnewuch, U., Morana, S., Hinz, O., Kellner, R., & Maedche, A. (2024). More than a bot? The impact of disclosing human involvement on customer interactions with hybrid service agents. *Information Systems Research, 35*(3), 936–955. https://doi.org/10.1287/isre.2022.0152

He, B., Li, H., Jang, Y. K., Jia, M., Cao, X., Shah, A., & Lim, S. N. (2024). Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13504–13514.

Hlee, S., Park, J., Park, H., Koo, C., & Chang, Y. (2023). Understanding customer's meaningful engagement with AI-powered service robots. *Information Technology & People, 36*(3), 1020–1047. https://doi.org/10.1108/ITP-10-2020-0740

Huang, T. L., Liao, G. Y., Dennis, A. R., & Teng, C. I. (2025). High efficiency or easy troubleshooting? Human use of autonomous Mobile Healthcare Robots. *Decision Support Systems, 193*, Article 114453. https://doi.org/10.1016/j.dss.2025.114453

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., & Zou, J. (2023). A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine, 29*(9), 2307–2316. https://doi.org/10.1038/s41591-023-02504-3

Isbister, K., Cottrell, P., Cecchet, A., Dagan, E., Theofanopoulou, N., Bertran, F. A., & Slovak, P. (2022). Design (not) lost in translation: A case study of an intimate-space socially assistive "robot" for emotion regulation. *ACM Transactions on Computer-Human Interaction, 29*(4), Article 38. https://doi.org/10.1145/3491083

Izumi, K., Tanaka, H., Shidara, K., Adachi, H., Kanayama, D., Kudo, T., & Nakamura, S. (2024). Response generation for cognitive behavioral therapy with large language models: a comparative study with Socratic questioning. *arXiv preprint arXiv:2401.15966*

Kim, W., Choi, C., Lee, W., & Rhee, W. (2024). An image grid can be worth a video: Zero-shot video question answering using a VLM. *IEEE Access, 12*, 12345–12356. https://doi.org/10.1109/ACCESS.2024.3517625

Kim, Y., Kim, D., Choi, J., Park, J., Oh, N., & Park, D. (2024). A survey on integration of large language models with intelligent robots. *Intelligent Service Robotics, 17*(5), 1091–1107. https://doi.org/10.1007/s11370-024-00550-5

Lee, D. W., Kim, Y., Guvenoz, D., Jeong, S., Malachowsky, P., Morency, L. P., & Park, H. W. (2025). The Human Robot Social Interaction (HSRI) Dataset: Benchmarking Foundational Models' Social Reasoning. *arXiv preprint arXiv:2504.13898*

Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., & Han, S. (2024). Vila: On pre-training for visual language models. *The IEEE Conference on Computer Vision and Pattern Recognition*, 26689-26699.

LLaVA Team. (2025). LLaVA-NeXT-Video-34B-hf [Model]. Hugging Face.

Magrabi. (2021). Realizing AI in healthcare: Challenges appearing in the wild. In *Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*.

Meta. (2025). LLaMA-3.1-8B-Instruct [Model]. Hugging Face.

Mitchell, J. J., & Jeon, M. (2025). Exploring emotional connections: A systematic literature review of attachment in human-robot interaction. *International Journal of Human–Computer Interaction, 41*(3), 311–325. https://doi.org/10.1080/10447318.2024.2445100

Obrenovic, B., Gu, X., Wang, G., Godinic, D., & Jakhongirov, I. (2024). Generative AI and human–robot interaction: Implications and future agenda for business, society and ethics. *AI & Society, 40*(2), 677–690. https://doi.org/10.1007/s00146-024-01889-0

Robert Jr, L. P., Fantinato, M., You, S., & Hung, P. C. (2024). Social robotics business and computing. *Information Systems Frontiers, 26*(1), 1-8. https://doi.org/10.1007/s10796-023-10413-6

Salimpour, S., Fu, L., Keramat, F., Militano, L., Toffetti, G., Edelman, H., & Queralta, J. P. (2025). Towards Embodied Agentic AI: Review and Classification of LLM-and VLM-Driven Robot Autonomy and Interaction. *arXiv preprint arXiv:2508.05294*

Sartor, S., & Thompson, N. (2024). Neural Scaling Laws in Robotics. *arXiv preprint arXiv:2405.14005*.

Shrestha, S., Zha, Y., Banagiri, S., Gao, G., Aloimonos, Y., & Fermuller, C. (2024). Natsgd: A dataset with speech, gestures, and demonstrations for robot learning in natural human-robot interaction. *arXiv preprint arXiv:2403.02274*

Song, X., Chen, J., Wu, Z., & Jiang, Y. G. (2021). Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia, 24*, 2914–2923. https://doi.org/10.1109/TMM.2021.3090595

Tang, C., Tang, C., Gong, S., Kwok, T. M., & Hu, Y. (2025). Robot Character Generation and Adaptive Human-Robot Interaction with Personality Shaping. *arXiv preprint arXiv:2503.15518*

Techatassanasoontorn, A. A., Waizenegger, L., & Doolin, B. (2023). When Harry, the human, met Sally, the software robot: Metaphorical sensemaking and sensegiving around an emergent digital technology. *Journal of Information Technology, 38*(4), 416-441. https://doi.org/10.1177/02683962231157426

Turja, T., Aaltonen, I., Taipale, S., & Oksanen, A. (2020). Robot acceptance model for care (RAM-care): A principled approach to the intention to use care robots. *Information & Management, 57*(5), Article 103220. https://doi.org/10.1016/j.im.2019.103220

Wirtz, J., & Stock-Homburg, R. (2025). Generative AI meets service robots. *Journal of Service Research, 28*(2), 123–141. https://doi.org/10.1177/10946705251340487

You, S., & Robert, L. P. (2018). Emotional attachment, performance, and viability in teams collaborating with embodied physical action (EPA) robots. *Journal of the Association for Information Systems, 19*(5), 377-407. https://doi.org/10.17705/1jais.00496

You, S., & Robert, L. (2024). Trusting and working with robots: A relational demography theory of preference for robotic over human co-workers. *MIS Quarterly, 48*(4): 1297–1330. https://doi.org/10.25300/MISQ/2023/17403