Designing Socially Grounded Data Pipelines for Training and Operating Socially Intelligent Robots: Challenges and Future Directions

Abstract

Designing socially intelligent robots presents a frontier challenge in human–robot interaction and information systems research, where technical architectures must align with the embodied, context-rich demands of social life. Current Vision–Language–Action (VLA) models integrate multimodal inputs but fall short in supporting socially coherent behavior, limiting their value as training pipelines. Using a clinical problematization approach, this study examines experimental pipelines to identify structural limitations in temporal coherence, continuity, multimodal integration, and affective inference. These shortcomings reveal a deeper misalignment between prevailing training paradigms and the sociotechnical requirements of interactional integrity. In response, we highlight perception–language models (PLMs) as a more socially attuned substrate, capable of extracting contextual signals and enhancing situational grounding for behavior modeling. We conclude by outlining a research agenda that advances spatiotemporal reasoning, affective modeling, and embodied coherence, thereby contributing to IS discourse on the design of trustworthy, socially adaptive robotic systems.

Keywords: Perception-Language Models, Socially Intelligent Robots, Robot Training Models, Multimodal AI Systems, Human-Robot Interaction.

1 Introduction

Social robots—autonomous agents engineered for purposeful, affective, and socially meaningful interaction—are undergoing a significant evolution fueled by breakthroughs in foundation models (Chi et al., 2023; Isbister et al., 2022; Andersen et al., 2021). At the heart of this transformation lies the integration of perception, language, and action within expansive AI frameworks designed to manage the intricacies of human-centered environments (Salinas-Martínez et al., 2024; Shrestha et al., 2024). This shift is driven by the emergence of large-scale architectures, alongside a new class of large behavior models aimed at capturing complex social dynamics (e.g., Kim et al., 2024).

Together, the rise of these models marks a pivotal step toward endowing robots with the ability to follow instructions, interpret dynamic environments, and acquire skills through demonstration. Early systems—such as RT-1 and RT-2—have begun to exhibit limited task generalization in structured scenarios like tabletop manipulation. However, translating these capabilities into open-ended, socially complex environments remains a formidable frontier (Obrenovic et al., 2024). Most training pipelines still rely on variations of supervised learning, reinforcement learning, and imitation learning, often enhanced through synthetic and simulated data to support scalability. State-of-the-art systems, such as RFM-1 and Gemini Robotics, demonstrate the potential of multimodal integration, showing how the coordinated processing of sensory, linguistic, and action inputs can lay the groundwork for more fluid and context-sensitive behavior.

This study redirects attention from task-level performance to a more foundational question: *how to help robots receive meaningful inputs for understanding social cues—a prerequisite for interpreting human intent and sustaining social interaction*. The issue lies not only in generating outputs but in preparing upstream pipelines that deliver context-rich signals to the model. Nevertheless, most current learning paradigms emphasize downstream execution, leaving the design of socially grounded inputs underdeveloped.

Vision–Language–Action models, such as PaLM-E, have emerged as unified architectures to facilitate robotic learning. Yet, their utility in social settings is hampered by persistent structural training deficiencies. Building reliable upstream pipelines requires expansive, high-fidelity feeds of socially situated interactions (Wirtz & Stock-Homburg, 2025). Without such data, models struggle with temporal alignment, identity continuity, and affective attunement—a bottleneck that continues to constrain the advancement of socially intelligent robotics.

Emerging architectures promise to scale training pipelines by aggregating multimodal traces across tasks and contexts (Wirtz & Stock-Homburg, 2025; Bartalesi et al., 2024). Their success, however, depends on upstream models that feed reliable data for both training and operation, capturing the full spectrum of embodied, socially situated interactions (Wirtz & Stock-Homburg, 2025). These models must translate raw sensory inputs (visual, audio, and spatial) into structured social signals. When this translation falters, robots misinterpret context, fragment temporally, or overlook affect, undermining social responsiveness. Problematizing input preparation, therefore, shifts attention upstream: the challenge is not only to design behavioral architectures but to ensure they are nourished with socially grounded, reliable data. Meeting this challenge is labor-intensive and technically demanding, requiring robust architectures that can generalize across contexts while avoiding overfitting.

This study examines cost-efficient training strategies by foregrounding their limitations. We adopt clinical problematization, a data-driven diagnostic method that interrogates existing systems to surface hidden assumptions, contradictions, and structural blind spots. Unlike abstract forms of problematization (Alvesson & Sandberg, 2011), this approach is empirically anchored and oriented toward actionable reframing. Its goal is not critique for its own sake but the catalytic identification of unresolved issues that can inspire innovation. Such a lens is particularly vital in an era of autonomous, agentic systems, where both problems and solutions evolve unpredictably beyond conventional design boundaries. In our context, clinical problematization enables us to empirically challenge prevailing assumptions and chart new lines of inquiry for advancing sociotechnical design. Through five experimental training pipelines, we expose representational gaps that hinder alignment between technical architectures and social interaction. These gaps reveal design priorities for building cost-effective and scalable pipelines that deliver socially grounded data inputs for socially intelligent robots.

This study makes a dual contribution. First, it provides a practical foundation for refining training methodologies in socially demanding contexts such as assistive care, therapeutic engagement, and collaborative service roles (Huang et al., 2025). These applications require more than task efficiency; they demand sensitivity to affective cues, adaptability to shifting dynamics, and the capacity to sustain coherent, context-aware interaction (Gnewuch et al., 2025; Mitchell & Jeon, 2025). Our findings surface the prerequisites for such fluency and inform future research on socially grounded training pipelines. Second, we advance the Information Systems (IS) literature by positioning clinical problematization as a robust methodological tool for the AI era. As robots increasingly act autonomously within social settings (You & Robert, 2024), the boundaries between technical artifacts, social actors, and ethical agents become increasingly blurred (Hlee et al., 2023; Turja et al., 2020). Clinical problematization enables IS researchers to interrogate these entanglements, pinpoint systemic shortcomings, and open pathways for novel inquiry, especially when artifacts remain immature and their functions unsettled.

2 Background

Social robots are autonomous agents designed to interact with humans through socially meaningful behavior and multimodal communication (Huange et al, 2025). By combining modalities such as speech, gaze, gesture, and spatial positioning, these robots aim to support naturalistic interactions that feel intuitive and responsive (You & Robert, 2024). Their use spans domains such as education, healthcare, and service environments, where social fluency is not merely ancillary but essential to user engagement and system efficacy (Breazeal, 2003).

Among these systems, socially intelligent robots stand out for their ability to navigate the subtleties of human affect and interactional norms (Fang et al., 2025). They must go beyond basic input-output routines, continuously sensing, interpreting, and adapting to dynamic social cues. Training such nuanced behavior has often relied on standard VLA models, which fuse visual, textual, and behavioral inputs to simulate socially responsive outputs. However, while VLA models offer a modular approach to multimodal learning, they often fall short in capturing the richness and temporality of social interaction. The reason is that they are not designed to train on socially grounded data (Sapkota et al., 2025).

Emerging behavior models (LBMs) offer a promising alternative due to their capacity for learning holistic, context-sensitive patterns of behavior. However, these models are also notoriously difficult to train, owing to their need for dense, structured, and socially grounded data. In this context, we argue that perception-language models (PLMs)—which integrate fine-grained perceptual understanding with language-based reasoning—offer a more socially attuned pathway for training language-based models.

2.1 Training Socially Intelligent Robots

Earlier approaches to training foundational robotic models, including supervised and reinforcement learning, have shown promise in constrained settings but fall short when confronted with the fluid, situated, and ambiguous nature of real-world social interactions. These methods lack both the adaptability and the semantic depth needed to support generalized social fluency (Zolanvari et al., 2025).

Architectures, such as the fusion of Large Language Models (LLMs) and Vision-Language Models (VLMs), offer new pathways by encoding relationships across modalities (Chen & Huang, 2024). This multimodal grounding has enabled researchers to construct behavioral systems that can operate flexibly across a range of social contexts, making interactions more robust and generalizable (Sabo et al., 2024). Studies have increasingly emphasized the importance of such integrative approaches for enhancing the quality and responsiveness of social robot behavior (Lin et al., 2024; Mahmud et al., 2025). These models approximate human-like perceptual capacities, enabling robots to understand social cues and engage in socially coherent interactions that mirror natural human communication (Izumi et al., 2024).

In this study, we assume socially intelligent robots are enabled by LBMs (e.g., transformer-based architectures fine-tuned to encode, anticipate, and generate coherent behavioral sequences across multimodal, temporally extended social scenarios) (Salimpour et al., 2025; Stipancic et al., 2016). Our focus, however, is not on evaluating these models directly but on the upstream data pipelines that make them viable. Building such pipelines—capable of capturing temporally ordered social episodes and encoding their interactional meaning—remains the fundamental bottleneck. Without reliable, socially grounded inputs, even the most advanced architectures misinterpret context, fragment temporally, and fail to sustain adaptive, human-centered interaction. Addressing this bottleneck is therefore not a technical afterthought but the core design challenge for advancing socially intelligent robotics.

2.2 Socially Grounded Inputs for Large Behavior Models in Social Robots

Social robots must navigate fluid, ambiguous interactions shaped by cultural norms, interpersonal dynamics, and situational cues (Breazeal, 2003). To succeed, their internal representations must align with the perceptual and temporal realities of social life. However, foundational robotic models, even when enhanced by LBMs, are not plug-and-play solutions for real-world interaction. Their dependency on context makes socially grounded input essential—not merely as raw sensory data, but as the basis for grounding meaning, maintaining temporal coherence, and disambiguating social intent (Duncan et al., 2024). Without reliable input pipelines, robots are prone to fragmentation, identity drift, and semantic errors during live interaction (Firoozi et al., 2025).

To address this challenge, we foreground the role of upstream architectures and assess how existing models can support socially grounded input preparation (Table 1). Socially grounded data are inherently event-driven, requiring models to integrate signals across multiple modalities and timescales (Wang et al., 2023). Yet even state-of-the-art techniques such as multimodal pretraining or reinforcement learning from human feedback (RLHF) struggle to produce socially coherent, adaptive behavior—not because models cannot process such data, but because upstream pipelines fail to reliably capture and encode it from the environment (Ouyang et al., 2022; Kim et al., 2024). The central challenge, therefore, is not merely building more powerful models, but ensuring that the subtleties of social interaction are made intelligible to them—both during training and in real-time operation.

Table 1. Foundational Models Supporting Socially Intelligent Robots

Model	Core Function	Function in Social Robots	Training Input Requirements	Operational Input Requirements	Key Limitations
LLMs (Large Language Models)	Language understanding & generation	Interpret intent, sustain dialogue, and generate contextually appropriate responses.	Massive text corpora (books, web data, dialogues)	User prompts, conversational history	Strong in language but ungrounded; lack perceptual and situational awareness.
VLMs (Vision-Language Models)	Visual–linguistic grounding	Recognize people, objects, and actions; connect what robots "see" with language.	Image–text or video–text pairs (captioned datasets)	Real-time visual frames, contextual snapshots	Capture static scenes well but weak in temporality and affective nuance.
VLAs (Vision-Language-Action Models)	Multimodal perception–action mapping	Translate sensory input and instructions into motor behaviors (object handover, navigation…).	Paired perception–action traces, robot demonstrations	Continuous sensorimotor streams during tasks	Effective for task execution but brittle in fluid, socially complex contexts.
LBMs (Large Behavior Models)	Social behavior modeling	Generate coherent multimodal behavioral sequences across time and context.	Rich, annotated datasets (dialogue, gestures, affect, temporal sequences)	Live multimodal streams capturing evolving social cues	Data-hungry, costly to build; generalization to real-world social contexts remains limited.
PLMs (Perception-Language Models)	Context-rich social perception	Extract subtle cues (gaze, posture, prosody, affect) and integrate them with language for situational awareness.	Aligned perceptual–linguistic corpora; weak supervision and narrative scaffolds	Real-time perception across modalities (vision, audio, spatial…)	Promising for grounding, but integration into full behavioral pipelines is still emerging.

3 Methodology

This study adopts a clinical problematization approach to interrogate the limitations of current training pipelines for socially intelligent robots. This empirically based approach builds on Foucault's (1972) idea of problematization as a "critical history of the present." It expands Alvesson and Sandberg's (2011) method for developing new research questions and is implemented using Ulrich's (1994) critical systems heuristics, which focus on boundary critique and reflective practice. Clinical problematization involves empirically grounded analysis of system breakdowns, representational gaps, and structural misalignments—not to critique for its own sake, but to surface possibilities for redesign. Unlike abstract theorizing, this method is oriented toward transformation: it treats persistent failures and contextual mismatches as diagnostic signals for rethinking foundational assumptions. The process unfolds in three steps:

(a) Uncovering Assumptions. The initial step is to identify what systems take for granted by revealing hidden expectations, overlooked signals, and implicit boundaries that shape their presumed functioning. In our case, this involves examining multimodal pipelines to specify the types of inputs they favor, the cues they miss, and the interactional contexts they consider standard.

(b) Exposing Breakdowns. The second step involves challenging these assumptions by testing them against actual data, uncovering tensions, contradictions, and flaws in the structure. For our research, this means investigating where multimodal upstream pipelines fail to meet their own assumptions and reframing the problem in clear, analyzable terms.

(c) Formalizing Testable Problems. The final step is to turn diagnostic insights into concrete problem statements with clear evaluation standards for future research. In our context, this entails defining measurable goals—such as temporal consistency, identity preservation, or multimodal coherence—to ensure critiques lead to actionable solutions instead of remaining abstract.

This process enables us to evaluate and reconfigure training strategies for socially intelligent robots with particular attention to how different upstream architectures can enrich social grounding. However, rather than optimizing toward a fixed performance goal, our methodology foregrounds experimental variation as a discovery tool and diagnostic probe—a way to surface hidden costs, trade-offs, and affordances across architectures. In this sense, we treat experimentation not as the pursuit of immediate solutions, but as a structured site of design inquiry that reframes problems and guides innovation.

3.1 Study Design for Behavior Modeling

To explore how socially intelligent robots can be better trained, we developed a sequence of interconnected experiments that use VLMs, LLMs, and PLMs to generate structured input streams for behavior modeling. The focus was on producing inputs that are usable in both training and real-time operation, thereby testing the robustness of the upstream pipelines. The experimental pipeline relied on egocentric and fixed-camera recordings of common social gestures—such as waving, nodding, and displaying attentiveness—constructed to resemble everyday human–robot encounters. Across experimental setups, we systematically varied input formats, architectural combinations, and perceptual scaffolds to assess how different configurations shape the quality of socially grounded signals. Each experiment was therefore positioned within a clinical problematization framework, not as a stand-alone evaluation but as part of a cumulative diagnostic process. This allowed us to expose recurring limitations in identity preservation, action sequencing, and multimodal contextualization. In doing so, we positioned failure not as an endpoint but as a generative signal for reframing the design space.

3.2 Identity-Aware Frame Preprocessing

To enable context-sensitive behavior modeling, we implemented an identity-encoding pipeline embedded directly into the visual stream. Video data were downsampled to two frames per second, a rate selected to preserve temporal continuity while reducing computational overhead. Each frame was processed through a facial detection and recognition module, leveraging a pre-trained embedding model for feature extraction. Candidate faces were compared against a reference dictionary of known identities using cosine similarity, with matches confirmed at a confidence threshold of 0.9.

Frames containing recognized individuals were tagged with labeled bounding boxes, while unrecognized faces were labeled as "unknown." This preprocessing step ensured that subsequent VLM modules could reference identity-specific anchors in generating behavioral narratives (e.g., "Phillip nods in agreement" rather than "a person nods in agreement"). By embedding persistent identity markers across sequential frames, the pipeline preserved temporal continuity. It enabled more accurate tracking of social dynamics—an essential requirement for modeling behavior in human–robot interaction.

3.3 Diagnostic Experiment Configurations

Building on the identity-aware preprocessing stage, we implemented and evaluated a series of unified upstream pipelines designed to provision LBMs with socially grounded inputs. Each pipeline combined VLMs, LLMs, and PLMs to approximate end-to-end processing of socially salient cues. To ensure comparability, all experiments were conducted on a series of short, controlled clips representing identical social scenarios—such as waving, object handovers, and drinking gestures embedded within conversational settings.

The purpose of these experiments was not to optimize benchmark scores but to expose structural limitations in current modeling pipelines. Specifically, we sought to diagnose representational breakdowns, including identity ambiguity, contextual fragmentation, and temporal discontinuity. Because outputs were generated as open-text behavioral narratives rather than categorical labels, conventional accuracy metrics were unsuitable. Instead, we adopted qualitative evaluation criteria aligned with the requirements of human–robot interaction: temporal ordering of actions, continuity of identity tracking, and richness of contextual detail. This approach with a referential anchor allowed us to systematically assess whether socially relevant information was preserved, degraded, or lost during multimodal transformations.

Experiment (a): Frame-Level Action Interpretation. This baseline test evaluated the ability of image–text transformer models to infer social behavior from isolated frames. Short clips ranging from 3 to 30 seconds were uniformly downsampled (≈2 FPS), yielding representative frames. Each frame was processed with LLaVA-1.5-7B (LLaVA Team, 2024), prompted to extract objects, agents, and actions. Because no temporal continuity was available, this setup tested whether static visual descriptions could still yield semantically coherent behavioral cues. Outputs were examined for consistency, descriptive granularity, and ability to incorporate identity labels introduced during preprocessing.

Experiment (b): Sequential Summarization with LLM Fusion. To approximate an evolving social context, we combined VLM-derived frame descriptions with incremental updates from LLaMA-3.1-8B-Instruct (Meta, 2024). At each step, the LLM received the current frame description and the accumulated summary, generating a running narrative. This experiment tested whether language-based reasoning layers could mitigate the temporal blindness of frame-based perception and produce progressively enriched accounts of an activity sequence—an essential capability for long-horizon social interaction modeling.

Experiment (c): Temporal Conditioning with Explicit Timestamps. Expanding on Experiment (b), we introduced explicit temporal metadata into the summarization process. Each frame description was paired with its capture timestamp before being passed to the LLM (LLaMA-3.1-8B-Instruct; Meta, 2024). This conditioning was intended to anchor the generated narrative in temporal order, enabling the model to distinguish event progression (e.g., reaching → grasping → withdrawing) rather than collapsing them into static labels. Evaluation focused on whether explicit time cues improved continuity, sequencing, and contextual fidelity.

Experiment (d): End-to-End Video-to-Text Inference. In this configuration, we tested LLaVA-NeXT-Video-34B-hf (LLaVA Team, 2024), a large video–text model capable of processing raw video directly. Unlike the staged approaches above, this model attempted holistic video interpretation without frame-level decomposition or external summarization. We varied clip duration (≈3 to 30 seconds) to assess how input length influenced narrative fidelity. The goal was to investigate whether direct spatiotemporal modeling could yield more accurate, temporally aware behavioral accounts, or whether it would exacerbate common issues such as generic summarization and omitted micro-actions.

Experiment (e): PLM-Driven Contextual Inference. The final experiment explored perception–language models as a candidate upstream substrate for socially grounded input preparation. Leveraging recent advances in unified perception–localization–reasoning architectures, the PLM (FB's Perception-LM-1B) was tasked with generating behaviorally rich narratives from the same interaction clips. We focused on whether PLM outputs demonstrated improved temporal grounding, sensitivity to identity cues, and contextual richness compared to VLM–LLM or direct video–text pipelines. This experiment provided an initial assessment of how PLMs might address persistent architectural gaps and better support the demands of socially intelligent HRI.

3.4 Evaluation Criteria

Model outputs were evaluated qualitatively based Model outputs were assessed qualitatively using four social-reasoning criteria (Gandhi et al., 2023; Lee et al., 2025). First, *identity consistency* was evaluated to determine whether individuals were recognized and referenced coherently across frames, thereby enabling continuity in personalization. Second, *action accuracy* reflected the interpretive alignment between model descriptions and observed gestures, such as nodding or waving. Third, *temporal coherence* examined whether the generated narratives preserved the logical flow of unfolding events rather than collapsing them into static accounts. Finally, *narrative utility* was considered in terms of the descriptive richness and situational relevance of outputs for informing downstream interactions, such as dialogue or decision-making. These criteria were applied interpretively across experiments to highlight comparative tendencies and systemic shortcomings in different architectures, rather than reducing outcomes to quantitative scores. From an HRI perspective, a qualitative lens is especially appropriate: socially intelligent interaction is inherently open-ended, context-dependent, and relational, and thus better evaluated through interpretive diagnostics than via fixed benchmarks.

4 Results

Our diagnostic evaluation identified multiple recurrent challenges in the processing of socially relevant video input feeding LBMs by current foundational models. These patterns were consistently observed across five pipeline configurations. Nonetheless, they should be regarded as diagnostic tendencies rather than definitive measurements.

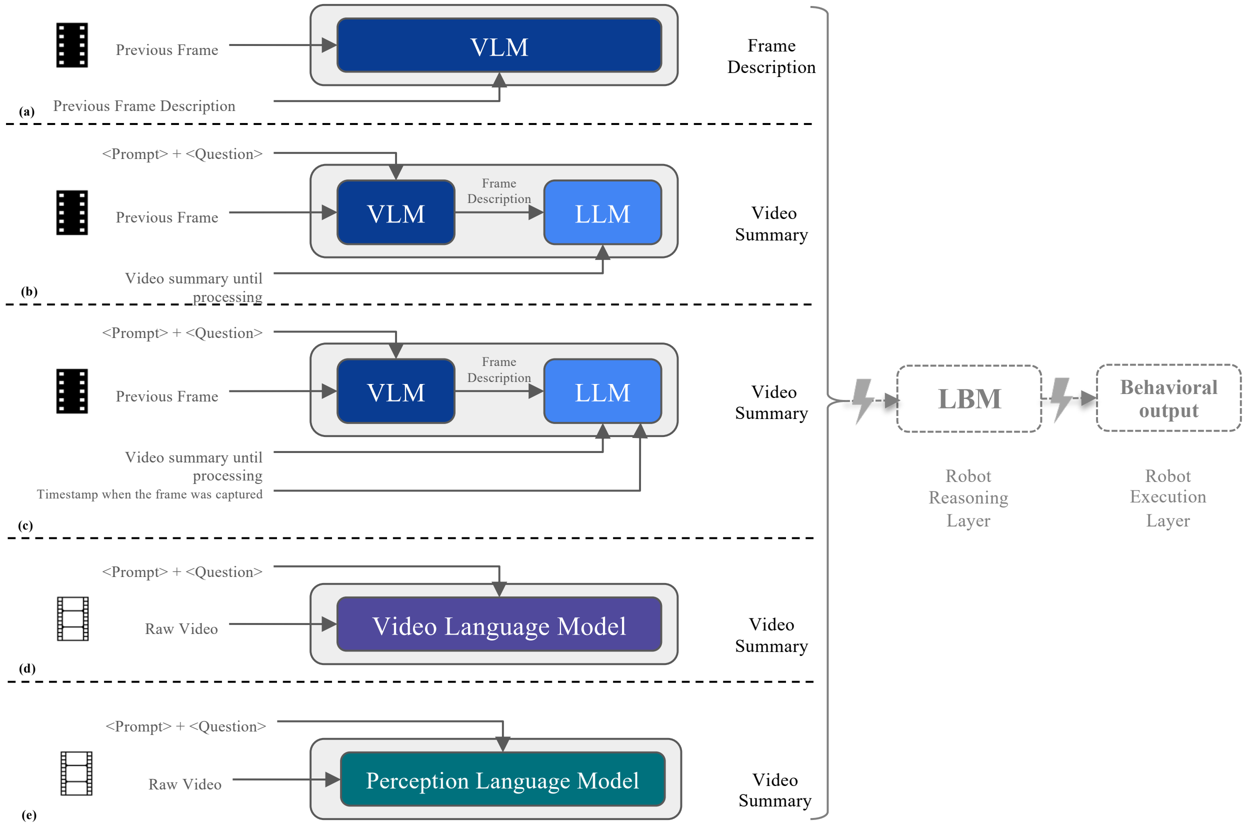

Figure 1. Experimental Settings

Temporal Blindness and Sequential Amnesia. Our experiments revealed that current pipelines lack the capacity to accurately represent the temporal progression of social interactions. In Experiments (a) and (b), frame-level analysis and incremental summarization produced nearly identical captions across sequences, treating each input as an isolated event rather than as part of an unfolding trajectory. Similarly, the end-to-end video–text approach in Experiment (d) often compressed extended action sequences into generic accounts (e.g., “waving hands”) while ignoring key transitional details (e.g., “standing”). These outcomes reflect a broader weakness in encoding temporality: models perceive snapshots but not flows. In HRI contexts, this limitation directly undermines the robot’s ability to recognize social cues, anticipate turn-taking, and maintain synchrony, which are essential for interactional fluency.

Action Representation. Action recognition across the experiments demonstrated partial success but revealed substantial gaps in behavioral fidelity (e.g., Experiment (e)). Models consistently identified broad categories such as “waving” or “holding a cup,” but often omitted micro-actions or conflated distinct movements into a single label. For instance, object handover sequences (reach → stabilize → transfer → withdraw) were rarely captured in full, with outputs collapsing into simplified descriptors. This flattening of action detail matters profoundly in HRI: nuanced differences in gesture timing, posture, and sequencing frequently signal intent and affect. Robots that cannot register or respond to these subtleties risk misaligned or socially inappropriate behaviors, reducing their perceived sensitivity and trustworthiness in human-facing roles.

Identity Continuity. Identity-aware preprocessing embedded in the videos successfully detected and annotated individuals across frames; however, this information was inconsistently carried through by downstream models. In Experiment (c), annotated names were frequently replaced with “Unknown,” and even the PLM tested in Experiment (e) did not reliably reference identified individuals despite explicit cues. These discontinuities illustrate a breakdown in representational persistence across the pipeline. For HRI, the implications are significant: continuity of identity is foundational for personalization, rapport-building, and sustained trust. A robot that fails to recognize or misattributes its interlocutors risks undermining not only user confidence but also the relational bonds that give social interaction its depth and richness.

Redundancy and Plausibility Bias. A further diagnostic pattern was the prevalence of repetitive or plausibility-biased outputs. In Experiment (b), incremental summarization frequently produced redundant phrasings across successive frames, offering little narrative advancement. In other cases, particularly with PLM outputs in Experiment (e), descriptions were fluent but factually incomplete or inaccurate—for example, fabricating an action (“pouring into a bottle”) or overlooking a salient gesture. These tendencies suggest that models optimize for statistical plausibility and linguistic fluency rather than behavioral fidelity. In HRI settings, such outputs risk eroding user trust: interactions that feel repetitive can quickly become disengaging, while plausible but inaccurate accounts create a perception of inattentiveness or even social incompetence.

The diagnostic experiments reveal that the upstream preparation of inputs is the central bottleneck in achieving socially grounded behavior. Identity consistency broke down not because recognition models failed at detection, but because upstream pipelines could not preserve those signals across representational layers, leaving robots unable to maintain personalization over time. Action accuracy was weakened when upstream processing collapsed fine-grained gestures into coarse categories, stripping away nuances that convey intent in social contexts. Temporal coherence faltered when frame-based and video-based inputs were not encoded with mechanisms to capture event order, resulting in static or repetitive narratives rather than continuous interactional flows. Finally, narrative utility was degraded when upstream architectures optimized for plausible text rather than socially meaningful detail, yielding outputs that were redundant or socially irrelevant. Taken together, these findings indicate that current pipelines fragment and distort socially grounded cues before they reach behavioral models. For HRI, this means that robots are fed inputs that are too brittle to support adaptive, context-aware engagement.

5 Limitations and Future Research

This study should be read as an exploratory probe rather than a definitive solution for training socially intelligent robots. Our objective was to identify weaknesses in upstream input pipelines and critique how current models handle socially grounded signals, rather than delivering benchmark-optimized architectures. This diagnostic orientation introduces several limitations that future work should address.

First, our experiments were conducted under controlled laboratory conditions using short, scripted scenarios. While these clips offered ecological plausibility, they inevitably lacked the richness, ambiguity, and variability of spontaneous real-world interactions where social robots are expected to operate. Second, the evaluation was qualitative and guided by our rubric, which assessed identity consistency, action accuracy, temporal coherence, and narrative utility, rather than relying on standardized community benchmarks. Although this was intentional, it limits comparability with broader machine learning evaluations. Third, some failures likely reflect design decisions in our pipeline itself, including choices in preprocessing, model prompting, and integration strategies, which complicates causal attribution. Finally, our analysis covered a narrow slice of available models; additional baselines, fine-tuned variants, or hybrid architectures may reveal different strengths and limitations.

These constraints mean that our findings should be understood as diagnostic signals. They highlight that the central challenge for socially intelligent robotics lies upstream—in how raw perceptual inputs are curated, contextualized, and transmitted into learning architectures. Future research must therefore shift from piecemeal fine-tuning toward unified pipelines that embed temporal order, preserve identity, and distill socially relevant detail across modalities. Perception–language models offer a promising pathway, but further exploration is necessary to investigate alternative architectures, richer datasets, and evaluation protocols that more accurately reflect the dynamics of HRI. Building on this orientation, we identify five sociotechnical domains where assumptions must be challenged (Table 2). Together, these domains form the foundation of a research agenda aimed at robust upstream architectures.

Table 2. PLM-Centered Research Agenda for Socially Intelligent Robotics

Domain	Focus Area	Research Direction	Implications for HRI
Temporal Grounding	*Spatiotemporal Attention*	Design pipelines that preserve event order by embedding temporal attention and anchoring frames to sequential shifts and visual cues.	Enables robots to anticipate turns, segment actions accurately, and sustain interactional rhythm.
	Incremental Context Integration	Incorporate hierarchical memory so inputs are updated and retained across frames and turns without temporal drift.	Supports coherent multi-turn interaction, fluid turn-taking, and continuity in social narratives.
Contextual Continuity	*Dynamic Context Windowing*	Implement saliency-weighted filtering to prioritize socially relevant cues while discarding noise across interaction streams.	Prevents context loss and reinforces discourse consistency in extended interactions.
	Episodic Memory Embedding	Build pipelines with long-term memory structures that preserve context across sessions and episodes.	Enables longitudinal tracking of user goals and rapport over repeated encounters.
Personalization Modeling	*Identity-Aware Encoding*	Maintain identity signals across modalities by fusing facial, gestural, and linguistic embeddings in the input stage.	Facilitates personalization and trust by ensuring robots consistently recognize and adapt to individuals.
	Semantic Matching via Embeddings	Apply embedding-based resolution methods to disambiguate identities in noisy or multi-user environments.	Improves reliability of personalization and alignment of responses with the correct interlocutor.
Narrative Utility	*Prompt Adaptation Strategies*	Structure pipelines to foreground salient social details while minimizing redundancy and generic phrasing.	Produces socially useful outputs that enrich conversations and support decision-making.
	Feedback-Driven Refinement	Integrate feedback loops during input preparation to iteratively refine relevance and reduce error propagation.	Aligns robot narratives with user expectations, enhancing perceived attentiveness and social naturalness.
Multimodal Coherence	*Cross-Modal Fusion Architecture*	Develop upstream fusion layers that unify gaze, gesture, prosody, and language before downstream modeling.	Strengthens interpretation of intent in complex, emotionally nuanced, or group interactions.
	Concise Output Optimization	Apply compression techniques that preserve key social signals while reducing input redundancy.	Improves clarity, reduces cognitive load, and enhances robot comprehensibility in real time.

Temporal Grounding. A persistent weakness in current pipelines is their inability to encode spatiotemporal attention, treating events as disjointed snapshots rather than evolving trajectories. Upstream architectures must incorporate temporal attention and incremental context integration to process sequences with causal progression and interactional rhythm. This matters because social robots need to anticipate turns, align gestures with speech, and sustain joint attention. Evaluation should test whether robots segment actions into meaningful boundaries (e.g., reach → grasp → withdraw), respond at socially appropriate moments, and adjust pacing in collaborative tasks.

Contextual Continuity. Most input systems reset context at every turn, lacking mechanisms for dynamic windowing or episodic memory. Future designs must embed memory architectures that retain salient cues across frames, turns, and sessions. In HRI, such continuity underpins trust and long-term engagement—for example, remembering a student's learning trajectory or a patient's daily routine. Evaluation should therefore extend beyond brief interactions to examine whether robots maintain discourse consistency and goal alignment over time.

Identity Awareness. Identity signals frequently dissipate as data move through the pipeline. To address this, upstream architectures must incorporate identity-aware encoding and semantic matching that preserve user-specific traces across modalities. In practice, continuity of recognition is essential for personalization: a robot that forgets or misattributes identities risks undermining relational trust. Testing should include multi-user scenarios where robots must consistently disambiguate individuals, adapt their responses to the correct partner, and maintain user histories across repeated encounters.

Narrative Utility. Redundancy and shallow plausibility reveal pipelines that prioritize fluency over utility. Upstream processing must employ adaptive prompting and feedback-driven refinement to foreground socially relevant cues and suppress trivial repetition. For HRI, this ensures robots enrich interactions rather than burdening them with monotony. Evaluation should focus on whether outputs advance the encounter—for instance, by adding contextually meaningful detail, supporting decision-making, or framing feedback appropriately.

Multimodal Coherence. Finally, upstream architectures must unify heterogeneous signals through cross-modal fusion, capturing how meaning emerges from the alignment of gaze, gesture, posture, and prosody. A single modality rarely carries social intent; it depends on its integration. Robots that process modalities in isolation miss the fluency that makes interaction socially credible. Evaluation should place robots in multimodal conflict scenarios (e.g., verbal "yes" paired with a head shake) to test whether they prioritize socially intended meaning through coherent fusion.

Future architectures can be tested through a mix of controlled simulations, lab-based interaction studies, and in-the-wild deployments. Controlled simulations allow systematic manipulation of input variables to assess how pipelines handle temporal order, identity persistence, and multimodal fusion under varied conditions. Lab-based HRI experiments offer ecologically valid yet manageable settings where user trust, engagement, and interactional fluency can be observed in real-time. Finally, field trials in naturalistic environments—such as classrooms, clinics, or public spaces—are essential to evaluate how upstream designs generalize to the ambiguity, diversity, and unpredictability of everyday social encounters. Together, these methods ensure that upstream architectures are not only technically robust but socially credible in practice.

6 Conclusion

This study investigated how upstream pipelines can be reimagined to furnish socially intelligent robots with inputs that preserve the richness of human interaction. Using a clinical problematization lens, we treated experimental breakdowns as diagnostic signals that exposed blind spots in current pipelines. Our findings show that essential dimensions of social interaction—temporal flow, action detail, identity continuity, and narrative utility—are often lost before reaching downstream behavioral and reasoning models. We argue that future progress lies not in scaling architectures but in designing upstream pipelines that embed spatiotemporal reasoning, contextual continuity, and multimodal coherence, enabling robots to sustain coherent, adaptive, and socially meaningful engagement.

7 References

Alvesson, M., & Sandberg, J. (2011). Generating research questions through problematization. *Academy of Management Review, 36*(2), 247–271.

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), 1–12.

Bachelard, G. (1938). *The Formation of the Scientific Mind* (M. M. Jones, Trans.). Clinamen Press.

Bartalesi, V., Lenzi, E., & De Martino, C. (2024). Using large language models to create narrative events. *PeerJ Computer Science, 10*, e2242.

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. https://arxiv.org/abs/2412.16698

Breazeal, C. (2003). Toward sociable robots. *Robotics and autonomous systems*, *42*(3-4), 167-175.

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*, 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 preprint. https://arxiv.org/abs/2502.02046

Duncan, J. A., Alambeigi, F., & Pryor, M. W. (2024). A Survey of Multimodal Perception Methods for Human–Robot Interaction in Social Environments. *ACM Transactions on Human-Robot Interaction*, *13*(4), 1–50.

Fang, J., Zhou, W., Xiong, L., & Song, G. (2025). The Role of Social Robots in Alleviating Anxiety and Enhancing Mental Well-Being. *International Journal of Mental Health and Addiction*, 1–34.

Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., & Schwager, M. (2025). Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, *44*(5), 701-739

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.

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.

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.

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*, 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.

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), 1–36.

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.

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.

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

LLaVA Team. (2024). *LLaVA-NeXT-Video-34B-hf* [Model]. Hugging Face.

LLaVA Team. (2024). *LLaVA-1.5-7B* [Model]. Hugging Face.

Meta. (2024). *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*, 1–22.

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*, 1–14.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730–27744.

Pool, J., Indulska, M., & Sadiq, S. (2024). Large language models and generative AI in telehealth: a responsible use lens. *Journal of the American Medical Informatics Association*, *31*(9), 2125–2136.

Sabo, R., Beňuš, Š., Kevická, V., Trnka, M., Rusko, M., Darjaa, S., & Kejriwal, J. (2024). Towards the Use of Social Robot Furhat and Generative AI in Testing Cognitive Abilities. *Human Affairs, 34*(2), 224-243.

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*.

Salinas-Martínez, Á. G., Cunillé-Rodríguez, J., Aquino-López, E., & García-Moreno, A. I. (2024). Multimodal Human–Robot Interaction Using Gestures and Speech: A Case Study for Printed Circuit Board Manufacturing. *Journal of Manufacturing and Materials Processing*, *8*(6), 274.

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.

Sapkota, R., Cao, Y., Roumeliotis, K. I., & Karkee, M. (2025). Vision-language-action models: Concepts, progress, applications and challenges. arXiv preprint arXiv:2505.04769.

Stipancic, T., Jerbic, B., & Curkovic, P. (2016). A context-aware approach in realization of socially intelligent industrial robots. *Robotics and computer-integrated manufacturing*, *37*, 79-89

Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, & Farah 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.Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, and Takeo Igarashi (Eds.), ACM, New York, NY, 1–5.

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), 103220.

Ulrich, W. (1994). *Critical heuristics of social planning: A new approach to practical philosophy*. Wiley. (Original work published 1983)

Wang, W., Wang, R., Mao, L., & Min, B. C. (2023, October). Navistar: Socially aware robot navigation with hybrid spatio-temporal graph transformer and preference learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 11348–11355). IEEE.

Wirtz, J., & Stock-Homburg, R. (2025). Generative AI Meets Service Robots. *Journal of Service Research,* forthcoming.

Yang, Y., Zhou, T., Li, K., Tao, D., Li, L., Shen, L., & Shi, Y. (2024). Embodied multi-modal agent trained by an LLM from a parallel text world. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

You, S., & Robert Jr, L. P. (2024). Trusting and working with robots: A relational demography theory of preference for robotic over human co-workers. *MIS Quarterly*, *48*(4), 1297-1330.

Zolanvari, S. M. J., Taheri, A., Meghdari, A. F., & Alemi, M. (2025). Social robot recognition of human social movements: teaching a robot social etiquette using cognitive architecture. *International Journal of Social Robotics*, 1–26.