From command to symbiosis: Integration paradigms and conceptual frameworks for seamless multimodal Human-Robot Interaction

Jiangke WU

School of Informatics, Computing and Cyber Systems

Northern Arizona University, Flagstaff, Arizona, United States

jw2966@nau.edu

Abstract: Human-Robot Interaction (HRI) is currently undergoing a paradigm shift from discrete command-based control to seamless, symbiotic communication. However, achieving this symbiosis requires overcoming significant challenges in temporal alignment and computational efficiency, particularly when processing conflicting multimodal signals. This paper presents a hybrid study combining a semi-systematic literature review (2020–2025) with a novel conceptual framework. The study critically evaluates the architectural evolution from standard Transformers to linear-complexity State Space Models (SSMs), such as Mamba, highlighting the trade-offs between long-term semantic context and real-time responsiveness. Furthermore, the Multimodal Perception-Driven Decision-Making (MPDDM) framework is introduced as a conceptual model designed to resolve conflicts between explicit commands and implicit physiological cues via a dynamic confidence-weighting mechanism. To address the limitations of traditional latency-based benchmarks, a new set of evaluation metrics, specifically Modality-Specific Fluidity (MSF), is proposed to quantify the smoothness of interaction. Finally, recent integration paradigms are categorized into modular and end-to-end Vision-Language-Action (VLA) models, offering a critical synthesis of their respective safety and efficiency profiles. This work provides a roadmap for developing verifiable, low-latency HRI systems capable of operating in dynamic, unstructured environments.

Keywords: Human-Robot Interaction, Seamless symbiosis, Multimodal fusion, State space models, Interaction repair.

1. Introduction

Human-Robot Interaction (HRI) is ushering in a major turning point. Traditionally, robots have been mainly regarded as passive tools, relying on clear instructions issued by humans through keyboards, joysticks or preset scripts. This mode is remarkably effective in the factory environment, but its limitations are revealed when robots enter social places such as homes, hospitals and public spaces (Thomaz, Hoffman & Cakmak, 2016; Zhao, Gangaraju & Yuan, 2025). These scenarios require robots that can assist the elderly (Mois & Beer, 2020), educate children (Belpaeme et al., 2018), and build long-term relationships (Van Straten, Peter & Kühne, 2020; Zhao & McEwen, 2025).

The current research is turning to the "seamless symbiosis" mode. This new type of communication is no longer limited to the transmission of simple instructions in classical information theory (Shannon & Weaver, 1949), but into a continuous two-way process aimed at fostering the mutual understanding of things, as shown in Figure 1. In this proposed model, the Human (Red) and Robot (Blue) operate within a shared physical context. Solid arrows represent explicit interactions (e.g., speech, action), while dashed lines indicate implicit cues (e.g., gaze, physiological signals) and feedback loops. The "seamless" here not only refers to the integration of sensor data at the technical level, but also requires a smooth interactive experience. To achieve this goal, the system needs to perform well in three core areas: timing coordination, semantic matching and error correction (Ong, Seet & Sim, 2008). The system also needs to have the ability to handle a variety of input types (e.g., sound, visual clues and physiological data) (Gupta et al., 2025; Yoshida et al., 2025). By integrating these signals, the system can clearly present the user's intentions and emotional states.

However, achieving this objective presents significant challenges due to the complexity of human multimodal communication and the inherent ambiguity of intent. Without the contextual grounding provided by gestures or gaze, deictic commands such as "pick up that" are prone to

https://doi.org/10.33436/v36i1y202607

misinterpretation. Furthermore, multisensory signals are rarely perfectly synchronous; for instance, individuals typically fixate on a target object prior to initiating verbal communication. Failure to resolve these temporal misalignments or identify gesture onsets inevitably leads to interaction failures (Pramanick & Rossi, 2024).

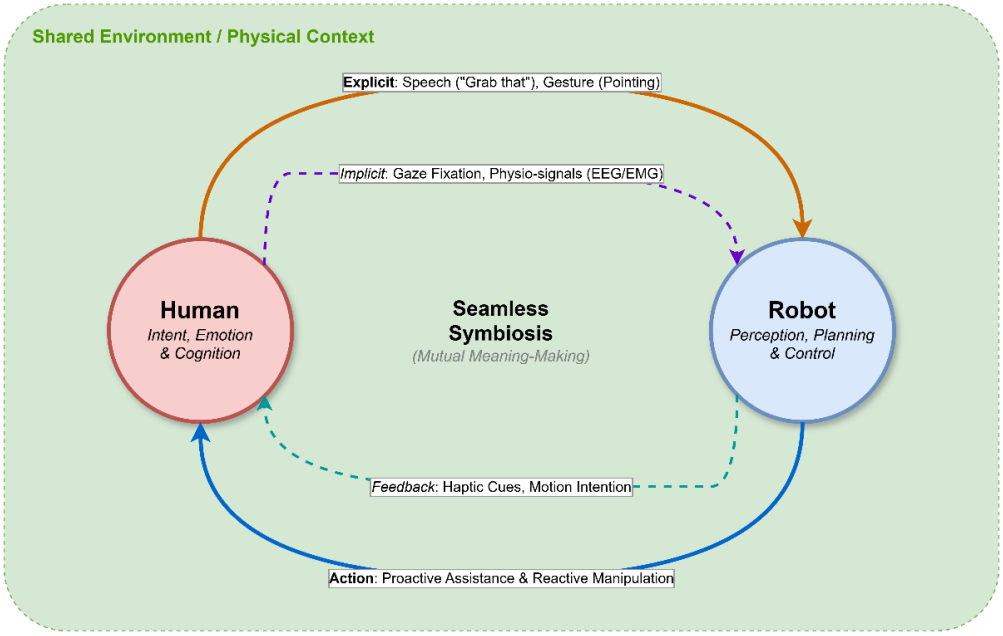

Figure 1. The cycle of seamless symbiosis (Source: author).

To bridge the gap between perception and actuation in dynamic environments, this article presents a hybrid study combining a critical literature review (2020‑2025) with a novel conceptual framework. The primary objective is to evaluate emerging computational architectures, specifically contrasting Transformers with State Space Models (Mamba), while proposing the Multimodal Perception-Driven Decision-Making (MPDDM) framework to resolve sensory conflicts. Unlike recent surveys that focus largely on the accuracy of deep learning perception modules (Zhao, Gangaraju & Yuan, 2025), this work differentiates itself by prioritizing interaction fluidity, temporal alignment, and computational latency as the core metrics for success. Consequently, the following sections explicitly distinguish between established methodologies found in the literature and the author's proposed theoretical contributions, providing a roadmap for verifiable, low-latency HRI systems.

1.1. Contributions

To clarify the scope and novelty of this work, this study is explicitly defined as a hybrid inquiry that combines a literature review with conceptual proposals. Specifically, this paper delivers three distinct contributions:

- **A Targeted Review of Integration Architectures:** This paper presents a comparative analysis of recent integration technologies (2024–2025), specifically contrasting the context-aware capabilities of Transformers with the linear computational efficiency of State Space Models (SSMs/Mamba) in the context of HRI.
- **The MPDDM Conceptual Framework:** The Multimodal Perception-Driven Decision-Making (MPDDM) framework is introduced. This theoretical model is designed to resolve conflicts between explicit commands (e.g., speech) and implicit cues (e.g., physiological signals) using a confidence-weighting mechanism.
- **Novel Fluidity Metrics:** A new set of evaluation metrics, specifically Modality-Specific Fluidity (MSF) and Physiological Synchrony, is proposed. These serve as a theoretical baseline to move beyond standard latency benchmarks and measure the "seamlessness" of human-robot symbiosis, pending future empirical validation.

Differentiation from Existing Surveys: While recent reviews have extensively covered deep learning in HRI (Zhao, Gangaraju & Yuan, 2025), they primarily focus on accuracy maximization within static Transformer architectures. This study differentiates itself by: (1) investigating the emerging computational shift toward linear-complexity models (specifically Mamba/SSMs) for real-time edge processing, and (2) shifting the evaluation focus from traditional "command recognition rates" to "interaction fluidity" and "repair mechanisms" (Spitale et al., 2024). This creates a unique perspective on handling temporal misalignments in dynamic environments.

The rest of this paper is organized as follows. Section 2 details the semi-systematic review methodology and differentiates this study from existing surveys. Section 3 expounds the MPDDM theoretical framework. Section 4 discusses the main computing models, specifically contrasting Transformers with Mamba architectures. Section 5 categorizes integration methods into modular and end-to-end paradigms. Section 6 analyzes the impact of temporal dynamics on these strategies. Section 7 explores the societal impact, proposes new assessment criteria, and discusses real-world deployment challenges. Finally, Section 8 concludes the full paper.

2. Related work and survey methodology

2.1. Review strategy and scope

To effectively bridge the gap between technical architectures and interaction design, this study adopts a semi-systematic review approach. Unlike strict systematic reviews that prioritize quantitative meta-analysis, this approach allows for the critical synthesis of diverse literature across computer vision, robotics, and cognitive science. The literature search was conducted using major databases, including IEEE Xplore and the ACM Digital Library. Papers published between 2020 and 2025 were screened, with priority given to research containing keywords such as "Multimodal HRI", "Vision-Language-Action (VLA) Models", and "Interaction Repair".

2.2. Inclusion and exclusion criteria

To ensure the scientific rigor and relevance of this review, specific selection criteria were applied to screen literature retrieved from major databases (IEEE Xplore, ACM Digital Library).

Inclusion Criteria:

- **Publication Type:** Peer-reviewed journal articles and conference proceedings published between January 2020 and early 2025.
- **Domain Relevance:** Studies explicitly addressing "Multimodal Integration", "Vision-Language-Action (VLA) Models", or "Interaction Repair" within the context of Human-Robot Interaction.
- **Validation Level:** Research that provides verifiable architectural implementations (e.g., Transformer or SSM-based) or empirical user studies.

Exclusion Criteria:

- **Unimodal Studies:** Research focusing solely on a single modality (e.g., voice-only assistants) without sensor fusion mechanisms.
- **Non-Peer-Reviewed Material:** Extended abstracts, posters, and pre-prints lacking significant community citation.
- **Language:** Publications not written in English.

2.3. Comparison with existing surveys

To highlight the unique contribution of this work, Table 1 compares this study with recent related surveys. While prior works focus heavily on deep learning accuracy, this paper uniquely addresses the computational shift toward linear complexity (Mamba) and the metrics of interaction fluidity.

Table 1. Comparison with recent surveys in multimodal HRI

Survey	**Primary Focus**	**Key Architecture Analyzed**	**Addressed Gap**
(Zhao, Gangaraju & Yuan, 2025)	Perception-Driven Decision Making	Deep Learning (General)	Comprehensive taxonomy of perception algorithms.
(Chaabene et al., 2025)	Healthcare Data Fusion	General Fusion Techniques	Methods for fusing heterogeneous medical data.
(Teoh et al., 2024)	Healthcare Data Fusion	Traditional ML & Fusion	Application-specific (Medical) data handling.
This Work	**Seamless Symbiosis & Fluidity**	**Hybrid (Transformer + Mamba)**	**Temporal alignment, interaction repair, and linear-complexity processing.**

3. Theoretical framework

3.1. Multimodal Perception-Driven Decision-Making (MPDDM)

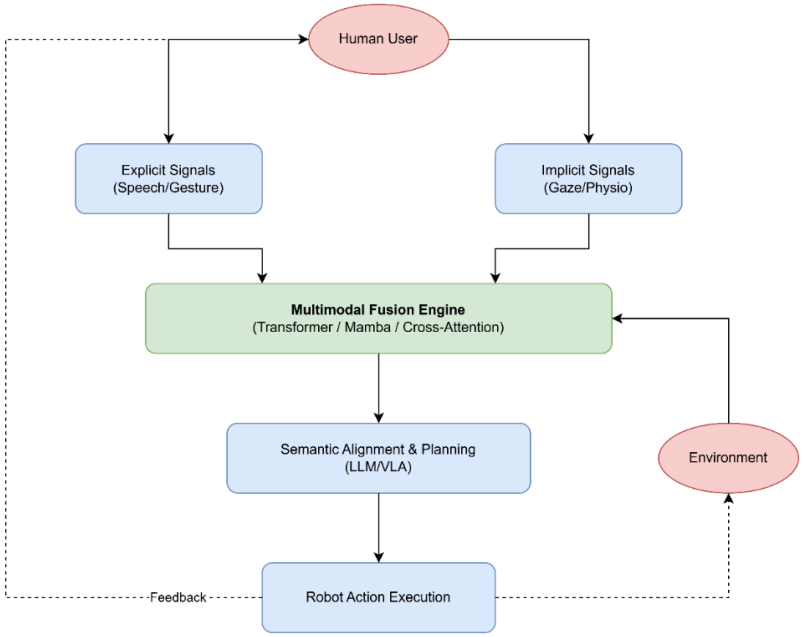

Figure 2. The Multimodal Perception-Driven Decision-Making (MPDDM) framework (Source: author)

Multimodal perception is not only a way to collect sensor data, but also a way for robots to understand complex environments and human states. Recent research refers to this as Multimodal Perception-Driven Decision-Making (MPDDM) (Teoh et al., 2024). As illustrated in **Error! Reference source not found.**, the framework is structured to process explicit and implicit signals in parallel, feeding them into a unified fusion engine. The architecture delineates the flow from Input Acquisition to the Multimodal Fusion Engine (highlighted in Green), and finally to Action Execution (Blue). This structure emphasizes how the fusion engine integrates conflicting signals to inform the semantic planning layer, subject to environmental constraints. The main idea of MPDDM is that successful HRI depends on the robot's ability to see more than just physical actions. It must also understand mental states, intentions (Adebayo, McLoone & Dessing, 2024), and hidden preferences (Dennler, Nikolaidis & Matarić, 2025). It is important to note that human signals do not happen at the same time. For example, a look at one moment might explain a spoken command that follows it. Therefore, the standard decision-making model is extended to incorporate time alignment.

3.1.1. Confidence weighting and conflict resolution

To address the ambiguity inherent in conflicting multimodal signals, such as when a user's verbal command contradicts their physiological stress markers, the MPDDM framework employs a weighted arbitration mechanism. It is important to clarify that the following formulation is a conceptual model intended to illustrate the logic of decision-making, rather than a formally derived mathematical theorem.

The robot selects the optimal action A^* from the available action space $\mathcal{A}$ by maximizing a joint confidence function, subject to environmental safety constraints:

$$A^* = *arg\,max_{a\in\mathcal{A}} \left[\lambda \cdot S_{explicit}(a) + (1-\lambda) \cdot S_{implicit}(a) \right] \quad (1)$$

$$\text{subject to} \quad C_{env}(a) = \text{True} \quad (2)$$

Where:

- A^*: The optimal action selected for execution.
- a: A candidate action within the robot's action space $\mathcal{A}$.
- λ: The dynamic weight factor ($\lambda \in [0,1]$). This value is inversely proportional to the entropy of the physiological signal; high entropy (uncertainty) in implicit cues results in a higher λ, shifting reliance toward explicit commands.
- $S_{explicit}(a)$: The normalized confidence score [0,1] derived from explicit signals (e.g., speech, gesture).
- $S_{implicit}(a)$: The normalized confidence score [0,1] derived from implicit cues (e.g., gaze, EEG/EMG).
- $C_{env}(a)$: A boolean function representing environmental constraints (e.g., collision avoidance), where $C_{env}(a) = \text{True}$ indicates the action is safe.

3.2. Taxonomy of modalities and interaction bandwidth

Interaction modalities are categorized according to the mode of information flow and the degree of user participation to address the integration problem. Table 2 provides the detailed information of the framework. This classification distinguishes not only based on the physical form of the modality but also considers its semantic richness and temporal dynamics.

Table 2. Taxonomy of interaction modalities in HRI

Category	Specific Forms	Data Features	Role in HRI
Explicit Interaction (Active)	Speech commands, Hand movements, Touch input	Distinct steps, Clear symbols, Rich meaning	Users give clear orders to control the robot.
Implicit Perception (Passive)	Eye contact, Facial expressions, Body language, EEG/EMG	Continuous, Complex, Unclear, Hard to interpret	The robot guesses attention, emotions, and mental effort.
Contextual Awareness	Recognizing objects, Distance between items, Background noise	Related to space, Organized	Sets physical limits and explains the background.

4. Computational backbones: From attention to state spaces

In the field of data integration of HRI, the key is to strike a balance between understanding the context and saving computing power. This section discusses two main models used by modern

HRI systems: Transformer's global attention mechanism and new SSMs with linear complexity. Table 3 also provides a comparative summary of the main frameworks discussed in this review.

Table 3. Comparative analysis of integration architectures in HRI

Architecture	Core Mechanism	Complexity	Key Strength	Limitation
M3ET (Zhang et al., 2025)	Hybrid Transformer + Mamba (SSM)	$\mathcal{O}(L)$ (Linear)	Handling long-sequence physiological / audio data on edge devices	May lose fine-grained spatial attention compared to pure ViTs.
TransforMerger (Vanc & Stepanova, 2025)	Probabilistic Embeddings	$\mathcal{O}(L^2)$	Resolving ambiguity in noisy inputs via joint probability modeling	Computationally expensive; harder to scale to continuous streams.
GhostShell (Gong et al., 2025)	Streaming XML Token Parser	Stream-based	Minimizing latency; "Acting-while-thinking" capability	Risk of executing premature actions if LLM hallucinates mid-stream.
SemanticVLA (Li et al., 2025)	End-to-End VLA with Sparsification	Variable	Generalizable world knowledge; semantic-aware filtering	High memory footprint; difficult to interpret "black box" decisions.

4.1. The dominance of Transformer architectures

Because its self-attention mechanism can connect distant parts of the sequence (Vaswani et al., 2017), the Transformer model has become the preferred scheme for multimodal processing. This technology originated from early technologies: such as ResNet for processing static images (He et al., 2016) and BERT for interpreting language (Devlin et al., 2019). Bottom-Up and Top-Down attention mechanism (Anderson et al., 2018) and Swin Transformer (Liu et al., 2021) are two ways to associate original pixels with the meaning of the visual world. This ability is crucial to the long-distance association in human-robot interaction, for example, allowing the system to associate the pronoun "it" in the current sentence with the object of the premise of the three-round dialogue.

However, standard self-attention scales quadratically with sequence length ($\mathcal{O}(L^2)$).

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Due to the high computing cost, the real-time robot system takes a long time to respond. This is especially true when the robot processes fast-moving physiological data or continuous video streams on devices with limited power.

Critical Synthesis: While Transformer architectures have solved the problem of long-term semantic dependency, they face a fundamental computational bottleneck due to quadratic complexity in robotics. What is known is that global attention is indispensable for grounding complex language commands. However, it remains uncertain how to adapt this discrete token-based mechanism for continuous, high-frequency sensory streams (e.g., 1kHz tactile data) without prohibitive computational costs. The open challenge lies in developing "linear-attention" variants that retain semantic depth while meeting the millisecond-level latency constraints of physical control loops.

4.2. The rise of linear complexity: Mamba and SSMs

To address the slow response times of Transformers, State Space Models (SSMs) like Mamba have appeared as a strong alternative (Gu & Dao, 2024). Mamba uses a selective scan

mechanism. This mechanism can be turned into a step-by-step process defined by the discretized State Space Model (SSM) equations:

$$h_t = \overline{\mathbf{A}} h_{t-1} + \overline{\mathbf{B}} x_t \tag{4}$$

$$y_t = \mathbf{C} h_t \tag{5}$$

Where:

- h_t : The hidden state vector at time step t , compressing the historical context.
- x_t : The input signal vector (e.g., sensor token) at time t .
- y_t : The output response vector.
- $\overline{\mathbf{A}}, \overline{\mathbf{B}}$: The discretized state transition and input matrices, derived from the continuous system parameters using the zero-order hold (ZOH) method.
- $\mathbf{C}$: The projection matrix mapping the hidden state to the output.

The M3ET (Mamba-Enhanced Transformer) architecture, visually detailed in Figure 3 demonstrates this hybrid approach (Zhang et al., 2025). The diagram distinguishes between data flow (solid arrows) and control flow (dashed lines). Specifically, the Transformer Block (highlighted in Blue) is designed to handle sparse semantic retrieval, focusing on specific local details, while the Mamba Block (Orange) processes the high-throughput linear sensor stream to compress long sequences of data. Additionally, the dashed 'Selective Gating' line illustrates how semantic context modulates the state-space compression mechanism. The processing speed of M3ET is 2.3 times that of the Transformer model alone, which is due to the hybrid architecture.

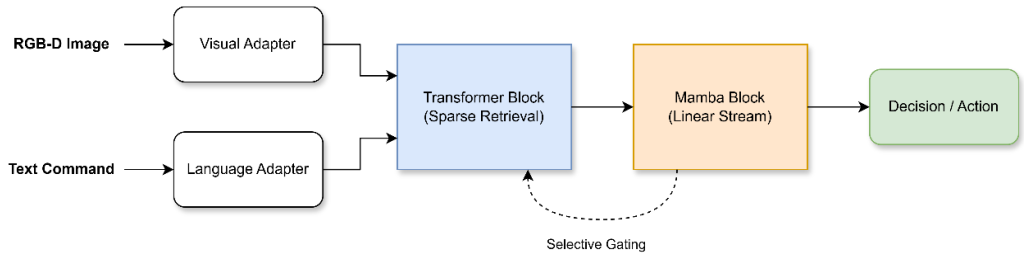

Figure 3. The M3ET architecture (Source: Zhang et al., 2025)

Critical Synthesis: While Mamba offers exceptional compression efficiency, its recurrent structure is prone to information loss or 'forgetting.' The algorithm compresses historical data into a fixed-size state. Although this allows for the efficient processing of real-time sensor data, retrieving specific past information, such as a user's coffee preference from a week ago, remains challenging. Conversely, the Transformer architecture excels at capturing global dependencies through its attention mechanism. Given this complementarity, a hybrid design, M3ET, is proposed. This architecture employs the Mamba module to compress continuous, high-throughput data streams (e.g., visual and physiological signals), while utilizing a sparse Transformer module to extract long-term semantic dependencies. This strategy effectively balances the requirement for real-time responsiveness with the necessity of long-term memory.

5. Integration paradigms: From perception to action

Based on these computing models, two primary integration methods are identified. The first category is the modular approach, which achieves integration through the fusion of specific signals. The second category employs an end-to-end process to directly map perception to action.

5.1. Modular approaches: Handling uncertainty

The modular system separates perception from planning and focuses on reliably integrating signals. Foundation-model Assisted Multi-modal Human-Robot Interaction (FAM-HRI) (Lai et al.,

2025) and TransforMerger are the two most important examples in this group. TransforMerger (Vanc & Stepanova, 2025) adopts probability embedding technology. When the data is not clear, it will not make a strict judgment, but expresses the joint probability of the user's intention by integrating multimodal likelihoods:

$$P(I \mid V, A) \propto P(V \mid I) \cdot P(A \mid I) \cdot P(I) \tag{6}$$

Where:

- I : The specific user intention being evaluated (e.g., "Handover").
- V, A: The observed feature vectors from Visual (V) and Audio (A) modalities.
- $P(I \mid V, A)$: The posterior probability of the intention given the observed signals.
- $P(V \mid I)$ and $P(A \mid I)$: The likelihood functions representing how probable the observed signals are for a given intention.
- $P(I)$: The prior probability of the intention, based on historical context or task constraints.

This method successfully clears up confusion when users point at things. However, it often struggles to handle the complex thinking needed for unpredictable places.

Critical Synthesis: Modular approaches offer the distinct advantage of explainability and safety, crucial for medical and industrial applications where "black box" errors are unacceptable. However, the primary limitation is the "information bottleneck" between isolated modules, where rich context often gets lost during data handover. Future research must address how to maintain the transparency of modular systems while approximating the "cross-modal intuition" found in end-to-end models, possibly through differentiable interfaces that allow gradient flow across module boundaries.

5.2. End-to-End streaming paradigms

The field is increasingly moving towards Vision-Language-Action (VLA) models, such as RT-2 (Zitkovich et al., 2023) and SemanticVLA (Li et al., 2025). These models learn to connect raw inputs directly to control actions. They use training on a massive scale from the internet to gain general knowledge about the world.

The GhostShell framework introduces "Stream Programming" (Gong et al., 2025) to narrow the gap between advanced reasoning and real-time control. This streaming mechanism is compared with traditional pipelines in Figure 4. The top row depicts the latency bottleneck in traditional "Perception-Planning-Action" pipelines. Conversely, the bottom row illustrates the concurrent execution flow, where Token Generation (highlighted in Orange) and XML Parsing (Blue) overlap with Motor Execution (Green), significantly reducing the total system delay (Δt). While traditional methods usually make a complete plan before performing any operation, GhostShell allows the robot to move synchronously during the planning process through the streaming XML function token parser. The Large Language Model (LLM) is based on GPT-3 (Brown et al., 2020) and other small sample learners, which can instantly convert words into function calls. This mechanism realizes the ability of "thinking while acting", making the interaction process smoother and significantly shortening the time required for robots to respond to human instructions.

The trade-off between security protection and latency: Although streaming execution can reduce latency, it also greatly weakens security. Before fully understanding the overall situation, the robot may issue wrong or harmful instructions, such as "throwing". Recently, the system has introduced a forward-looking security protection mechanism to solve this problem. This lightweight safety tool will prevent the generation of 3 to 5 instructions and quickly compare the "Forbidden Action List" before sending instructions to the motor controller.

This calculated delay ($T_{delay} \approx 100$ ms) is generally acceptable for voice communication, as conversational turn-taking gaps often exceed this threshold. However, this latency poses a challenge for high-frequency haptic feedback loops, which typically require sub-millisecond

response times to ensure stability. Assuming a token generation rate of R_{gen} (e.g., 50 tokens/sec), a safety buffer size of B (e.g., 5 tokens) will introduce a latency delay T_{delay}:

$$T_{delay} = \frac{B}{R_{gen}} \approx 100 \text{ ms} \quad (7)$$

Where:

- T_{delay}: The additional latency introduced by the safety verification layer.
- B: The size of the look-ahead token buffer.
- R_{gen}: The generation speed of the Large Language Model.

This calculated delay ($T_{delay} \approx 100$ ms) is generally acceptable for voice communication, as conversational turn-taking gaps often exceed this threshold. However, this latency poses a challenge for high-frequency haptic feedback loops, which typically require sub-millisecond response times to ensure stability.

Critical Synthesis: VLA models represent a leap toward general-purpose robot intelligence, capable of "thinking while acting". Yet, a critical uncertainty remains: the trade-off between semantic reasoning and kinematic safety. Since probabilistic models can "hallucinate" actions, the pressing open question is how to architect "Verifiable VLA Control", systems that leverage the generalization of LLMs but are constrained by deterministic safety barriers (like GhostShell) to prevent catastrophic failures in the real world.

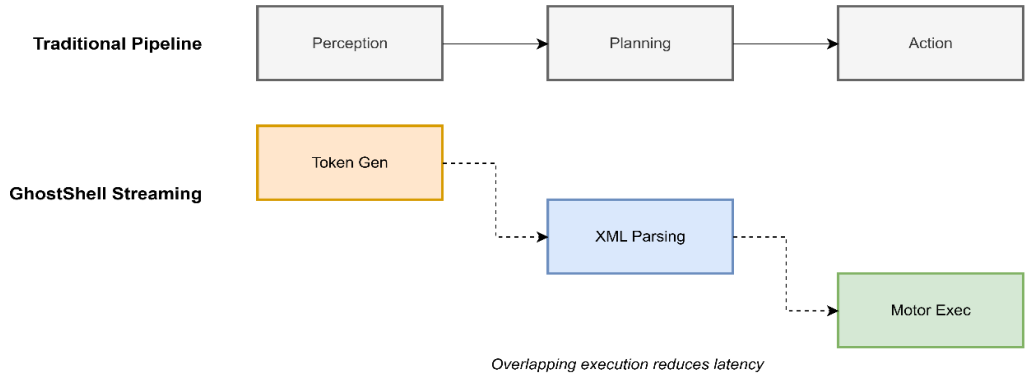

Figure 4. Comparison between traditional sequential execution and GhostShell's Streaming architecture (Source: Gong et al., 2025)

To synthesize the distinctions between these two dominant integration paradigms, Table 4 provides a comparative summary focusing on key deployment metrics.

Table 4. Comparative Summary of Integration Paradigms: Modular vs. End-to-End

Feature	**Modular Approaches (e.g., TransforMerger)**	**End-to-End VLA (e.g., SemanticVLA, GhostShell)**
Core Philosophy	Divide and Conquer: Separate perception, planning, and control modules.	Holistic Mapping: Direct mapping from raw pixels/tokens to robot actions.
Latency	High: Cumulative latency due to data serialization between modules.	Low: Optimized for streaming; reduced intermediate processing.
Interpretability	High: Errors can be traced to specific modules (e.g., "ASR failed").	Low: "Black box" nature makes it hard to diagnose specific failure points.
Robustness	Moderate: Prone to error propagation (cascading failures).	High: Joint optimization allows the model to recover from noisy inputs.
Deployment	Complex: Requires synchronizing multiple distinct software stacks.	Streamlined: Single model deployment, but high memory footprint.

6. Temporal dynamics and interaction ruptures

The art of timing makes seamless interaction possible. Delay or interference will destroy the immersive feeling and trust relationship.

6.1. Asynchrony and turn-taking

There is a significant difference in the speed of data processing in multimodal systems. For example, visual processing may only take milliseconds, but tactile feedback requires a microsecond response. Modern design meets this challenge by setting a differentiated time frame for different channels. At the same time, it is very important to accurately judge the appropriate timing of the conversation, which can prevent robots from disturbing humans (Bae & Bennett, 2025). Research shows that the waiting time (i.e. the response delay) affects the user's mood. However, there are differences in the appropriate thresholds for different senses. For example, a delay of more than 500 ms in a conversation will make people feel slow, while eye contact needs to be completed within 200 ms to keep both parties focused. On the other hand, if the robot uses "Let me think..." In transitional sentences, tasks that require in-depth thinking can withstand longer delays.

6.2. Architecture and rupture detection

The choice of computing model greatly affects how well the system detects breakdowns. The ERR@HRI challenge shows that slight timing errors, including short pauses, are often a signal of system failure (Spitale et al., 2024). The construction of SSM and Mamba may make them more advantageous than the Transformer model in this field. SSM has a hidden state h_t that changes with time but remains constant. Therefore, they can naturally perceive the mutation of time. Compared with the conventional Transformers model, which only processes fixed window data, this feature allows it to detect interaction errors faster.

6.3. Repair strategies matrix

The key to effective repair lies in situational consideration. Table 5 shows that the repair strategy needs to be adjusted according to the error type and relationship duration (Axelsson, Spitale & Gunes, 2024).

Table 5. Matrix of interaction ruptures and repair strategies

Error Type	**Multimodal Indicators**	**Recommended Strategy & Rationale**
Technical Error	User gaze fixation, cessation of motion, frowning.	**Technical Explanation + Retry**. In short-term tasks, users prefer transparency about system limits to understand what went wrong.
Social Friction	Increased speech rate, pitch jitter, face-covering gestures.	**Brief Apology.** Excessive apology disrupts flow; concise acknowledgement maintains task efficiency.
Moral Violation	Body withdrawal, negative sentiment, refusal to interact.	**Empathic Apology + Compensation.** For long-term trust (e.g., care robots), acknowledging feelings is more critical than explaining technical causes.
Deictic Ambiguity	Hesitation, gaze shifting between objects.	**Proactive Clarification.** Asking "Do you mean the red one?" prevents failure before it occurs.

7. Discussion

This section talks about critical benchmarks for measuring performance based on the technological review in the preceding sections. It also talks about the problems that come up when these systems are used in the real world and how they affect society as a whole.

7.1. Measuring symbiosis: Metrics and benchmarks

As HRI advances towards "seamlessness", traditional binary success metrics are no longer sufficient. To address this limitation, a new framework is proposed. This approach evaluates the quality of collaboration by synthesizing multidimensional factors.

7.1.1. Objective metrics

Modality-Specific Fluidity (MSF): Rather than focusing solely on the absolute Response Time Gap (RTG), this metric contextualizes latency by comparing it against the distinct thresholds of each modality: Gaze (<200 ms), Speech (200-500 ms), and Cognitive Reasoning (adaptive).

Concurrent Action Rate (CAR): The proportion of time when humans and robots collaborate at the same time. The higher the collaboration rate, the more people really achieve collaborative work, not simply take turns.

Intervention Ratio: The clear number of corrections required for each operation is adjusted according to the difficulty of the task.

7.1.2. Subjective and psycho-physiological metrics

Perceived Intelligence & Social Presence: Standardized questionnaires, such as the Godspeed questionnaire series, and basic usability principles based on common project design (Norman, 2013).

Cognitive Load: Use pupil measurement or electroencephalogram (e.g., Theta/Alpha ratios) for real-time evaluation to ensure that the "seamless" interface can indeed make it easier for users to think.

Physiological Synchrony: The degree of time synchronization between the user's physiological state (e.g., heart rate variability) and the internal emotional model of the robot. High synchronization shows that there is a deep symbiotic connection, which means that the robot can resonate well with the user's emotional state.

7.2. Real-world challenges and deployment

Although the development of algorithms seems exciting, it is difficult to put them into practice in the real world.

Hardware Constraints: The application of Large Language Models (LLMs) and Vision Language Actions (VLAs) on mobile robots with limited battery capacity and heat dissipation space is still challenging. At this time, the linear complexity of quantitative technology and M3ET algorithm is particularly important.

Network Latency & Integration: The cloud-based foundation model will introduce variable latency. In addition, optimizing perception pipelines for mixed reality and robot integration is still a challenging optimization project (Rathnayake et al., 2020).

Risk and Hallucination: The application of basic models needs to be rigorously tested, because the hallucination risk observed in other high-risk fields (e.g., radiology) (Wiggins & Tejani, 2022) is equally severe in the field of human-robot interaction.

Environmental Noise: Most data sets (e.g., M4Bench) are collected in a controlled environment. Audio and visual noise in the real world, such as blocking and crowd conversation, requires strict "real environment" verification.

7.3. Applications and societal impact

7.3.1. Shared autonomy and latent spaces

When people engage in hard physical labor, they may not always be able to provide accurate instructions. The shared autonomous system makes up for this shortcoming by integrating human instructions and robot capabilities. Researchers suggest using a potential action space model (Jeon, Losey & Sadigh, 2020). which can teach robots to transform simple signals into complex and practical actions. In order to improve the convenience of control, the key is to accurately match these implicit spaces with human intentions (Tucker, Zhou & Shah, 2022). This can significantly simplify practical operations, such as easy delivery tools (Malobický et al., 2025).

7.3.2. Ethical challenges: transparency

"Ethical transparency" is the core concept of SecuRoPS architecture. The concept advocates that robots should take the initiative to inform users when collecting information. This strategy builds an "interactive privacy protection" mechanism to build real trust (Kanbara, Murakawa & Nakanishi, 2024). This method also meets the requirements of the latest guidelines: computer systems must provide "right of interpretation" for their decision-making results (Goodman & Flaxman, 2017).

8. Conclusions

This study has systematically evaluated the paradigm shift in HRI from imperative command-based control to seamless symbiosis. The comparative analysis reveals a fundamental architectural trade-off: while Transformers provide the necessary semantic depth for complex intent understanding, their quadratic computational cost impedes the millisecond-level responsiveness required for high-frequency tactile feedback loops. In contrast, linear-complexity State Space Models (SSMs), such as Mamba, offer a viable path for efficient edge deployment but currently face challenges in retaining long-term historical context. Similarly, the juxtaposition of Modular versus End-to-End VLA paradigms highlights the ongoing tension between the interpretability required for safety-critical applications and the "acting-while-thinking" fluidity offered by streaming architectures like GhostShell.

To bridge these technical gaps, this paper articulated the Multimodal Perception-Driven Decision-Making (MPDDM) framework, providing a conceptual mechanism to resolve sensory conflicts (e.g., speech vs. physiology) via dynamic confidence weighting. Furthermore, the proposal of Modality-Specific Fluidity (MSF) metrics shifts the evaluation focus from binary task completion rates to the temporal synchronization of human-robot dyads.

However, this survey encompasses specific limitations. First, the analysis is primarily architectural and theoretical; the proposed MPDDM framework and MSF metrics currently lack extensive longitudinal empirical validation in unstructured, real-world environments. Second, while ethical transparency was addressed, the broader socio-psychological implications of deploying probabilistic "hallucinating" models in sensitive caregiving scenarios require deeper interdisciplinary investigation. Future research must prioritize bridging the gap between these theoretical models and physical hardware constraints, verifying whether linear-complexity architectures can robustly handle the unpredictability of human behavior in unconstrained real-world environments.

Author contributions

Conceptualization, Methodology, Investigation, Writing−original draft, Writing−review & editing: J.W. The author has read and agreed to the published version of the manuscript.

Submitted: 04 December 2025; Revised: 28 January 2026; Accepted: 03 February 2026; Published: 31 March 2026.

REFERENCES

Adebayo, S., McLoone, S. & Dessing, J. C. (2024) QUB-PHEO: A Visual-Based Dyadic Multi-View Dataset for Intention Inference in Collaborative Assembly. *IEEE Access*. 12, 157050-157066. https://doi.org/10.1109/ACCESS.2024.3485162.

Anderson, P., He, X., Buehler, C. et al. (2018) Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City UT, USA, 2018.* pp. 6077-6086. https://doi.org/10.1109/CVPR.2018.00636.

Axelsson, M., Spitale, M. & Gunes, H. (2024) "Oh, Sorry, I Think I Interrupted You": Designing Repair Strategies for Robotic Longitudinal Well-being Coaching. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, 11-14 March 2024, Boulder, Colorado, USA.* pp. 13-22. https://doi.org/10.1145/3610977.3634948.

Bae, Y. H. & Bennett, C. C. (2025) Multimodal Transformer Models for Turn-taking Prediction: Effects on Conversational Dynamics of Human-Agent Interaction during Cooperative Gameplay. *Human-Computer Interaction*. [Preprint] https://doi.org/10.48550/arXiv.2503.16432. [Accessed: 3rd Dec 2025].

Belpaeme, T., Kennedy, J., Ramachandran, A. et al. (2018) Social robots for education: A review. *Science robotics*. 3(21), eaat5954.

Brown, T., Mann, B., Ryder, N. et al. (2020) Language models are few-shot learners. *Advances in neural information processing systems*. 33, 1877-1901.

Chaabene, S., Boudaya, A., Bouaziz, B. et al. (2025) An overview of methods and techniques in multimodal data fusion with application to healthcare. *International Journal of Data Science and Analytics*. pp. 1-25.

Dennler, N., Nikolaidis, S. & Matarić, M. (2025) Contrastive Learning from Exploratory Actions: Leveraging Natural Interactions for Preference Elicitation. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 4-6 March 2025, Melbourne, Australia*. IEEE. pp. 778-788. https://doi.org/10.1109/HRI61500.2025.10974136.

Devlin, J., Chang, M.-W., Lee, K. et al. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*. Volume 1 (long and short papers). pp. 4171-4186.

Gong, J., Huang, Y., Yuan, B. et al. (2025) GhostShell: Streaming LLM Function Calls for Concurrent Embodied Programming. *arXiv* [preprint] arXiv:2508.05298. [Accessed: 2nd Dec 2025].

Goodman, B. & Flaxman, S. (2017) European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*. 38(3), 50-57.

Gu, A. & Dao, T. (2024) Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling, 7-9 October 2024, Philadelphia, USA*. https://doi.org/10.48550/arXiv.2312.00752.

Gupta, C., Gill, N. S., Gulia, P. et al. (2025) A multimodal fusion model for real-time environment emotion recognition using audio-visual-textual features. *Journal of Big Data.* 12(1), 256.

He, K., Zhang, X., Ren, S. et al. (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770-778.

Jeon, H. J., Losey, D. P. & Sadigh, D. (2020) Shared autonomy with learned latent actions. *Science and Systems*, 2020.

Kanbara, M., Murakawa, Y. & Nakanishi, I. (2024) Privacy-Secure HRI: Framework of Human-Robot Interaction Protecting User's Privacy. In *Proceedings of the 12th International Conference on Human-Agent Interaction*. pp. 382-383.

Lai, Y., Yuan, S., Zhang, B. et al. (2025) Fam-hri: Foundation-model assisted multi-modal human-robot interaction combining gaze and speech. *arXiv* [preprint] arXiv:2503.16492. [Accessed: 2nd Dec 2025].

Li, W., Zhang, R., Shao, R. et al. (2025) SemanticVLA: Semantic-Aligned Sparsification and Enhancement for Efficient Robotic Manipulation. *arXiv* [preprint] arXiv:2511.10518. [Accessed: 2nd Dec 2025].

Liu, Z., Lin, Y., Cao, Y. et al. (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 10012-10022.

Malobický, B., Hruboš, M., Kafková, J. et al. (2025) Towards Seamless Human–Robot Interaction: Integrating Computer Vision for Tool Handover and Gesture-Based Control. *Applied Sciences*. 15(7), 3575.

Mois, G. & Beer, J. M. (2020) Robotics to support aging in place. In *Living with robots*. Elsevier. pp. 49-74.

Norman, D. (2013) *The design of everyday things: Revised and expanded edition*. Basic books.

Ong, K. W., Seet, G. & Sim, S. K. (2008) An implementation of seamless human-robot interaction for telerobotics. *International Journal of Advanced Robotic Systems*. 5(2), 18.

Pramanick, P. & Rossi, S. (2024) PRISCA at ERR@ HRI 2024: Multimodal Representation Learning for Detecting Interaction Ruptures in HRI. In *Proceedings of the 26th International Conference on Multimodal Interaction, 4-8 November 2024, San Jose, Costa Rica.* pp. 666-670.

Rathnayake, D., De Silva, A., Puwakdandawa, D. et al. (2020) Jointly optimizing sensing pipelines for multimodal mixed reality interaction. In *Proceedings of the 2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), 10-13 December 2020, Delhi, India.* IEEE. pp. 309-317. https://doi.org/10.1109/MASS50613.2020.00046.

Shannon, C. E. & Weaver, W. (1949) A mathematical model of communication. Urbana, IL: University of Illinois Press. 11, 11-20.

Spitale, M., Parreira, M. T., Stiber, M. et al. (2024) Err@ hri 2024 challenge: Multimodal detection of errors and failures in human-robot interactions. In *Proceedings of the 26th International Conference on Multimodal Interaction, 4-8 November 2024, San Jose, Costa Rica.* pp. 652-656.

Teoh, J. R., Dong, J., Zuo, X. et al. (2024) Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. *PeerJ Computer Science*. 10, e2298.

Thomaz, A., Hoffman, G. & Cakmak, M. (2016) Computational human-robot interaction. *Foundations and Trends® in Robotics*. 4(2-3), 105-223.

Tucker, M., Zhou, Y. & Shah, J. A. (2022) Latent space alignment using adversarially guided self-play. *International Journal of Human–Computer Interaction*. 38(18-20), 1753-1771.

Van Straten, C. L., Peter, J. & Kühne, R. (2020) Child–robot relationship formation: A narrative review of empirical research. *International Journal of Social Robotics*. 12(2), 325-344.

Vanc, P. & Stepanova, K. (2025) TransforMerger: Transformer-based Voice-Gesture Fusion for Robust Human-Robot Communication. *arXiv* [preprint] arXiv:2504.01708. [Accessed: 2nd Dec 2025].

Vaswani, A., Shazeer, N., Parmar, N. et al. (2017) Attention is all you need. *Advances in neural information processing systems*. 30.

Wiggins, W. F. & Tejani, A. S. (2022) On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*. 4(4), e220119.

Yoshida, A., Dossa, R. F. J., Di Vincenzo, M. et al. (2025) A Multi-User Multi-Robot Multi-Goal Multi-Device Human-Robot Interaction Manipulation Benchmark. *Frontiers in Robotics and AI*. 12, 1528754.

Zhang, Y., He, L., Kang, Z. et al. (2025) M3ET: Efficient Vision-Language Learning for Robotics based on Multimodal Mamba-Enhanced Transformer. *arXiv* [preprint] arXiv:2509.18005. [Accessed: 2nd Dec 2025].

Zhao, W., Gangaraju, K. & Yuan, F. (2025) Multimodal perception-driven decision-making for human-robot interaction: a survey. *Frontiers in Robotics and AI*. 12, 1604472.

Zhao, Z. & McEwen, R. (2025) The robot that stayed: understanding how children and families engage with a retired social robot. *Frontiers in Robotics and AI.* 12, 1628089.

Zitkovich, B., Yu, T., Xu, S. et al. (2023) RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. PMLR*. pp. 2165-2183.

* * *

Jiangke WU is currently a Master's student at the School of Informatics, Computing and Cyber Systems, Northern Arizona University. He earned his Bachelor's degree in "Computer Science & Engineering" from Osmania University. He is an experienced software engineer with over eight years of professional experience in the fields of Information Technology and Artificial Intelligence. His main areas of interests include artificial intelligence applications, software development, system design, information security, and automation.

.