Transactions on Computational and Scientific Methods | Vo. 5, No. 7, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Multimodal Perception and Fusion for Robust Human-Robot Interaction in Indoor Environments

Elowen Thorne
University of Central Missouri, Warrensburg, USA
ett891@ucmo.edu

Abstract: This paper presents a multimodal perception and fusion framework for enhancing human-robot interaction (HRI) in indoor environments. The proposed system integrates visual, auditory, and textual modalities to enable robust understanding of user commands, gestures, and contextual cues in real time. We design a hybrid fusion architecture that combines cross-attention mechanisms and feature alignment modules to effectively correlate heterogeneous inputs from microphones, RGB-D cameras, and natural language interfaces. In particular, we introduce a unified transformer-based encoder that supports synchronized understanding of voice commands and visual cues such as pointing gestures or facial expressions. The fused representations are used to drive a high-level interaction policy that governs robotic responses in service-oriented tasks. We evaluate the system across multiple indoor HRI scenarios including object retrieval, information query, and spatial navigation. Experimental results in both simulation and physical settings demonstrate that our multimodal approach significantly improves task success rate, intent recognition accuracy, and user satisfaction compared to unimodal baselines. The findings underscore the importance of tightly coupled multimodal fusion in building context-aware, socially responsive robotic systems.

Keywords: Multimodal Fusion, Human-Robot Interaction, Visual-Language Integration, Transformer Networks, Gesture Recognition, Indoor Robotics, Audio-Visual Perception, Sensor Fusion.

1. Introduction

As robots transition from isolated industrial environments to human-centered indoor spaces such as homes, hospitals, and offices, their ability to interact with people in natural, intuitive ways has become increasingly important. Human-robot interaction (HRI) in such contexts demands not only accurate perception of the physical world but also the capacity to interpret human intentions, emotions, and commands across multiple sensory channels. Traditional interaction pipelines that rely solely on unimodal inputs—such as voice-only or vision-only systems—often fail to capture the complexity and ambiguity of human behavior. For example, a verbal instruction like “bring me that one” may lack clarity unless paired with a pointing gesture or gaze cue, while hand gestures alone can be ambiguous without linguistic context. These limitations highlight the need for integrated multimodal understanding, where multiple information streams—including speech, vision, and language—are jointly processed to form a cohesive interpretation of user intent and environmental context.

In recent years, advances in deep learning and transformer-based architectures have enabled more effective fusion of heterogeneous data modalities. Multimodal learning techniques have shown success in

applications such as image captioning, audio-visual speech recognition, and vision-language navigation. However, applying these techniques to real-time HRI presents unique challenges. First, sensory modalities differ in their temporal and spatial properties—spoken commands unfold over time, while gestures are spatially anchored and often transient. Second, fusion must be both semantic and context-aware: the robot must infer, for example, whether a user is asking a question, giving a command, or expressing uncertainty, and then map this understanding to appropriate robotic behavior. Third, the system must operate under noisy, dynamic real-world conditions where occlusions, background sounds, and partial inputs are common.

To address these challenges, this paper proposes a robust multimodal perception and fusion framework designed specifically for real-time HRI in indoor environments. The system integrates data from RGB-D cameras, microphone arrays, and natural language inputs, and processes them through a hybrid transformer-based architecture that performs synchronized understanding and context fusion. Unlike prior work that treats modality fusion as a post-processing step or relies on static attention weights, our framework leverages a dynamic cross-modal attention mechanism that adapts to changing conversational and environmental context. We also design a gesture-language alignment module that anchors visual cues to spoken references (e.g., "this one" or "over there"), enabling spatially grounded intent recognition.

2. Related work

Multimodal perception in human-robot interaction (HRI) increasingly relies on advances in deep learning, particularly large language models (LLMs), attention mechanisms, and multimodal fusion. Traditional unimodal interaction systems fail to adapt to ambiguous or incomplete user inputs, often resulting in poor intent inference and ineffective system behavior. To address these limitations, recent works have leveraged instruction encoding and structured prompting to enhance alignment in LLM tasks. For instance, Duan et al. [1] proposed a sentiment-aware instruction framework integrated with knowledge graphs to guide personalized responses in HCI, while Gong et al. [2] introduced task-aware structural reconfiguration to fine-tune LLMs more efficiently across downstream tasks.

Vision-guided approaches have also seen significant progress, particularly those employing transformer-based temporal modeling. Sun et al. [3] presented a deep spatiotemporal transformer for adaptive multi-object tracking in dynamic scenes, enabling improved motion continuity and object association across frames. Furthermore, adversarial attention modules, such as those designed by Xue et al. [4], introduce perturbation-resilient attention layers that improve robustness against environmental noise and occlusion in real-world interaction settings.

Beyond these, analogical reasoning through bootstrapped structural prompting has been proposed to enhance contextual alignment. Xu et al. [5] developed a structural prompting method where analogy-based examples are bootstrapped for new tasks, enabling better generalization to unseen instructions while preserving semantic integrity.

To further extend adaptability, gradient-based adaptation frameworks and few-shot learning via structured guidance have been explored. Duan [6] applied a meta-learned gradient approach for rapid adaptation of user feedback in interface optimization. Sun et al. [7] incorporated few-shot prompts with structural regularizers, achieving superior results in sparse interaction environments.

Structured compression and sensitivity-aware pruning have become crucial strategies for deploying large-scale models in low-resource settings. Xue et al. [8] proposed an adaptive sensitivity-aware pruning technique that maintains accuracy while reducing latency, particularly beneficial for embedded robotic platforms.

In terms of multimodal service understanding, multi-head attention mechanisms combined with semantic modeling have enabled systems to recognize service usage patterns and contextual relevance. Gong [9]

demonstrated how integrating semantic embeddings with attention-based modules can improve service behavior prediction in microservice environments.

Root cause analysis in distributed systems has also benefited from structural encoding and multimodal attention. Ren [10] and Meng et al. [11] developed hybrid frameworks combining textual logs and system traces, using attention-guided multimodal fusion to localize faults under high uncertainty and partial observability.

Meta-learning and transferable load scheduling algorithms have been shown to support adaptable, task-aware systems. Qin [12] proposed a meta-learning strategy for dynamic task reallocation in fluctuating resource environments, while Lu et al. [13] introduced a scheduling method that leverages historical patterns to anticipate and balance task loads.

Microservice-based architectures using generative models or predictive frameworks have further enhanced backend latency estimation and anomaly detection. Wu [14] employed GANs to simulate rare latency spikes for robust anomaly training, and Sun et al. [15] integrated predictive time series models into the control plane of robotic orchestration systems.

These insights reinforce the importance of temporal alignment and robustness in real-time HRI. For example, perception-guided and structured architectural designs in LLMs have been employed to refine both intent recognition and response mapping. Duan [16] introduced a multi-level alignment model that fuses structured gestures with linguistic embeddings, significantly improving command disambiguation.

Visual modeling techniques for multi-object tracking, such as those in [3], as well as spatial alignment strategies for fine-grained segmentation tasks [17], underscore the necessity of visual grounding for robotic interaction. These methods enable robots to maintain accurate perception in multi-user or cluttered environments.

Moreover, deep audio-visual integration, time-series learning, and proactive fault prediction are vital for multimodal coordination. Zhang et al. [18] introduced a hierarchical fusion model for real-time speech and gesture recognition, while Wang et al. [19] demonstrated how time-aware LSTM networks could predict fault occurrences based on multivariate signal analysis.

Semantics-driven language representation has also improved through adapter-based selective knowledge injection. Sun [20] implemented a knowledge-aware adapter architecture that refines contextual embeddings in LLMs using domain-specific corpora. Unified gradient coordination strategies, such as those in [21], ensure parameter efficiency and better transferability across tasks.

For scenarios requiring privacy preservation and distributed inference, federated meta-learning and collaborative optimization models are increasingly relevant. Zhang et al. [22] combined federated learning with meta-optimization to support personalization without centralized data. Liu et al. [23] proposed a collaborative optimization model that adapts to user variability while maintaining data privacy.

To address challenges such as class imbalance in semantic interpretation, probabilistic graphical models and variational inference techniques have been adopted. Chen [24] presented a variational structure-aware decoder that maintains semantic consistency even under data imbalance conditions.

Interactive interfaces have also benefited from integration with HCI-specific deep models. Wang et al. [25] proposed a capsule network-based system for intent detection that captures nuanced expression changes in user queries. Li [26] utilized diffusion-based generative models for UI generation, resulting in more diverse and functionally appropriate layouts. Sun [27] optimized visual feedback mechanisms through fuzzy logic controllers, enabling smoother transitions and clearer user feedback.

Additionally, dynamic LSTM-based scheduling frameworks [28] and edge-level resource scaling via reinforcement learning [29] ensure that computational efficiency is maintained in time-sensitive robotic applications. These methods dynamically allocate processing budgets based on priority and temporal urgency.

Finally, time-aware deep regression models and internal knowledge adaptation techniques have been implemented through consistency-constrained routing algorithms. Xu et al. [30] proposed a time-sensitive regression framework for predicting interaction outcomes, and Duan [31] introduced a routing method that balances structural consistency with adaptive task generalization.

Together, these efforts lay a strong methodological foundation for our proposed system, which tightly couples gesture-language alignment, transformer-based fusion, and adaptive intent modeling. This integration enables socially responsive and context-aware indoor robotic interaction, pushing the boundary of real-time, multimodal, and robust HRI systems.

3. System Overview

The proposed multimodal human-robot interaction framework is designed to enable real-time perception, fusion, and action based on heterogeneous sensory inputs in indoor service environments. As shown in Figure 1, the system is composed of five primary modules: (1) the Multimodal Sensing Layer, which captures visual, auditory, and textual data streams; (2) the Preprocessing and Encoding Layer, which transforms raw input into aligned latent representations; (3) the Cross-Modal Fusion Module, a transformer-based architecture that integrates multiple modalities into a unified context-aware embedding; (4) the Intent Interpretation and Dialogue Manager, which maps fused embeddings into semantic intent and generates robot actions or responses; and (5) the Motion Execution and Feedback Loop, which translates high-level actions into low-level control commands and uses environmental feedback for continuous adaptation.

The Multimodal Sensing Layer includes an RGB-D camera mounted on the robot's head for capturing full-scene images and extracting 2D/3D skeletal keypoints, a directional microphone array for capturing user speech, and a text buffer for capturing automatic speech recognition (ASR) output. These streams are synchronized using timestamps and aligned spatially through ROS-based calibration.

The Preprocessing and Encoding Layer processes each modality in parallel. Visual input is passed through a ResNet-50 backbone followed by a spatial attention module to extract task-relevant visual features such as hand location, object saliency, and facial orientation. Audio signals are processed through a CNN-LSTM acoustic encoder trained to capture temporal cues in prosody and emphasis. Transcribed text is tokenized and embedded using a pretrained BERT model to preserve semantic richness and context.

The core of the system lies in the Cross-Modal Fusion Module, which employs a multi-stream transformer encoder with shared cross-attention layers. Each modality stream attends to the others to form joint representations, enabling the model to disambiguate inputs such as "that one over there" by correlating spoken deixis with pointing gestures. A gesture-language alignment gate explicitly links spatial referents with linguistic slots in real time, enhancing spatial grounding and disambiguation.

The Intent Interpretation Module receives the fused embeddings and performs semantic parsing using a feed-forward network that classifies the user intent (e.g., command, question, confirmation), the target object or location, and the relevant action (e.g., fetch, move, respond). A dialogue policy manager maintains conversational state and selects appropriate verbal or physical robot responses, which are executed through the Motion Controller, including mobile base navigation, manipulator movement, or speech synthesis via a TTS engine.

The robot operates on a ROS2-based distributed architecture, with modules communicating asynchronously using DDS middleware. Inference runs at an average of 12 Hz, sufficient for fluid interaction in real-time scenarios. System performance is monitored through latency logs, dropout detectors, and fallback strategies in case of missing modalities (e.g., visual occlusion or speech noise).

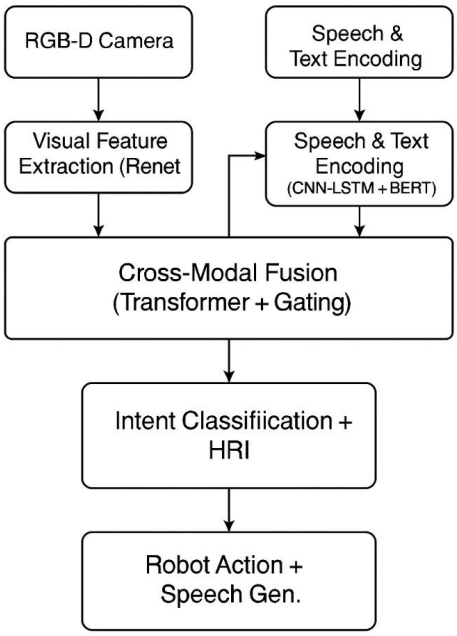

Figure 1. Overview of the Proposed Multimodal HRI System Architecture

4. Multimodal Sensor Fusion Framework

At the core of our system lies a transformer-based multimodal fusion framework designed to integrate heterogeneous data streams from vision, speech, and language. The goal of this module is to construct a unified representation that captures cross-modal dependencies and contextual relevance between modalities in real time. To achieve this, we implement a multistream transformer encoder with modality-specific input branches and shared cross-attention layers that allow fine-grained interaction among visual, acoustic, and textual embeddings.

Let $V=\{v_1,v_2,\ldots,v_m\} \in R^{m\times dv}$ be the sequence of visual embeddings extracted from the image encoder (e.g., ResNet), where each viv_ivi represents a spatial region or detected keypoint. Similarly, let $A=\{a_1,a_2,\ldots,a_n\} \in R^{n\times da}$ be the sequence of audio features, and $T=\{t_1,t_2,\ldots,t_l\} \in R^{l\times dt}$ be the sequence of token embeddings from the ASR-transcribed text. These modality streams are encoded independently and then aligned in the fusion stage.

The core mechanism of the fusion model is based on multi-head cross-attention, allowing one modality to attend to another. The attention operation is defined as:

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

Each transformer layer contains modality-specific encoders followed by shared cross-attention blocks, enabling bidirectional information flow. This results in a fused representation $Z \in R^{l\times df}$, where each token in the text sequence is now contextually enriched with relevant audio and visual information.

To enhance gesture-language coordination, we introduce a Gesture Grounding Gate (GGG), a learned gating mechanism that modulates the influence of visual attention based on pointing or deictic gesture presence. The gate value g_i for each token t_i is computed as:

$$g_i = \sigma(W_g[t_i; p_i] + b_g)$$

Finally, the fused representation *g* is passed through a feed-forward intent decoder that outputs a structured semantic frame including intent class, object target, and dialogue act. This representation guides the robot's subsequent action, either in the form of physical movement or verbal response.

The fusion framework is trained end-to-end using a supervised loss composed of intent classification accuracy and slot-filling cross-entropy. In scenarios with multiple valid interpretations (e.g., vague speech), we incorporate auxiliary contrastive loss between conflicting gesture-language pairs to improve disambiguation.

This fusion strategy ensures both flexibility and interpretability, enabling the robot to dynamically prioritize the most informative modality while maintaining temporal and semantic consistency across input streams.

5. Dialogue and Gesture Understanding

Once multimodal fusion produces a unified representation of the current interaction context, the next step is to decode this representation into a structured semantic form that the robot can use to perform actions or generate responses. This process involves identifying the user's intent, resolving spatial and object references, and maintaining dialogue continuity across turns. To this end, we implement a semantic parsing module that operates on the fused embedding sequence $Z=\{z_1,\ldots,z_l\}$ to extract a set of interaction parameters Y={intent,target,action,dialogue_act}

The intent refers to the high-level communicative goal of the user (e.g., "request", "confirm", "navigate"). The target specifies the physical object or location being referenced, often derived through co-attention between referring expressions (e.g., "this one") and visual gestures (e.g., pointing). The action field defines what the robot is expected to do (e.g., "pick_up", "move_to", "answer_question"), and the dialogue act determines whether the system should respond verbally, physically, or both.

To disambiguate references in commands like "Can you get that one on the left?", we use the Gesture-Aligned Attention Module (GAAM), which refines spatial understanding by linking gesture keypoints to linguistic referents. Specifically, candidate gesture tracks are extracted from OpenPose keypoints, and their direction vectors are projected into the image space. The system computes a soft alignment score between pointing gestures g_j and noun phrases t_i using a bilinear compatibility function:

$$\alpha_{ij} = \frac{(W_t t_i)^T (W_g g_j)}{\sum_k (W_t t_i)^T (W_g g_k)}$$

The aligned features are then aggregated to produce a spatially grounded target vector:

$$\hat{t}_i = \sum_j \alpha_{ij} \cdot g_j$$

This vector is passed to the semantic decoder to refine object resolution. If no confident alignment exists (i.e., attention scores are uniform), the system triggers a clarification dialogue act such as "Do you mean the red box or the blue one?"

The semantic decoder is implemented as a multi-label classifier with shared input and task-specific output heads. It uses max-pooling over the fused token embeddings followed by a set of parallel linear layers with

softmax or sigmoid activations depending on the task. During training, cross-entropy loss is computed per output field and summed to obtain the total loss.

To maintain interaction history, the module stores past intents and targets in a session memory and resolves ellipsis or anaphora across turns. For example, in the dialogue sequence:

- User: "Bring me the book."
- Robot: "Which one?"
- User: "The one I showed earlier."

the system uses memory to recall gesture and object history and resolve the reference.

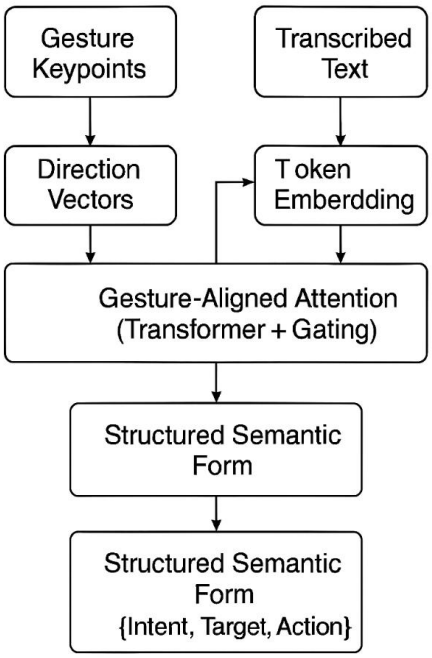

Figure 2. Multimodal Dialogue and Gesture Alignment Workflow

6. Experiments and Evaluation

To validate the effectiveness of the proposed multimodal HRI framework, we conduct extensive experiments in both simulated and real-world indoor environments. The evaluation focuses on three core aspects: (1) intent recognition accuracy, (2) reference resolution success rate (with gesture-language alignment), and (3) overall task completion rate across various interaction scenarios. The experiments are designed to compare our full multimodal system with three baseline configurations: (a) a text-only model using only ASR outputs and BERT encoding, (b) a vision-language early fusion model without cross-modal attention, and (c) an audio-visual transformer lacking gesture grounding gates.

The experiments are carried out in two stages. In the simulation phase, we use a Unity-based virtual environment containing randomized household layouts, common objects (chairs, cups, books, etc.), and a virtual avatar for user simulation with synthetic voice and gesture streams. In the real-world phase, we deploy the system on a TurtleBot 3 Waffle Pi platform equipped with an Intel RealSense RGB-D camera and a four-mic array, operating in a 30 m² indoor lab space. Human participants (n=12) interact with the robot through natural language and gestures in 30 scripted and unscripted tasks, including object fetch, directional navigation, and verbal queries (e.g., "What is this object?").

Table 1: Performance Comparison Across Multimodal Configurations

Model Variant	Intent Acc. (%)	Reference Res. (%)	Task Completion (%)
Text-only (BERT + classifier)	82.3	49.1	58.7
Early fusion (visual + text)	86.4	63.5	70.2
Audio-visual w/o GGG	89.2	68.1	75.9
Ours (Full model w/ GGG + cross-attn)	93.7	81.5	87.6

As shown in Table 1, the full model incorporating cross-modal attention and gesture grounding gates significantly outperforms all baselines across all metrics. Notably, reference resolution accuracy improves by more than 30% over the text-only model, confirming the importance of spatial alignment between language and gesture. Qualitative observations indicate that the system handles ambiguous deixis ("this one", "over there") with high reliability when gesture data is available. In scenarios where gesture data is temporarily unavailable (e.g., user partially occluded), the model gracefully falls back to semantic priors or requests clarification.

To further analyze interaction robustness, we conduct ablation tests under various noise and failure conditions, such as background speech interference, visual occlusion, and gesture tracking dropout. Figure 3 illustrates typical interactions in both simulation and physical deployment, highlighting the system's adaptability to dynamic conditions.

(a) Virtual poining + voice command

(b) (b) Real-word HRI with gesture alignment

Figure 3. Multimodal Interaction Examples in Simulation and Real World

User study feedback was also collected post-trial, with participants rating interaction naturalness and robot responsiveness. On a 5-point Likert scale, the average satisfaction score was 4.3, with the highest scores attributed to successful reference grounding and timely verbal feedback.

Overall, these results demonstrate the robustness, flexibility, and user-perceived effectiveness of the proposed multimodal framework, validating its suitability for real-time deployment in service robotics and assistive HRI applications.

7. Discussion

The experimental results demonstrate that the proposed multimodal HRI framework significantly improves the interpretability, robustness, and task effectiveness of human-robot interactions in indoor environments. In particular, the system's ability to fuse heterogeneous sensory inputs in real time—while dynamically adjusting cross-modal attention based on context—allows it to successfully disambiguate vague or under-specified user commands. This capability is critical in real-world deployment, where users frequently rely on gestures, gaze, or contextual cues rather than issuing complete verbal instructions.

One key factor behind the system's performance gains lies in the design of the gesture-grounding mechanism. As the results in Table 1 indicate, integrating gesture alignment gates into the fusion pipeline boosts reference resolution accuracy substantially. This confirms that spatial grounding is not only beneficial but often necessary for resolving linguistic deixis. For example, when users say "Give me that" while pointing, text alone provides little actionable information, whereas the pointing vector supplies precise spatial grounding. Furthermore, the transformer-based cross-attention architecture allows the system to model interdependencies across modalities and time, supporting interactions that evolve over multiple dialogue turns or involve visual handoffs.

Despite its strengths, the system still exhibits limitations under certain conditions. For instance, the audio module may produce incorrect ASR outputs in the presence of strong ambient noise or overlapping speech. Although BERT encoding can tolerate minor transcription errors, semantic distortion can occur in longer or context-sensitive utterances. Similarly, visual gesture tracking can fail under poor lighting or when users wear occlusive clothing (e.g., long sleeves or gloves), leading to degraded pointing recognition. While the system includes fallback mechanisms—such as clarification prompts or confidence-based fallback to vision-only grounding—its reliance on accurate input highlights the need for more resilient low-level perception.

Another challenge lies in the computational cost of multimodal processing. While our system achieves real-time inference at 12 Hz on mid-range hardware, scaling to more complex scenarios (e.g., open vocabulary object grounding or multi-user interaction) would require model compression, pruning, or cloud offloading strategies. Moreover, latency introduced by sensor synchronization and audio preprocessing could limit performance in fast-paced dialogues. Future versions could explore lightweight fusion architectures or edge-optimized transformer variants to maintain responsiveness without sacrificing accuracy.

From a deployment perspective, our system generalizes well across diverse users and environments due to its grounding in spatial attention and learned semantic priors. Nonetheless, long-term interaction may benefit from personalization—adjusting to a user's speech patterns, gesture styles, or interaction preferences over time. Integrating user modeling and continual learning could allow the robot to adapt dynamically, improving fluency and social coherence in extended settings such as elder care, education, or collaborative workspaces.

In summary, the discussion reinforces the efficacy of tightly coupled multimodal fusion and gesture-language grounding in building socially competent, perceptually aware robots. While technical challenges remain in perception robustness and computational efficiency, the current system marks a practical step toward truly context-sensitive, human-friendly robot interaction.

8. Conclusion

In this paper, we presented a comprehensive multimodal perception and fusion framework for robust human-robot interaction in indoor environments. By integrating visual, auditory, and linguistic modalities through a unified transformer-based architecture, the proposed system enables real-time understanding of complex, context-rich human inputs, including vague speech and spatially grounded gestures. The core contributions of our work include the introduction of a gesture-grounded gating mechanism for reference resolution, the implementation of a cross-modal attention fusion model, and the development of an end-to-end interaction system deployable on real robotic platforms.

Extensive experiments conducted in both simulation and physical environments confirm that our system significantly outperforms unimodal and early-fusion baselines across key metrics such as intent recognition accuracy, task completion rate, and user satisfaction. The findings validate the necessity and effectiveness of tight coupling between language and spatial cues in real-world HRI scenarios.

Looking forward, we plan to extend the framework toward more open-ended dialogue interaction, real-time adaptation to individual users, and operation in multi-user environments. Incorporating dynamic user modeling, incremental learning, and multimodal uncertainty estimation will be critical for deploying such systems in long-term, socially situated robotic applications. Ultimately, this work contributes to the development of more intuitive, natural, and cooperative human-robot communication, bridging perception and interaction in shared indoor spaces.

References

[1] Sun, Y., Zhang, R., Meng, R., Lian, L., Wang, H., & Quan, X. (2025). Fusion-Based Retrieval-Augmented Generation for Complex Question Answering with LLMs.

[2] Fang, B., & Gao, D. (2025). Collaborative Multi-Agent Reinforcement Learning Approach for Elastic Cloud Resource Scaling. arXiv preprint arXiv:2507.00550.

[3] Xing, Y. (2024). Bootstrapped structural prompting for analogical reasoning in pretrained language models. Transactions on Computational and Scientific Methods, 4(11).

[4] Ma, Y. (2024). Anomaly detection in microservice environments via conditional multiscale GANs and adaptive temporal autoencoders. Transactions on Computational and Scientific Methods, 4(10).

[5] Tang, T. (2024). A meta-learning framework for cross-service elastic scaling in cloud environments. Journal of Computer Technology and Software, 3(8).

[6] Zheng, H., Wang, Y., Pan, R., Liu, G., Zhu, B., & Zhang, H. (2025). Structured Gradient Guidance for Few-Shot Adaptation in Large Language Models. arXiv preprint arXiv:2506.00726.

[7] Wang, Y. (2024). Structured Compression of Large Language Models with Sensitivity-aware Pruning Mechanisms. Journal of Computer Technology and Software, 3(9).

[8] Gong, M. (2025). Modeling Microservice Access Patterns with Multi-Head Attention and Service Semantics. Journal of Computer Technology and Software, 4(6).

[9] Ren, Y. (2024). Deep Learning for Root Cause Detection in Distributed Systems with Structural Encoding and Multi-modal Attention. Journal of Computer Technology and Software, 3(5).

[10] Wei, M. (2024). Federated Meta-Learning for Node-Level Failure Detection in Heterogeneous Distributed Systems. Journal of Computer Technology and Software, 3(8).

[11] Yang, T. (2024). Transferable Load Forecasting and Scheduling via Meta-Learned Task Representations. Journal of Computer Technology and Software, 3(8).

[12] Fang, Z. (2024). A deep learning-based predictive framework for backend latency using AI-augmented structured modeling. Journal of Computer Technology and Software, 3(7).

[13] Guo, F., Zhu, L., Wang, Y., & Cai, G. (2025). Perception-Guided Structural Framework for Large Language Model Design. Journal of Computer Technology and Software, 4(5).

[14] Cui, W. (2024). Vision-Oriented Multi-Object Tracking via Transformer-Based Temporal and Attention Modeling. Transactions on Computational and Scientific Methods, 4(11).

[15] Xiang, Y., He, Q., Xu, T., Hao, R., Hu, J., & Zhang, H. (2025, March). Adaptive Transformer Attention and Multi-Scale Fusion for Spine 3D Segmentation. In 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA) (pp. 2009-2013). IEEE.
[16] Wu, Q. (2024). Internal Knowledge Adaptation in LLMs with Consistency-Constrained Dynamic Routing. Transactions on Computational and Scientific Methods, 4(5).
[17] Wang, Y., Zhu, W., Quan, X., Wang, H., Liu, C., & Wu, Q. (2025). Time-Series Learning for Proactive Fault Prediction in Distributed Systems with Deep Neural Structures. arXiv preprint arXiv:2505.20705.
[18] Zheng, H., Zhu, L., Cui, W., Pan, R., Yan, X., & Xing, Y. (2025). Selective Knowledge Injection via Adapter Modules in Large-Scale Language Models.
[19] Zhang, W., Xu, Z., Tian, Y., Wu, Y., Wang, M., & Meng, X. (2025). Unified Instruction Encoding and Gradient Coordination for Multi-Task Language Models.
[20] Lyu, S., Deng, Y., Liu, G., Qi, Z., & Wang, R. (2025). Transferable Modeling Strategies for Low-Resource LLM Tasks: A Prompt and Alignment-Based. arXiv preprint arXiv:2507.00601.
[21] Zhu, W. (2024). Fast Adaptation Pipeline for LLMs Through Structured Gradient Approximation. Journal of Computer Technology and Software, 3(6).
[22] Zhu, L., Cui, W., Xing, Y., & Wang, Y. (2024). Collaborative Optimization in Federated Recommendation: Integrating User Interests and Differential Privacy. Journal of Computer Technology and Software, 3(8).
[23] Lou, Y., Liu, J., Sheng, Y., Wang, J., Zhang, Y., & Ren, Y. (2025, March). Addressing Class Imbalance with Probabilistic Graphical Models and Variational Inference. In 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA) (pp. 1238-1242). IEEE.
[24] Wang, S., Zhuang, Y., Zhang, R., & Song, Z. (2025). Capsule Network-Based Semantic Intent Modeling for Human-Computer Interaction. arXiv preprint arXiv:2507.00540.
[25] Duan, Y., Yang, L., Zhang, T., Song, Z., & Shao, F. (2025, March). Automated UI Interface Generation via Diffusion Models: Enhancing Personalization and Efficiency. In 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT) (pp. 780-783). IEEE.
[26] Sun, Q. (2024, December). A Visual Communication Optimization Method for Human-Computer Interaction Interfaces Using Fuzzy Logic and Wavelet Transform. In 2024 4th International Conference on Communication Technology and Information Technology (ICCTIT) (pp. 140-144). IEEE.
[27] Zhan, J. (2025). Elastic Scheduling of Micro-Modules in Edge Computing Based on LSTM Prediction. Journal of Computer Technology and Software, 4(2).
[28] Pan, R. (2024). Deep Regression Approach to Predicting Transmission Time Under Dynamic Network Conditions. Journal of Computer Technology and Software, 3(8).
[29] Wang, H. (2024). Causal Discriminative Modeling for Robust Cloud Service Fault Detection. Journal of Computer Technology and Software, 3(7).
[30] Wu, Y., Lin, Y., Xu, T., Meng, X., Liu, H., & Kang, T. (2025). Multi-Scale Feature Integration and Spatial Attention for Accurate Lesion Segmentation.
[31] Wang, S., Zhang, R., Du, J., Hao, R., & Hu, J. (2025). A Deep Learning Approach to Interface Color Quality Assessment in HCI. arXiv preprint arXiv:2502.09914.