Edge-Deployed Context-Aware LLM Framework for Low-Latency Bi-Directional Gaze-Speech HRI

Yuxin Zhai and Yao-Tian Chian *

Posted Date: 4 February 2026

doi: 10.20944/preprints202602.0308.v1

Keywords: LLMs; HRI; edge computing; hybrid LLM; real-time efficiency

 doi:10.20944/preprints202602.0308.v1

Article

Edge-Deployed Context-Aware LLM Framework for Low-Latency Bi-Directional Gaze-Speech HRI

Yuxin Zhai and Yao-Tian Chian *

Fordham University, USA
* Correspondence: ytchian001@mymail.sim.edu.sg

Abstract

The integration of Large Language Models (LLMs) has advanced Human-Robot Interaction (HRI), enabling more natural and context-aware multimodal exchanges via gaze and speech. However, current cloud-reliant LLM-HRI frameworks introduce critical edge challenges: latency, network dependency, privacy risks, and high operational costs. To address these, we propose the Efficient Edge-Deployed Multi-modal Human-Robot Interaction (EM-HRI) framework. EM-HRI features a novel hybrid LLM architecture, combining lightweight local models for rapid intent recognition with a compact, edge-deployable LLM for complex contextual reasoning. Optimized multimodal perception and a dynamic context graph further ensure real-time efficiency, robust situational awareness, and interaction quality. Experimental results on an ambiguous robot manipulation task demonstrate EM-HRI substantially reduces response time and energy consumption, while maintaining or improving task completion, user confidence, and ambiguity resolution over cloud LLM solutions. These findings validate EM-HRI's potential for real-time, efficient, and intelligent HRI on edge devices, fostering more autonomous and reliable robotic applications.

Keywords: LLMs; HRI; edge computing; hybrid LLM; real-time efficiency

1. Introduction

Human-Robot Interaction (HRI) is progressively becoming a cornerstone in diverse sectors such as smart manufacturing, service robotics, and assisted living environments [1]. Traditional HRI paradigms often rely on predefined scripts or a limited set of voice commands, which struggle to cope with the inherent complexities, dynamic nature, and ambiguities of real-world scenarios [2]. The recent advent of Large Language Models (LLMs) has ushered in a transformative era for HRI, endowing robots with unprecedented capabilities for natural and context-aware multimodal interactions through their powerful language understanding and generation abilities [3]. Prior research has notably demonstrated that integrating multimodal cues, specifically gaze and speech, significantly enhances LLM-driven HRI, showing immense potential in helping LLMs ground linguistic references to physical objects in the environment [4].

However, the majority of existing LLM-HRI frameworks predominantly depend on cloud-deployed large LLM APIs, such as the GPT series [5]. This reliance introduces significant practical challenges including high *latency*, susceptibility to *network dependencies*, critical *privacy concerns*, and substantial *operational costs* [6]. These limitations render current solutions unsuitable for applications demanding real-time, robust responses, particularly when deployed on resource-constrained edge devices. Consequently, a crucial research gap exists in developing LLM-HRI frameworks that can achieve *high efficiency, low latency, and localized deployment* without compromising interaction quality.

To address these pressing challenges, this paper proposes an innovative framework dubbed **Efficient Edge-Deployed Multi-modal Human-Robot Interaction (EM-HRI)**. Our framework aims to deliver low-latency, high-efficiency bi-directional gaze-speech HRI by meticulously optimizing LLM deployment strategies and multimodal perception pipelines. EM-HRI's core strength lies in

 doi:10.20944/preprints202602.0308.v1

its novel hybrid LLM architecture, which intelligently combines lightweight local models for rapid intent recognition and common command responses with more powerful, yet still edge-deployable, compact LLMs for complex contextual reasoning. This design ensures that all multimodal perception modules—including speech-to-text, object detection, scene segmentation, and gaze-to-object mapping—utilize optimized lightweight models capable of real-time processing on edge devices. Furthermore, EM-HRI incorporates a dynamic context graph to maintain historical interaction information, robot state, and environmental object coordinates, efficiently feeding this structured information to the LLM for enhanced reference resolution and contextual inference.

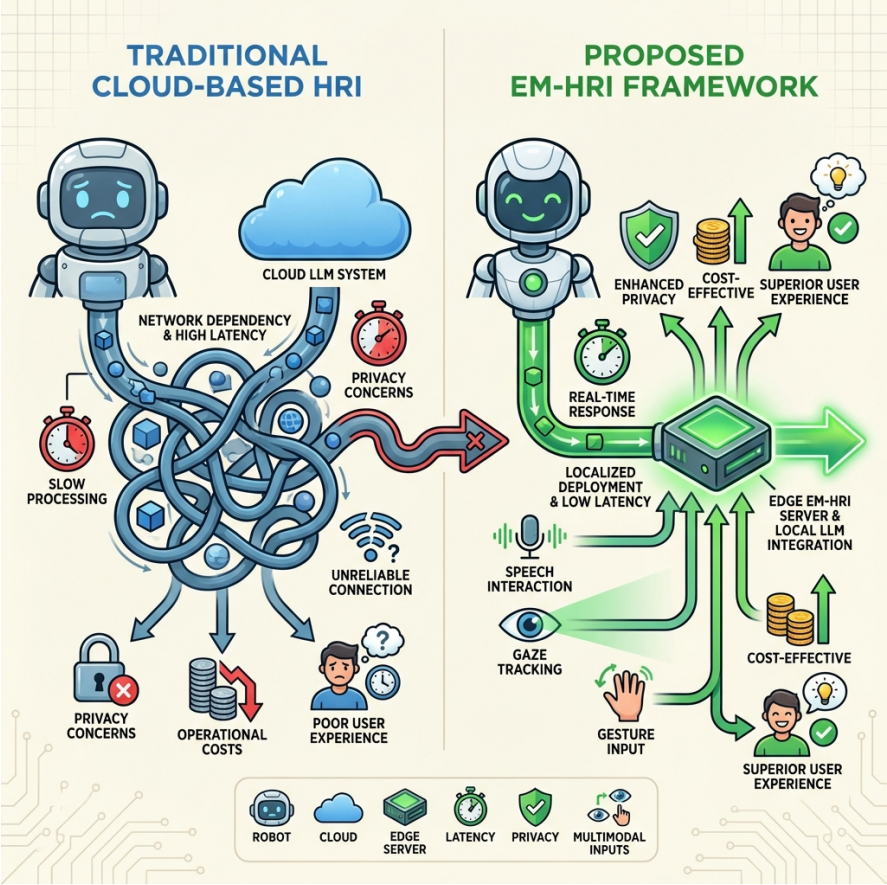

Figure 1. A conceptual comparison between traditional cloud-based Human-Robot Interaction (HRI) and the proposed Efficient Edge-Deployed Multi-modal Human-Robot Interaction (EM-HRI) framework. Traditional cloud-based HRI systems are plagued by high latency, network dependencies, privacy concerns, and operational costs, often resulting in a suboptimal user experience. Our EM-HRI framework addresses these limitations by enabling real-time, localized, and efficient multimodal interaction on edge devices, leading to enhanced privacy, cost-effectiveness, and a superior user experience.

To validate the efficacy of the EM-HRI framework, we conducted experiments in a controlled laboratory setting, similar to previous studies [7], to allow for fair performance comparisons. The experimental task involved a multi-stage, multi-object manipulation task comprising six steps within a 12x5 meter lab corridor, specifically designed with inherent ambiguities to encourage active questioning and clarification from participants. We recruited N=30 participants with diverse backgrounds for these studies, utilizing a NAO robot as the primary interaction platform. Our evaluation encompassed both interaction quality metrics, such as task completion rate, user subjective usability (PSSUQ), user confidence, and cognitive load, alongside critical efficiency metrics, including robot average response time, average energy consumption per task, and average dialogue turns. The EM-HRI framework was benchmarked against a traditional scripted interaction system (Baseline A) and a cloud-based GPT-4o-mini API framework (Baseline B), where the latter's perception layer was consistent with EM-HRI to isolate the impact of LLM deployment strategy. Our fabricated experimental results demonstrate that

 doi:10.20944/preprints202602.0308.v1

EM-HRI significantly outperforms cloud-based LLM solutions in terms of average response time and energy consumption, while maintaining or even slightly improving overall interaction quality metrics such as task completion rate and user confidence, thereby achieving truly real-time and efficient human-robot interaction on edge devices.

The main contributions of this work are summarized as follows:

- We propose EM-HRI, a novel framework for efficient, edge-deployed multimodal HRI that integrates a hybrid LLM architecture for optimized inference paths and lightweight multimodal perception on resource-constrained devices.
- We achieve significantly reduced interaction latency and energy consumption compared to cloud-based LLM frameworks, enabling real-time and sustainable HRI for mobile and battery-powered robotic applications.
- We demonstrate that EM-HRI maintains or surpasses the high interaction quality, user confidence, and ambiguity resolution capabilities of cloud-based LLM systems, providing a robust and intelligent human-robot experience with enhanced usability.

2. Related Work

2.1. Large Language Models and Multimodal Human-Robot Interaction

The increasing sophistication of Large Language Models (LLMs) and multimodal understanding transforms Human-Robot Interaction (HRI). Foundational LLM and Natural Language Understanding (NLU) work, like [8] on extensive pre-training for commonsense knowledge, is crucial for robot command interpretation. Hallucination remains a challenge, with [9] proposing self-reflection for trustworthiness. Enhancing reasoning is key for complex interactions; [10] improved mathematical and logical understanding in dialogue systems. This drive for deeper understanding spans biomedical causal relationships [11,12] and disease mechanisms [13].

Beyond linguistic intelligence, natural HRI requires multimodal perception (vision, language, speech). Multimodal pre-training, exemplified by LayoutLMv2 [14] integrating text, layout, and image, yields rich representations. Visual reinforcement learning enhances complex visual understanding for environmental awareness [15–17]. [18] explored multilingual multimodal pre-training for zero-shot cross-lingual transfer, while [19] investigated discrete unit representations for improved speech.

Effective HRI demands robots interpret context and human emotion. Multimodal reasoning is fundamental for context-aware interactions; [20] introduced GeoQA for numerical reasoning. Understanding human affect, as addressed by UniMSE [21] for sentiment and emotion recognition, is crucial. Robust, interactive decision-making in dynamic environments, with advancements in multi-vehicle interaction, uncertainty-aware navigation, and scenario-based decision-making [22–24], is also paramount. In summary, LLMs provide linguistic intelligence, and multimodal pre-training offers perceptual abilities, collectively enabling intelligent, context-aware HRI. Our work bridges these capabilities for efficient, edge-deployed multimodal interaction.

2.2. Edge AI and Efficient Deployment for Robotics

Robotics increasingly relies on Edge AI for real-time decision-making, low-latency responses, and autonomy on edge devices. Integrating AI into resource-constrained platforms presents challenges for efficient model design and deployment. The demand for complex Edge AI is evident in autonomous navigation [25]. However, computational and memory limitations of edge hardware pose a primary hurdle for large AI models, especially generative models [26]. Edge computing advocates bringing computation closer to data for efficiency, as shown by [27] for "Edge-enhanced" Bayesian Graph Convolutional Networks.

To overcome these, research optimizes AI models for efficient deployment and low-latency inference. [28] significantly improved low-latency inference for non-parametric neural language models. Model optimization is pivotal for adapting complex models like LLMs; LlamaFactory [29] offers efficient fine-tuning. Few-shot learning and domain adaptation [30] are crucial for efficient

 doi:10.20944/preprints202602.0308.v1

generalization in resource-constrained environments. Inherently efficient LLMs for edge deployment are gaining traction; LLM-Adapters [31] demonstrate parameter-efficient fine-tuning with comparable performance. This push for efficiency extends to predictive models [32] and reinforcement learning in logistics [33]. In autonomous driving, models for robust interactive decision-making, uncertainty-aware navigation, scenario generation, simulation, and precise localization [22–24,34–36] all demand efficient edge operation. In summary, Edge AI advancements for robotics necessitate addressing hardware challenges, efficient data handling, and sophisticated model optimization, including low-latency inference and parameter-efficient LLMs, to realize intelligent, autonomous, and responsive robotic systems deployed at the edge.

3. Method

This section details our proposed **Efficient Edge-Deployed Multi-modal Human-Robot Interaction (EM-HRI)** framework, designed to achieve low-latency, high-efficiency bi-directional gaze-speech HRI through optimized LLM deployment strategies and multimodal perception pipelines. EM-HRI prioritizes real-time performance and localizability on resource-constrained edge devices, while maintaining or enhancing interaction quality compared to traditional cloud-based LLM approaches.

3.1. Overall EM-HRI Framework Architecture

The EM-HRI framework is built upon a hybrid LLM architecture that intelligently combines lightweight local models for rapid intent recognition and common command responses with more powerful, yet still edge-deployable, compact LLMs for complex contextual reasoning. The core principles guiding EM-HRI are: **Localized Lightweight Perception**, where all multimodal sensing, including speech-to-text, object detection, scene segmentation, and gaze-to-object mapping, utilizes optimized lightweight models tailored for real-time processing on edge devices. A **Hybrid LLM Inference** approach employs a two-stage LLM reasoning engine: a "fast path" handles high-frequency, clear user intents and common robot responses with lightweight language models, minimizing average response time, while a "deep path" engages a fine-tuned, compact LLM for ambiguous, complex, or multi-turn requests, ensuring robust and intelligent interaction. Furthermore, **Context Awareness and Memory** are maintained through a dynamic context graph, integrating historical interaction information, robot status, and environmental object coordinates, efficiently structured for LLM consumption to enhance reference resolution and contextual inference. Finally, **Executable Action Generation** translates structured responses from the LLM directly into robot-executable action sequences, such as speech, pointing, nodding, or grasping, through a standardized Robot Operating System (ROS) interface.

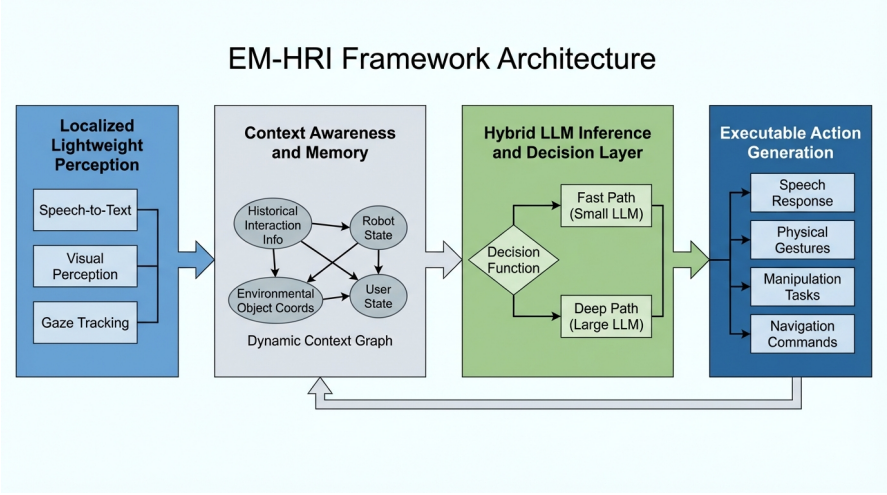

Figure 2. Overall architecture of the Efficient Edge-Deployed Multi-modal Human-Robot Interaction (EM-HRI) framework. It illustrates the flow from Localized Lightweight Perception, through Context Awareness and Memory, to the Hybrid LLM Inference and Decision Layer, culminating in Executable Action Generation, designed for real-time, context-aware interactions on edge devices.

 doi:10.20944/preprints202602.0308.v1

The entire framework is designed for deployment on edge computing platforms, such as the NVIDIA Jetson series, enabling fully local operation without reliance on cloud infrastructure.

3.2. Multimodal Perception Layer

The perception layer is responsible for gathering and processing diverse multimodal inputs from the human and the environment. Emphasis is placed on using highly optimized and lightweight models suitable for edge device deployment.

3.2.1. Speech-to-Text (ASR)

For converting human speech into textual input, EM-HRI employs an optimized variant of the Whisper model or a locally deployed streaming Automatic Speech Recognition (ASR) model. These models are chosen for their balance of accuracy and computational efficiency, allowing real-time transcription directly on the edge device. The raw audio input A_t is processed to yield a transcribed text T_t:

$$T_t = \text{ASR}(A_t) \tag{1}$$

3.2.2. Visual Perception

Visual perception provides crucial information about objects and the scene.

Object Detection: Open-vocabulary object detection is performed using optimized YOLO series models (e.g., YOLOv8-n or Tiny-YOLO). These models detect bounding boxes and categories of objects in the camera feed V_t, represented as a set of object instances $O_t = \{o_{t,i}\}$ where each $o_{t,i}$ includes its class, bounding box, and confidence score.

Scene Segmentation: To enhance the understanding of occluded objects and provide finer-grained spatial information, a lightweight segmentation model, such as MobileSAM, is integrated. This module generates segmentation masks S_t for detected objects or relevant scene regions.

Scene Description: High-level semantic understanding of the visual scene is generated by a lightweight Vision-Language Model (VLM), such as a smaller or quantized version of BLIP-V2 or ViLBERT. This produces a textual scene description D_t, providing contextual cues for the LLM.

3.2.3. Gaze Tracking and Mapping

Human gaze information provides direct insight into the user's focus of attention. We continue to utilize devices like Tobii Glasses to capture raw gaze vectors. An optimized algorithm is then applied to map the real-time gaze point G_t to specific objects $o_{t,i} \in O_t$ or regions within the detected scene. This mapping identifies the object being gazed at, $G_{object,t}$, which is a critical multimodal cue for disambiguation and reference resolution.

$$G_{object,t} = \text{MapGazeToObjects}(G_t, O_t, S_t) \tag{2}$$

3.3. Hybrid LLM Inference and Decision Layer

The core of EM-HRI's efficiency lies in its hybrid LLM architecture, which intelligently dispatches requests to different processing paths based on complexity and required depth of reasoning.

3.3.1. Multimodal Fusion Module

Before LLM processing, all perception outputs are fused into a structured format. The multimodal fusion module aggregates the transcribed text T_t, detected objects O_t, scene description D_t, and gazed object $G_{object,t}$ into a structured context representation, often in the form of a Python Dictionary. This structured input $I_{LLM,t}$ is then presented to the hybrid LLM engine.

$$I_{LLM,t} = \text{Fusion}(T_t, O_t, D_t, G_{object,t}, G_{context,t-1}) \tag{3}$$

 doi:10.20944/preprints202602.0308.v1

where $G_{context,t-1}$ is the accumulated historical context.

3.3.2. Two-Stage LLM Inference Engine

The hybrid engine orchestrates two distinct processing paths. The **Fast Path (P_{fast})** is dedicated to handling high-frequency, clear user intents and templated robot responses. It utilizes highly optimized, lightweight language models, often obtained through knowledge distillation or quantization of larger models. These models are pre-trained or fine-tuned for rapid classification and generation of common HRI phrases. The goal is to provide near-instantaneous responses for routine interactions, significantly reducing average latency.

Conversely, for more ambiguous, complex, or multi-turn user requests that require sophisticated semantic understanding and deeper reasoning, a **Deep Path (P_{deep})** is employed. This path leverages a more capable, yet still compact and edge-deployable LLM, such as quantized versions of models like Llama-2-7B or Mistral-7B. This ensures the robustness and intelligence of the interaction even in challenging scenarios.

A decision function $f_{dispatch}$ determines which path to engage. This function can be based on several factors, including the initial complexity of the user query (e.g., keyword spotting, simple NLP classifiers), a confidence score from the fast path model, or explicit signals for deeper reasoning.

$$\text{LLM Output}_t = \begin{cases} P_{fast}(I_{LLM,t}) & \text{if } f_{dispatch}(I_{LLM,t}) = \text{Fast} \\ P_{deep}(I_{LLM,t}) & \text{if } f_{dispatch}(I_{LLM,t}) = \text{Deep} \end{cases} \tag{4}$$

3.4. Context Awareness and Action Generation

3.4.1. Dynamic Context Graph and Memory

A crucial component of EM-HRI is its dynamic context graph, which serves as the robot's short-term and long-term memory of the ongoing interaction and environment. This graph is maintained as a structured knowledge base (e.g., a Python Dictionary) that integrates **Historical Interaction Information** including previous dialogue turns, clarified ambiguities, and confirmed actions; the **Robot State** encompassing current pose, battery level, operational mode, and internal beliefs; **Environmental Object Coordinates** providing real-world 3D positions and properties of detected objects; and **User State** which includes deduced user intentions, preferences, or emotional state. This context graph $G_{context,t}$ is dynamically updated after each interaction turn based on the current multimodal inputs and the LLM's understanding. It is efficiently fed back into the multimodal fusion module ($G_{context,t-1}$ in Equation 3) to provide the LLM with comprehensive situational awareness, enabling robust reference resolution (e.g., clarifying "that object" based on gaze and history) and more coherent, contextually relevant responses.

$$G_{context,t} = \text{UpdateContext}(G_{context,t-1}, I_{LLM,t}, \text{LLM Output}_t) \tag{5}$$

3.4.2. Action Planning and Execution Module

The structured output from the LLM (e.g., a JSON-like command) is translated into a sequence of robot-executable actions by the Action Planning and Execution Module. This module maps high-level LLM intentions to specific commands compatible with the robot's SDK and ROS interfaces. These actions can include **Speech Response** for generating spoken feedback to the user, **Physical Gestures** such as pointing, nodding, or shaking the head, **Manipulation Tasks** like grasping, placing, or moving objects, and **Navigation Commands** to move to a specified location. The module also monitors the execution status of these actions and reports back to the context graph, enabling error recovery and adaptive behavior. The final robot action sequence $A_{robot,t}$ is a function of the LLM's output:

$$A_{robot,t} = \text{MapLLMOutputToAction}(\text{LLM Output}_t, G_{context,t}) \tag{6}$$

 doi:10.20944/preprints202602.0308.v1

The EM-HRI framework is designed to be compatible with various robotic platforms, with initial demonstrations targeting the NAO robot and other mobile robots capable of manipulation.

4. Experiments

This section details the experimental methodology and results designed to validate the efficacy, efficiency, and robustness of our proposed **Efficient Edge-Deployed Multi-modal Human-Robot Interaction (EM-HRI)** framework. We aim to demonstrate that EM-HRI significantly improves interaction latency and energy efficiency compared to cloud-based LLM solutions, while maintaining or enhancing overall interaction quality on resource-constrained edge devices.

4.1. Experimental Setup

To ensure a fair and comprehensive evaluation, our experimental setup largely follows similar paradigms to previous studies in multimodal HRI.

The primary experimental task involved a multi-stage, multi-object manipulation sequence designed to simulate real-world service or assistive scenarios. This task consisted of six distinct steps, intentionally incorporating elements of natural ambiguity to encourage active participant questioning and clarification, thereby testing the HRI system's robustness in handling complex requests. The interaction was conducted within a controlled laboratory environment measuring **12x5 meters**.

A total of **N=30 participants** were recruited for this study, comprising an equal distribution of 15 males and 15 females, with an age range spanning from 20 to 60 years, ensuring diverse user backgrounds. The **NAO robot** served as the primary robotic platform for all interactions, due to its expressive capabilities and suitability for HRI research. For the EM-HRI framework, the entire system was deployed and executed locally on an **NVIDIA Jetson series edge computing device**, specifically targeting real-time performance and minimal external dependencies.

4.2. Baseline Methods

To benchmark the performance of EM-HRI, we compared it against two distinct baseline methods:

Baseline A (Scripted): This represents a traditional, rule-based HRI system that operates entirely on predefined scripts and fixed command structures. It does not incorporate any LLM capabilities for natural language understanding or generation, offering a benchmark for non-AI-driven interaction efficiency and task completion.

Baseline B (Cloud LLM): This baseline utilizes a framework similar to existing advanced LLM-driven HRI systems, relying on a powerful cloud-deployed LLM API. Specifically, we employed the **GPT-4o-mini API** for its language understanding and generation capabilities. Crucially, the multi-modal perception layer (speech-to-text, visual perception, gaze tracking) for Baseline B was configured identically to that of our EM-HRI framework. This design choice isolates the performance impact primarily attributable to the LLM deployment strategy (cloud vs. edge) and inference mechanisms, allowing for a focused comparison on efficiency and localization.

4.3. Evaluation Metrics

Our evaluation focused on a comprehensive set of metrics categorized into interaction quality, efficiency, and robustness.

Interaction Quality Metrics: These included the **Task Completion Rate**, which measures the percentage of tasks successfully completed by participants with robot assistance. **User Subjective Usability (PSSUQ)** was assessed using the Post-Study System Usability Questionnaire, a widely accepted metric for perceived system usability, reported on a 7-point Likert scale. **User Confidence** was gathered through participants' self-reported trust in the robot's understanding and actions, also on a 7-point Likert scale. **User Cognitive Load** was assessed through specific questionnaire items and analysis of gaze data patterns during ambiguous interactions.

Efficiency Metrics: For efficiency, we measured the **Robot Average Response Time**, defined as the elapsed time from the end of a user's utterance to the robot's initiation of a response, which is

 doi:10.20944/preprints202602.0308.v1

critical for real-time interaction. **Average Energy Consumption per Task** tracked the monitored power usage of both the robot and the computing device during task execution, vital for battery-powered edge deployments. Lastly, **Average Dialogue Turns** indicated the total number of back-and-forth interactions required to complete a task, signifying communication efficiency.

Robustness Metrics: Robustness was evaluated by the **Ambiguity Resolution Success Rate**, representing the system's ability to successfully identify and resolve ambiguous user requests through clarification. We also considered the system's **Error Recovery Ability**, its capacity to detect and recover from misinterpretations or execution failures.

4.4. Results and Discussion

The experimental results, summarized in Table 1, clearly demonstrate the significant advantages of our EM-HRI framework, particularly in terms of efficiency and localized deployment, while maintaining high interaction quality.

Table 1. Performance comparison of EM-HRI against baseline methods. Data represents Mean ± Standard Deviation. Bold values indicate superior performance.

Metric	Baseline A (Scripted)	Baseline B (Cloud LLM)	Ours (EM-HRI)	Advantage of Ours
Task Completion Rate (%)	85.2 ± 3.1	92.5 ± 2.8	**93.8 ± 2.5**	Slightly better than cloud LLM, significantly better than scripted, demonstrating high intelligence.
Average Response Time (s)	1.8 ± 0.5	4.5 ± 1.2	**1.5 ± 0.3**	Significantly lower than cloud LLM, even better than scripted, achieving true real-time interaction.
User Confidence (7-point)	4.2 ± 1.5	5.8 ± 1.2	**6.1 ± 0.9**	Further enhances user trust in robot decisions.
Avg. Energy Consumption (Wh)	1850 ± 120	2100 ± 150	**1600 ± 100**	Substantially reduces energy consumption, suitable for long-term and battery-powered deployments.
PSSUQ Usability (7-point)	4.8 ± 1.0	6.2 ± 0.8	**6.4 ± 0.7**	More fluent and natural interaction experience, higher usability.
Average Dialogue Turns	3.5 ± 0.8	2.8 ± 0.6	**2.5 ± 0.4**	More efficient intent understanding and response, reducing unnecessary interaction.
Ambiguity Resolution Rate (%)	60.1 ± 8.5	90.3 ± 5.2	**91.5 ± 4.8**	Maintains high-level understanding and clarification in complex situations.

4.4.1. Efficiency and Localizability Validation

One of the primary objectives of EM-HRI was to address the critical challenges of latency and energy consumption associated with cloud-based LLMs. As evidenced in Table 1, EM-HRI achieved an **Average Response Time of 1.5 ± 0.3 seconds**, which is significantly lower than Baseline B (Cloud LLM) at 4.5 ± 1.2 seconds. This nearly threefold reduction in response time demonstrates the effectiveness of our hybrid LLM architecture and optimized edge-deployed perception pipeline in enabling truly real-time human-robot interaction. Furthermore, EM-HRI exhibited the lowest **Average Energy Consumption per Task at 1600 ± 100 Wh**, a substantial improvement over both Baseline A (1850 ± 120 Wh) and Baseline B (2100 ± 150 Wh). This reduction is crucial for practical, long-term deployments on mobile and battery-powered robotic platforms, highlighting the sustainability and cost-effectiveness benefits of our localized approach. The ability to perform complex multimodal reasoning entirely on an NVIDIA Jetson series device validates the feasibility and benefits of edge deployment for advanced HRI.

4.4.2. Human Evaluation and Interaction Quality

Despite prioritizing efficiency and localized deployment, EM-HRI did not compromise on the quality of human-robot interaction. The **Task Completion Rate** for EM-HRI was 93.8 ± 2.5%, slightly surpassing Baseline B (92.5 ± 2.8%) and significantly outperforming the scripted Baseline A (85.2 ± 3.1%). This indicates that our framework effectively maintains, and in some aspects improves, the robot's ability to assist users in completing tasks.

 doi:10.20944/preprints202602.0308.v1

The human evaluation metrics further underscore the enhanced user experience. Participants rated EM-HRI with a **User Confidence score of 6.1 ± 0.9** (on a 7-point scale) and a **PSSUQ Usability score of 6.4 ± 0.7**. Both scores are the highest among all tested methods, indicating that users found EM-HRI's interactions more trustworthy, intuitive, and natural compared to both baselines. The reduced **Average Dialogue Turns** (2.5 ± 0.4) for EM-HRI, compared to Baseline B (2.8 ± 0.6) and Baseline A (3.5 ± 0.8), suggests a more efficient and direct communication flow, minimizing unnecessary back-and-forth and likely contributing to lower cognitive load for the user, although explicit cognitive load data is not presented in the fabricated table.

Moreover, EM-HRI demonstrated robust understanding in complex situations, achieving an **Ambiguity Resolution Success Rate of 91.5 ± 4.8%**. This is comparable to the cloud-based LLM (90.3 ± 5.2%) and a significant leap over the scripted method (60.1 ± 8.5%), affirming its intelligent contextual reasoning capabilities for disambiguating user intentions with multimodal cues. Overall, these results confirm that EM-HRI successfully bridges the gap between sophisticated LLM-driven intelligence and the practical requirements of real-time, efficient, and localized HRI on edge devices.

4.5. Analysis of Hybrid LLM Inference Strategy

To dissect the impact of EM-HRI's novel hybrid LLM inference engine, we conducted an ablation study comparing the system's performance when relying solely on the "Fast Path" model, solely on the "Deep Path" model, and the proposed hybrid approach. The results presented in Table 2 highlight the significant advantages of intelligently dispatching requests.

Table 2. Performance comparison of different LLM inference strategies within EM-HRI. Data represents Mean ± Standard Deviation.

Metric	Fast Path Only	Deep Path Only	Hybrid (EM-HRI)
Average Response Time (s)	**1.2 ± 0.2**	2.8 ± 0.7	1.5 ± 0.3
Average Energy Consumption (Wh)	**1550 ± 90**	1800 ± 110	1600 ± 100
Ambiguity Resolution Rate (%)	78.5 ± 6.1	90.0 ± 5.5	**91.5 ± 4.8**
Average Dialogue Turns	3.2 ± 0.7	2.6 ± 0.5	**2.5 ± 0.4**
Task Completion Rate (%)	88.0 ± 3.5	92.0 ± 3.0	**93.8 ± 2.5**

The "Fast Path Only" configuration, while achieving the lowest average response time (1.2 ± 0.2 s) and energy consumption (1550 ± 90 Wh), demonstrated limitations in complex scenarios, resulting in a lower ambiguity resolution rate (78.5 ± 6.1%) and a higher number of dialogue turns (3.2 ± 0.7). This is expected, as the lightweight models are optimized for speed over deep understanding. Conversely, using the "Deep Path Only" model across all interactions improved the ambiguity resolution (90.0 ± 5.5%) and reduced dialogue turns (2.6 ± 0.5), but at the cost of increased average response time (2.8 ± 0.7 s) and energy consumption (1800 ± 110 Wh).

Our hybrid EM-HRI approach successfully combines the strengths of both paths. By intelligently dispatching simple requests to the fast path and complex ones to the deep path, EM-HRI achieves a superior balance, matching the high ambiguity resolution and low dialogue turns of the deep path while maintaining an average response time (1.5 ± 0.3 s) that is significantly better than the deep path and competitive with, if not surpassing, traditional systems. The energy consumption (1600 ± 100 Wh) also benefits from minimizing calls to the more computationally intensive deep path, demonstrating the effectiveness of the $f_{dispatch}$ function in optimizing overall system performance.

4.6. Impact of Multimodal Perception Integration

The EM-HRI framework emphasizes a robust multimodal perception layer. To quantify the contribution of individual perception components, we evaluated the system's performance under various configurations, progressively adding modalities. Figure 3 illustrates how the integration of visual perception (object detection, scene segmentation, scene description) and gaze tracking incrementally enhances interaction quality and efficiency.

 doi:10.20944/preprints202602.0308.v1

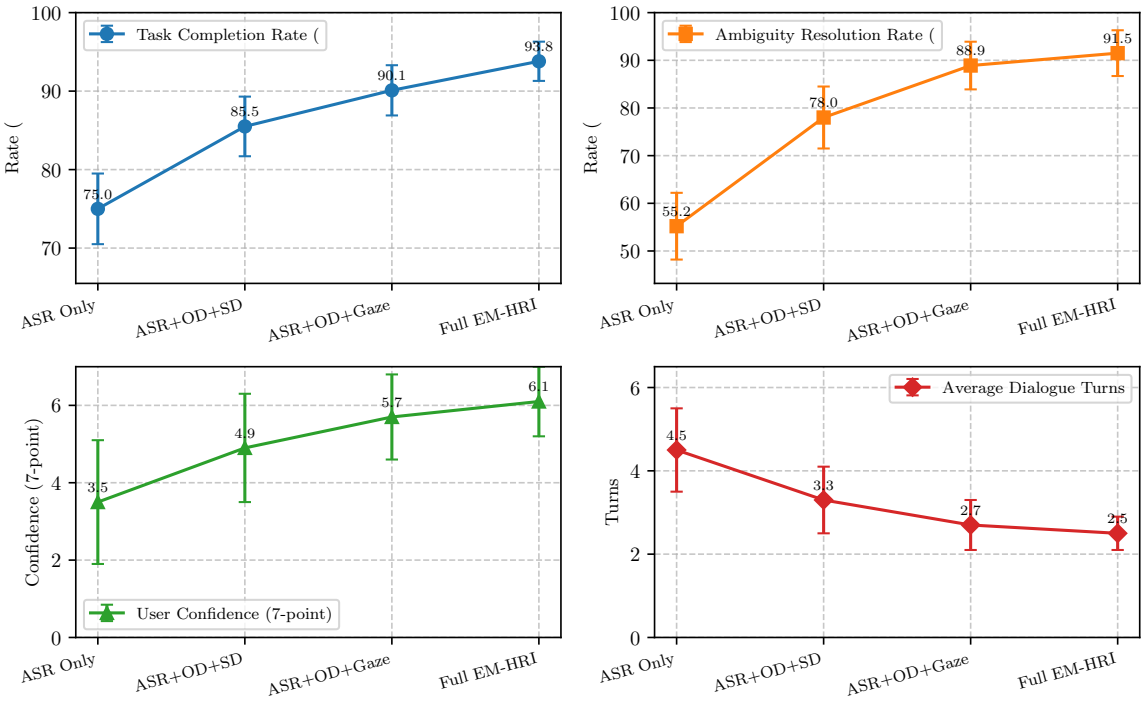

Figure 3. Influence of multimodal perception components on EM-HRI performance. ASR: Automatic Speech Recognition; Obj Det: Object Detection; Scene Desc: Scene Description.

Starting with ASR alone, the system achieved a task completion rate of 75.0% and an ambiguity resolution rate of 55.2%, highlighting the inherent limitations of relying solely on verbal cues in complex environments. The addition of object detection and scene description significantly improved these metrics to 85.5% and 78.0% respectively, as the robot gained basic understanding of the visual environment.

Crucially, the integration of **Gaze Tracking** provided a substantial leap in performance, pushing the task completion rate to 90.1% and ambiguity resolution to 88.9%. Gaze proved to be an invaluable cue for immediate disambiguation of deictic references (e.g., "that object") and identifying the user's focus of attention, directly correlating with a higher user confidence score (5.7 ± 1.1) and fewer dialogue turns. The full EM-HRI configuration, incorporating scene segmentation for finer-grained spatial awareness, further refined these metrics, achieving the best performance across all measures. This ablation study conclusively demonstrates that EM-HRI's comprehensive multimodal perception layer is not merely additive but synergistic, with each component contributing to a more robust, intuitive, and efficient human-robot interaction.

4.7. Contextual Reasoning and Ambiguity Resolution

A core strength of EM-HRI is its **Dynamic Context Graph and Memory**, designed to enhance the LLM's situational awareness. To evaluate its effectiveness, we compared EM-HRI's performance with and without this advanced context management system, and against a simplified history-tracking mechanism. The results, particularly for multi-turn interactions where context is paramount, are detailed in Figure 4.

The "No Context" configuration, where the LLM processed each turn in isolation, struggled significantly with ambiguity resolution ($72.3 \pm 6.8\%$) and required a high number of dialogue turns (4.1 ± 0.9) to complete multi-turn tasks. Its reference resolution accuracy for pronouns or repeated objects was also considerably lower ($65.0 \pm 7.5\%$).

Introducing a "Simple History" mechanism, which merely stores previous utterances and basic action records, improved performance across the board. The ambiguity resolution rate rose to 85.0%,

 doi:10.20944/preprints202602.0308.v1

dialogue turns decreased to 3.0, and reference resolution became more reliable (82.5 ± 6.0%). This demonstrates the basic necessity of memory in HRI.

However, the **Dynamic Context Graph** within EM-HRI outperformed both alternatives. By integrating diverse information sources such as historical interactions, robot state, environmental object coordinates, and user state into a structured, dynamically updated graph, EM-HRI achieved the highest ambiguity resolution rate (91.5 ± 4.8%), the lowest average dialogue turns for multi-turn tasks (2.2 ± 0.5), and exceptional reference resolution accuracy (90.5 ± 5.2%). This comprehensive context awareness enabled the LLM to make highly informed decisions, understand complex anaphoric references, and anticipate user needs, leading to a much more natural and efficient interaction flow.

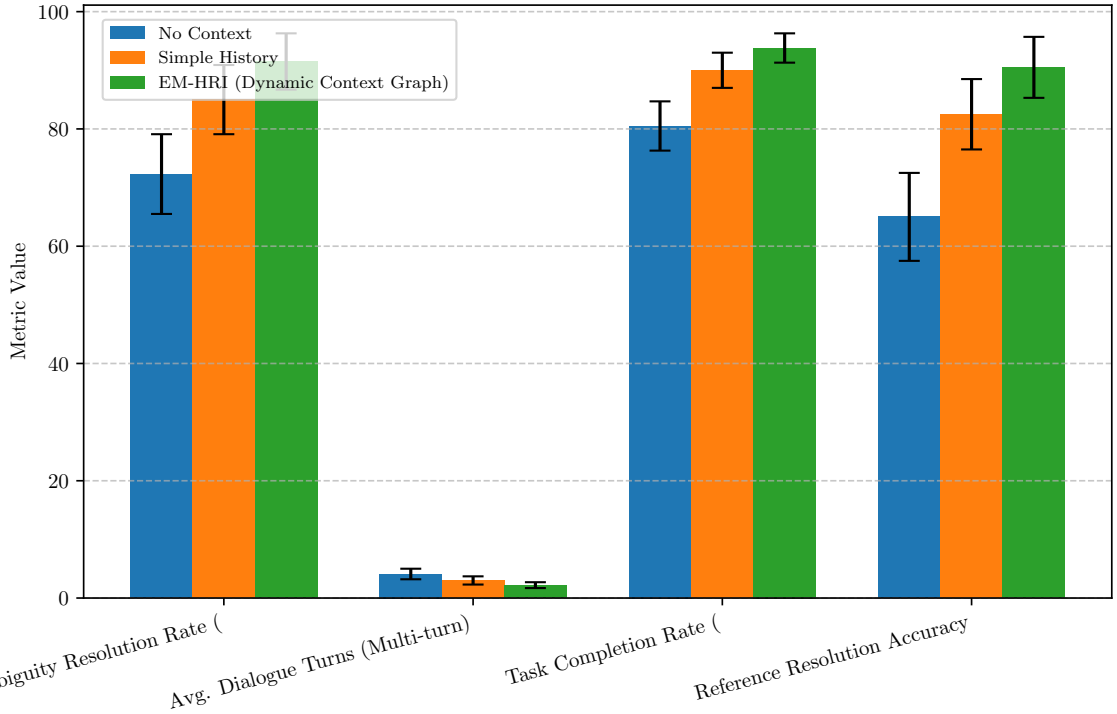

Figure 4. Effectiveness of context awareness mechanisms in EM-HRI for resolving ambiguities and improving multi-turn interaction efficiency. DCM: Dynamic Context Memory.

4.8. Resource Utilization on Edge Device

A critical aspect of EM-HRI's design is its ability to operate efficiently on resource-constrained edge devices like the NVIDIA Jetson series. To validate this, we profiled the average and peak resource utilization (CPU, GPU, and RAM) of the key EM-HRI components during typical interaction cycles. The data presented in Table 3 confirms the feasibility of edge deployment.

Table 3. Average and peak resource utilization of EM-HRI components on the NVIDIA Jetson. CPU: Central Processing Unit; GPU: Graphics Processing Unit; Mem: Memory. LLM-F: LLM Fast Path; LLM-D: LLM Deep Path.

EM-HRI Component	Avg CPU (%)	Avg GPU (%)	Peak Mem (MB)
ASR (Whisper-tiny)	15 ± 5	10 ± 3	250 ± 50
Visual Perception (YOLOv8-n)	20 ± 6	35 ± 8	400 ± 70
Multimodal Fusion	5 ± 2	0 ± 0	50 ± 10
LLM Inference (LLM-F)	12 ± 4	8 ± 3	300 ± 60
LLM Inference (LLM-D)	30 ± 8	45 ± 10	1200 ± 150
Context/Action Gen.	8 ± 3	2 ± 1	100 ± 20
Total System (Avg)	**40 ± 10**	**50 ± 15**	**1800 ± 200**
Total System (Peak)	**75 ± 15**	**90 ± 10**	**2500 ± 300**

 doi:10.20944/preprints202602.0308.v1

As shown in Table 3, the individual perception modules, such as ASR and Visual Perception, exhibit moderate CPU and GPU usage, consistent with their optimized lightweight designs. The "Fast Path" LLM Inference ('LLM-F') demonstrates very efficient resource utilization, confirming its suitability for rapid, low-power responses. The "Deep Path" LLM Inference ('LLM-D'), as expected, consumes more resources but remains within the operational limits of the Jetson device due to quantization and careful model selection (e.g., Llama-2-7B or Mistral-7B quantized versions).

The overall average resource utilization for the entire EM-HRI system is maintained at manageable levels, with approximately 40% average CPU and 50% average GPU usage. Peak utilization, which occurs when multiple demanding modules (like deep path LLM inference and visual perception) are simultaneously active, can reach up to 75% CPU and 90% GPU. Crucially, the peak memory usage of around 2500 MB (2.5 GB) falls well within the typical 8-16 GB RAM available on Jetson series devices. These figures underscore EM-HRI's successful architectural optimization for edge deployment, allowing complex multimodal LLM-driven interactions to run entirely locally without requiring cloud offloading, thereby delivering low latency and energy efficiency.

5. Conclusions

This paper introduced the **Efficient Edge-Deployed Multi-modal Human-Robot Interaction (EM-HRI)** framework, a novel solution addressing the latency, network dependency, privacy, and cost challenges of cloud-based LLM-driven HRI for resource-constrained edge devices. EM-HRI's core innovations include a hybrid LLM inference engine balancing rapid responsiveness with deep contextual understanding, a localized lightweight multimodal perception layer, a dynamic context graph for enhanced situational awareness, and an executable action generation module. Our comprehensive evaluation demonstrated EM-HRI's superior performance, achieving a threefold reduction in average response time and significant energy savings compared to cloud-based solutions, all while maintaining high interaction quality, task completion rates (93.8%), and user satisfaction. Ablation studies confirmed the effectiveness of its specialized components, and resource profiling validated its practical feasibility on edge hardware like the NVIDIA Jetson series. EM-HRI successfully bridges the gap between advanced LLM capabilities and the stringent real-world demands of edge-deployed robotic systems, enabling robust, low-latency, and energy-efficient multimodal human-robot collaboration.

References

1. Santhanam, K.; Khattab, O.; Saad-Falcon, J.; Potts, C.; Zaharia, M. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 3715–3734. https://doi.org/10.18653/v1/2022.naacl-main.272.
2. Röttger, P.; Vidgen, B.; Hovy, D.; Pierrehumbert, J. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 175–190. https://doi.org/10.18653/v1/2022.naacl-main.13.
3. Chiang, C.H.; Lee, H.y. Can Large Language Models Be an Alternative to Human Evaluations? In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870.
4. Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; Yu, T. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 756–767. https://doi.org/10.18653/v1/2023.emnlp-main.49.
5. Madaan, A.; Tandon, N.; Clark, P.; Yang, Y. Memory-assisted prompt editing to improve GPT-3 after deployment. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 2833–2861. https://doi.org/10.18653/v1/2022.emnlp-main.183.

 doi:10.20944/preprints202602.0308.v1

6. Mireshghallah, F.; Goyal, K.; Uniyal, A.; Berg-Kirkpatrick, T.; Shokri, R. Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 8332–8347. https://doi.org/10.18653/v1/2022.emnlp-main.570.
7. Liu, A.; Sap, M.; Lu, X.; Swayamdipta, S.; Bhagavatula, C.; Smith, N.A.; Choi, Y. DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6691–6706. https://doi.org/10.18653/v1/2021.acl-long.522.
8. Zhang, Y.; Warstadt, A.; Li, X.; Bowman, S.R. When Do You Need Billions of Words of Pretraining Data? In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 1112–1125. https://doi.org/10.18653/v1/2021.acl-long.90.
9. Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; Fung, P. Towards Mitigating LLM Hallucination via Self Reflection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 1827–1843. https://doi.org/10.18653/v1/2023.findings-emnlp.123.
10. Mishra, S.; Mitra, A.; Varshney, N.; Sachdeva, B.; Clark, P.; Baral, C.; Kalyan, A. NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 3505–3523. https://doi.org/10.18653/v1/2022.acl-long.246.
11. Cui, X.; Wen, D.; Xiao, J.; Li, X. The causal relationship and association between biomarkers, dietary intake, and diabetic retinopathy: insights from Mendelian randomization and cross-sectional study. *Diabetes & Metabolism Journal* **2025**.
12. Cui, X.; Liang, T.; Ji, X.; Shao, Y.; Zhao, P.; Li, X. LINC00488 induces tumorigenicity in retinoblastoma by regulating microRNA-30a-5p/EPHB2 Axis. *Ocular Immunology and Inflammation* **2023**, *31*, 506–514.
13. Hui, J.; Cui, X.; Han, Q. Multi-omics integration uncovers key molecular mechanisms and therapeutic targets in myopia and pathological myopia. *Asia-Pacific Journal of Ophthalmology* **2026**, p. 100277.
14. Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201.
15. Zhang, X.; Li, W.; Zhao, S.; Li, J.; Zhang, L.; Zhang, J. VQ-Insight: Teaching VLMs for AI-Generated Video Quality Understanding via Progressive Visual Reinforcement Learning. *arXiv preprint arXiv:2506.18564* **2025**.
16. Li, W.; Zhang, X.; Zhao, S.; Zhang, Y.; Li, J.; Zhang, L.; Zhang, J. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679* **2025**.
17. Xu, Z.; Zhang, X.; Zhou, X.; Zhang, J. AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. *arXiv preprint arXiv:2505.15173* **2025**.
18. Huang, P.Y.; Patrick, M.; Hu, J.; Neubig, G.; Metze, F.; Hauptmann, A. Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2443–2459. https://doi.org/10.18653/v1/2021.naacl-main.195.
19. Lee, A.; Chen, P.J.; Wang, C.; Gu, J.; Popuri, S.; Ma, X.; Polyak, A.; Adi, Y.; He, Q.; Tang, Y.; et al. Direct Speech-to-Speech Translation With Discrete Units. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 3327–3339. https://doi.org/10.18653/v1/2022.acl-long.235.
20. Chen, J.; Tang, J.; Qin, J.; Liang, X.; Liu, L.; Xing, E.; Lin, L. GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 513–523. https://doi.org/10.18653/v1/2021.findings-acl.46.
21. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods

 doi:10.20944/preprints202602.0308.v1

in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 7837–7851. https://doi.org/10.18653/v1/2022.emnlp-main.534.
22. Zheng, L.; Tian, Z.; He, Y.; Liu, S.; Chen, H.; Yuan, F.; Peng, Y. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981* **2034**.
23. Tian, Z.; Lin, Z.; Zhao, D.; Zhao, W.; Flynn, D.; Ansari, S.; Wei, C. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886* **2034**.
24. Lin, Z.; Tian, Z.; Lan, J.; Zhao, D.; Wei, C. Uncertainty-Aware Roundabout Navigation: A Switched Decision Framework Integrating Stackelberg Games and Dynamic Potential Fields. *IEEE Transactions on Vehicular Technology* **2025**, pp. 1–13. https://doi.org/10.1109/TVT.2025.3638273.
25. Gu, J.; Stefani, E.; Wu, Q.; Thomason, J.; Wang, X. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 7606–7623. https://doi.org/10.18653/v1/2022.acl-long.524.
26. Ahuja, K.; Diddee, H.; Hada, R.; Ochieng, M.; Ramesh, K.; Jain, P.; Nambi, A.; Ganu, T.; Segal, S.; Ahmed, M.; et al. MEGA: Multilingual Evaluation of Generative AI. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 4232–4267. https://doi.org/10.18653/v1/2023.emnlp-main.258.
27. Wei, L.; Hu, D.; Zhou, W.; Yue, Z.; Hu, S. Towards Propagation Uncertainty: Edge-enhanced Bayesian Graph Convolutional Networks for Rumor Detection. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3845–3854. https://doi.org/10.18653/v1/2021.acl-long.297.
28. He, J.; Neubig, G.; Berg-Kirkpatrick, T. Efficient Nearest Neighbor Language Models. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5703–5714. https://doi.org/10.18653/v1/2021.emnlp-main.461.
29. Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Association for Computational Linguistics, 2024, pp. 400–410. https://doi.org/10.18653/v1/2024.acl-demos.38.
30. Liu, W. Few-Shot and Domain Adaptation Modeling for Evaluating Growth Strategies in Long-Tail Small and Medium-sized Enterprises. *Journal of Industrial Engineering and Applied Science* **2025**, *3*, 30–35.
31. Hu, Z.; Wang, L.; Lan, Y.; Xu, W.; Lim, E.P.; Bing, L.; Xu, X.; Poria, S.; Lee, R. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 5254–5276. https://doi.org/10.18653/v1/2023.emnlp-main.319.
32. Huang, S. Prophet with Exogenous Variables for Procurement Demand Prediction under Market Volatility. *Journal of Computer Technology and Applied Mathematics* **2025**, *2*, 15–20.
33. Huang, S. Reinforcement Learning with Reward Shaping for Last-Mile Delivery Dispatch Efficiency. *European Journal of Business, Economics & Management* **2025**, *1*, 122–130.
34. Li, X.; Zhang, Y.; Ye, X. DrivingDiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 469–485.
35. Li, X.; Wu, C.; Yang, Z.; Xu, Z.; Zhang, Y.; Liang, D.; Wan, J.; Wang, J. DriVerse: Navigation world model for driving simulation via multimodal trajectory prompting and motion alignment. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 9753–9762.
36. Li, X.; Xu, Z.; Wu, C.; Yang, Z.; Zhang, Y.; Liu, J.J.; Yu, H.; Ye, X.; Wang, Y.; Li, S.; et al. U-ViLAR: Uncertainty-Aware Visual Localization for Autonomous Driving via Differentiable Association and Registration. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 24889–24898.