Towards MeluBot: A Multimodal AI Agent Integrating Text, Voice, Image, and Automation for Education and Health

Gabriel Henrique Alencar Medeiros

SeaFortress / INSA Rouen Normandie

DOI: https://dx.doi.org/10.51584/IJRIAS.2025.10100000123

Received: 20 October 2025; Accepted: 26 October 2025; Published: 13 November 2025

ABSTRACT

This paper presents MeluBot, a multimodal AI agent that integrates text, voice, and image modalities, combined with workflow automation, for interactive applications in education and healthcare. We describe the architectural design, enabling technologies, use-case scenarios, and discuss the potential, lim- itations, and future directions. We also position MeluBot with respect to related work in multimodal agents and intelligent tutoring or medical assistants.

INTRODUCTION

In recent years, the rapid evolution of large language models (LLMs) and vision–language models (VLMs) has catalyzed a new generation of artificial intelligence (AI) systems capable of interacting with humans in increasingly natural and adaptive ways. These advances have enabled the development of **multi- modal agents**—systems that process and generate information across several sensory modalities such as text, voice, and image, and that can reason about and act upon the surround- ing environment. Unlike traditional unimodal conversational agents, which rely exclusively on textual dialogue, multimodal systems are able to integrate visual and auditory cues, leading to more robust understanding and richer contextual grounding of user intent.

The convergence between multimodal reasoning and **au- tomation frameworks** is equally transformative. When a multimodal agent is endowed with the ability to trigger ex- ternal workflows, execute code, or access external databases, it transcends static question–answer paradigms and becomes an active participant in a digital ecosystem. Workflow or- chestration tools such as **n8n** or Node-RED enable low- code integration between the agent and third-party services, bridging the gap between perception, reasoning, and action. This combination of multimodal cognition and automation constitutes the foundation of a new research direction in human–AI collaboration.

In this context, we introduce **MeluBot**, an embodied conver- sational agent that integrates text, voice, and image modalities with workflow automation. The system is designed to oper- ate in socially meaningful domains—most notably **education** and **healthcare**—where natural interaction and contextual understanding are essential. MeluBot relies on a modular architecture that fuses multiple input modalities, interprets them through a reasoning layer powered by large language and vision models, and dynamically interacts with external systems through automated workflows. The agent is repre- sented through a 3D avatar capable of verbal and non-verbal communication, fostering engagement and empathy in end users.

The design of MeluBot addresses a broader research ques- tion: *how can multimodal AI agents leverage automation pipelines to achieve persistent, context-aware reasoning in human-centric environments?* This inquiry sits at the inter- section of natural language understanding, computer vision, speech technologies, and software automation, creating an ex- perimental space that unifies cognitive modeling with practical implementation. The prefix “Towards” in the title reflects the ongoing nature of this exploration; the work presented here constitutes an initial step toward a comprehensive multimodal ecosystem for real-world educational and medical support.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

The motivation behind this research arises from several converging trends. First, the post-pandemic acceleration of remote education and telehealth services has revealed the limitations of purely textual or video-based interfaces. Users increasingly demand systems capable of perceiving emotions, interpreting gestures, and maintaining context across multiple sessions. Second, advances in generative AI have reduced the technical barriers to deploying such agents, allowing smaller organizations and research teams to experiment with multimodal fusion. Finally, workflow automation platforms have matured to a point where they can reliably manage asynchronous events, external data retrieval, and secure inter- service communication, making them ideal backbones for intelligent assistants.

Through the development of MeluBot, we aim to contribute both a conceptual framework and an experimental platform for **multimodal automation**. By combining perception, reason- ing, and actuation within a unified system, MeluBot demon- strates how modern AI can move from passive information retrieval toward active collaboration. The framework allows an agent to not only understand complex multimodal queries but also to autonomously trigger sequences of actions that produce tangible outcomes—from scheduling an appointment to generating personalized feedback or compiling a medical summary.

The main contributions of this paper can thus be summa- rized as follows:

1. We propose an architectural blueprint for a multimodal agent that tightly couples text, voice, and vision process- ing with an automation pipeline.
2. We demonstrate its applicability in two critical domains— education and healthcare—showing how multimodality enhances engagement and contextual accuracy.
3. We discuss the challenges that arise from multimodal fu- sion, latency constraints, robustness, privacy, and ethical considerations, outlining potential future directions for research and deployment.

The remainder of this paper is organized as follows. Sec- tion II reviews related work in multimodal agents and automa- tion frameworks. Section III details the system architecture of MeluBot, including perception, fusion, and workflow integra- tion. Section IV presents application scenarios in education and healthcare. Section V discusses challenges and opportunity- ties, and Section VI concludes with directions for future work.

RELATED WORK

The development of multimodal intelligent agents is rapidly emerging as one of the central research directions in artificial intelligence. The fusion of perception, reasoning, and action across modalities represents a shift from static model inference to continuous, embodied interaction with the environment. This trend aligns with the growing literature on **agentic AI**, which seeks to endow models with autonomy, goal orientation, and tool-use capabilities.

A comprehensive survey by [1] explores the design space of **large multimodal agents**, providing an overview of systems that integrate text, image, audio, and video processing into unified architectures. The authors identify key challenges in multimodal alignment, temporal reasoning, and adaptive control, all of which are directly relevant to MeluBot's devel- opment. They also emphasize that future agents must move beyond mere perception toward autonomous coordination and task execution—an idea central to MeluBot's combination of multimodal understanding and workflow automation.

Complementary insights come from the survey [2], which reviews the emerging field of agentic AI and multimodal interaction. This work highlights the convergence of three major paradigms: large language models as cognitive cores, multimodal fusion as perceptual grounding, and automation or tool use as motor control. MeluBot aligns with this syn- thesis by combining perception and reasoning with actionable outputs via external workflow orchestration.

In the **healthcare domain**, multimodal learning has become a cornerstone of diagnostic and decision-support systems. The **HAIM framework** introduced by Soenksen et al. [3] illus- trates how integrating heterogeneous data—such as radiology images, clinical notes, and structured patient records—can improve diagnostic accuracy

and interpretability. Their work demonstrates that combining modalities provides complemen- tary signals that enrich decision-making. Similarly, Krones et al. [4] provide an extensive review of fusion techniques and missing-data handling methods in multimodal healthcare models, underscoring both the promise and the technical com- plexity of such systems. MeluBot inherits these principles by merging vision, speech, and text to deliver coherent, context- sensitive interactions in medical or educational settings.

Recent progress has also been made in developing **in- teractive medical agents** capable of visual reasoning. The **AMIE system** [5] exemplifies an agent that combines visual analysis with diagnostic dialogue, leveraging medical imaging and natural language conversation to support clinical decision- making. These advances validate the feasibility of multimodal medical dialogue but remain confined to domain-specific con- texts. In contrast, MeluBot aims for generalizability, providing a flexible framework that can adapt to multiple domains through modular workflows.

From a methodological standpoint, multimodal agents have evolved to include explicit **tool-use capabilities**. The work on *multi-modal agent tuning* [6] shows that vision-language models (VLMs) can learn to invoke external tools such as search engines, databases, or calculators, improving their ef- fectiveness in real-world scenarios. Similarly, Chen et al. [7] explore *embodied decision-making* through multi-modal large language models (MLLMs), enabling them to perceive, reason, and act within simulated environments. These developments provide foundational insights into how MeluBot leverages automation pipelines to transform reasoning into concrete actions, thereby reducing human cognitive load.

In the **education domain**, multimodal learning agents are gaining traction as intelligent tutoring systems (ITS) capable of interpreting multimodal input from learners—voice tone, facial expressions, handwriting—and providing adaptive feedback. Early prototypes integrate generative AI with voice and gesture recognition to create interactive virtual tutors capable of guid- ing students through problem-solving exercises. Such systems, however, are often limited by the lack of integration with work- flow automation and back-end educational platforms. MeluBot addresses this limitation by coupling its conversational and perceptual modules with an automation layer (e.g., n8n) that can schedule lessons, evaluate student progress, or generate personalized reports.

Beyond these domain-specific works, a broader set of studies examines the theoretical underpinnings of multimodal cognition, highlighting how fusion mechanisms can enhance interpretability, grounding, and robustness. Recent literature in multi-agent AI [8] also explores collaborative frameworks where multiple specialized agents cooperate, exchange con- text, and collectively reason over multimodal information streams. Such paradigms foreshadow future extensions of MeluBot, where multiple instances or sub-agents could operate in coordination to manage complex educational or medical workflows.

In summary, existing research provides valuable foundations for multimodal perception, dialogue, and automation, yet no integrated framework has been proposed that explicitly unites these elements in a general-purpose agentic system. MeluBot seeks to fill this gap by bridging three complementary re- search threads: (1) multimodal input processing for natural interaction, (2) LLM-driven reasoning for semantic coherence, and (3) workflow-based automation for practical execution.

This synthesis defines MeluBot's unique contribution to the evolving landscape of multimodal and agentic AI.

SYSTEM ARCHITECTURE

Overview

Figures 1 and 2 illustrate the conceptual design of **MeluBot**. The system combines a virtual avatar interface for user interaction (Figure 1) with a workflow automation pipeline (Figure 2) that connects the agent to external services through an orchestration layer. Together, these components enable the agent to perceive, understand, reason, and act upon multimodal input in real time.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

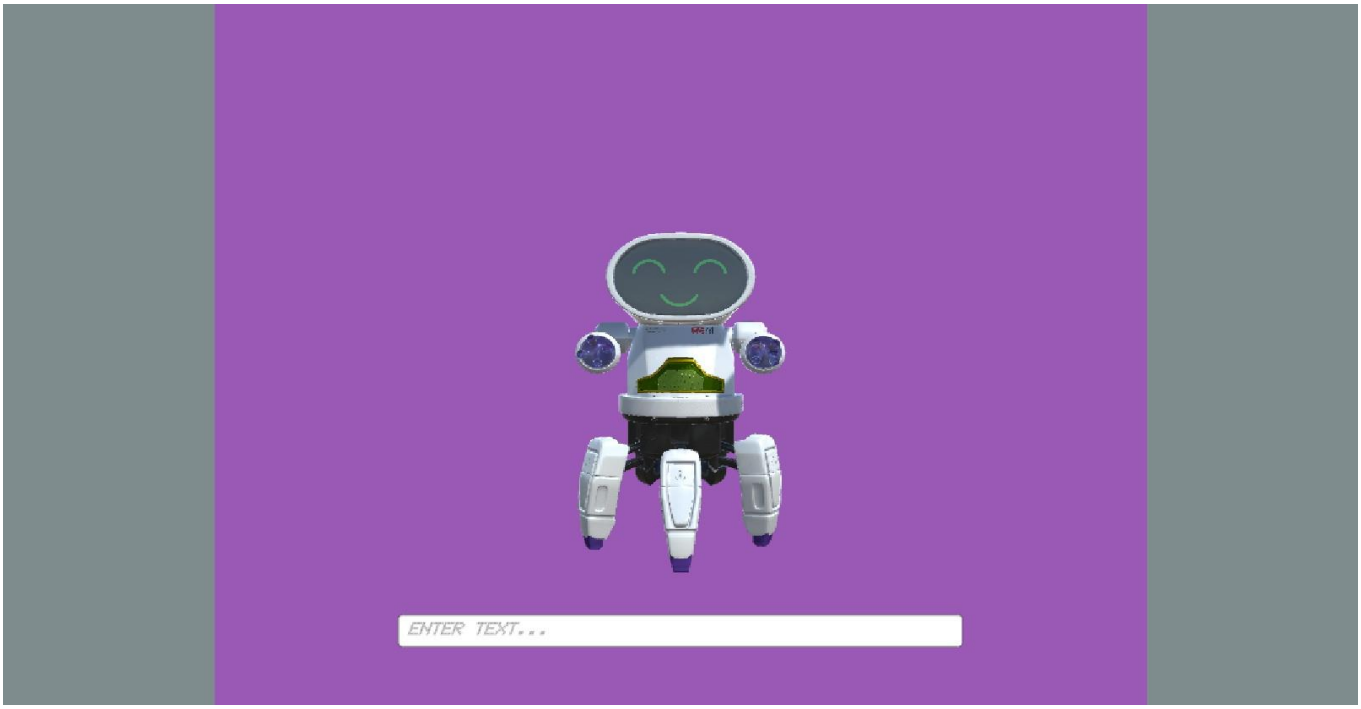

Fig. 1. Placeholder: 3D avatar and interface of MeluBot, showing speech, gesture, and visual interaction.

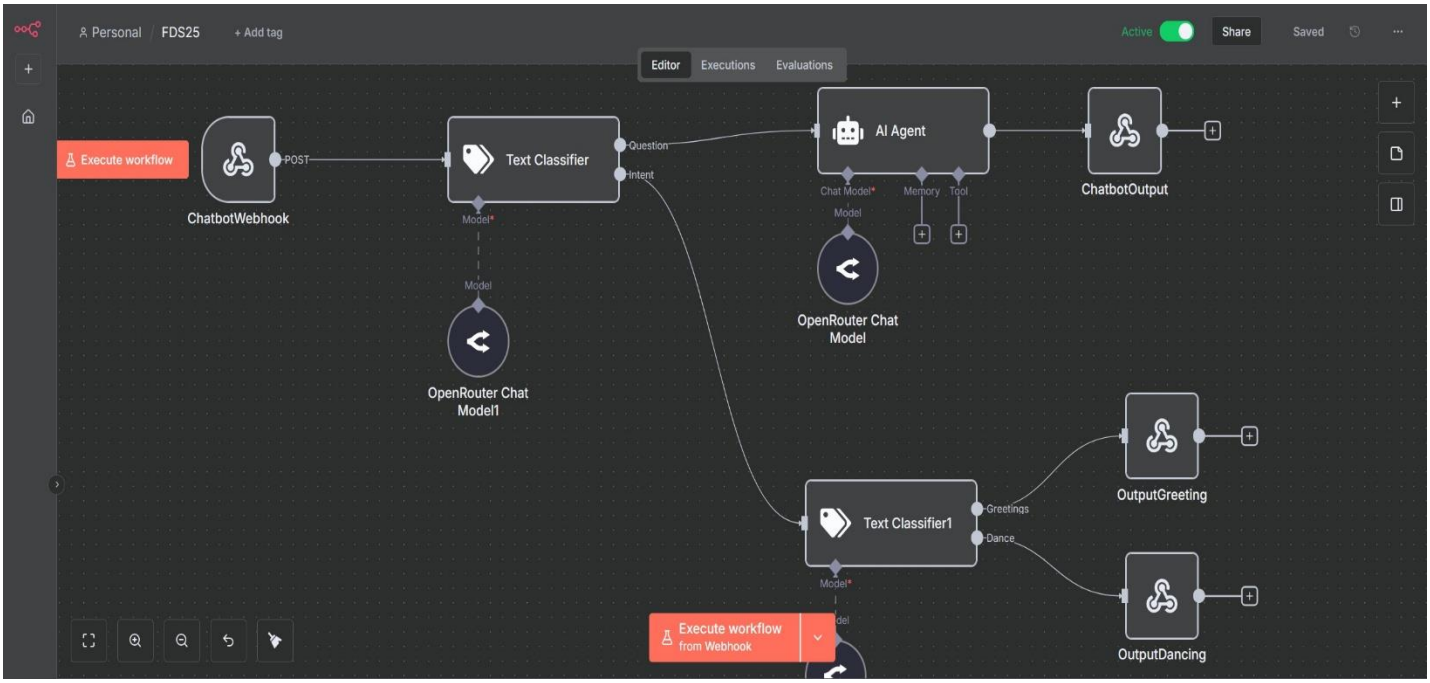

Fig. 2. Placeholder: Workflow automation pipeline (e.g., n8n orchestration) integrated with the agent.

At a conceptual level, MeluBot follows a layered architec- ture inspired by cognitive models of perception and action. The system is divided into five major components: (1) **Input Pro- cessing / Perception**, (2) **Multimodal Fusion / Reasoning**,

(3) **Dialogue & Generation**, (4) **Tool / Automation Con- nectors**, and (5) **State & Memory Module**. These modules operate sequentially yet asynchronously, forming a continuous feedback loop that allows dynamic adaptation to user behavior and contextual changes.

Input Processing and Perception

The perception layer is responsible for acquiring and pre- processing data from multiple sensory channels. It serves as the foundation of MeluBot's understanding capability.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

Text/NLP. Textual input is processed through a natural language understanding (NLU) pipeline that includes tok- enization, part-of-speech tagging, and semantic role labeling. The agent relies on transformer-based encoders to generate contextual embeddings, which capture both local syntactic dependencies and global semantic meaning. This textual rep- resentation is later aligned with other modalities during fusion. **Automatic Speech Recognition (ASR).** Speech is a pri- mary interaction modality in MeluBot. The ASR module converts spoken language into textual transcriptions using a neural sequence-to-sequence architecture, such as Whisper or Deep Speech. In addition to transcribing content, the ASR component also extracts prosodic features including pitch, volume, and rhythm, which are used to infer emotional tone and user intent. These paralinguistic signals contribute to

adaptive dialogue generation and empathetic responses.

Vision / Image Processing. The visual module processes static images or live video frames using convolutional neural networks (CNNs) or vision transformers (ViTs). This com- ponent performs object detection, semantic segmentation, and visual scene understanding. For healthcare applications, the module may identify anomalies in medical images, while in educational contexts, it can analyze diagrams or handwritten notes submitted by learners. The extracted visual features are projected into a latent embedding space compatible with textual and auditory representations.

The output of this layer consists of synchronized modality embeddings:

$$E = \{E_{text}, E_{speech}, E_{vision}\},$$

each representing the semantic content of its respective chan- nel. These embeddings are forwarded to the fusion module for integration.

Multimodal Fusion and Reasoning

The **fusion module** is the core of MeluBot's intelligence. It integrates heterogeneous representations into a unified context vector that enables cross-modal reasoning. The architecture follows a hybrid fusion strategy combining both early and late fusion techniques.

At the early stage, modality-specific encoders project inputs into a shared embedding space through linear transformations and normalization layers. This ensures dimensional alignment and facilitates attention-based interactions among modalities. At the later stage, a **cross-modal transformer** applies at- tention mechanisms that dynamically weight each modality according to contextual relevance. For example, in a medical diagnostic scenario, visual features might dominate when analyzing an image of a skin lesion, whereas linguistic cues might take precedence during a dialogue about symptoms.

Formally, let X_m denote the embedding of modality m. The fusion operation is modeled as:

$$Z = \text{Attention}(\{X_m\}; W_m),$$

where W_m are learnable parameters controlling the relative influence of each modality. The resulting fused vector Z is then passed to the **reasoning and planning module**, which performs intent recognition and determines the appropriate system response or external action. The reasoning stage uses an LLM backbone fine-tuned for dialogue management, retrieval-augmented reasoning, and intent classification.

This hierarchical design ensures that MeluBot is capable not only of understanding multimodal content but also of main- taining coherence and continuity across sessions. The modular nature of the fusion layer also allows new modalities—such as haptic feedback or biometric sensors—to be integrated in future iterations without re-engineering the full system.

Dialogue and Generation

Once the reasoning layer determines the intent and desired response, MeluBot generates multimodal outputs

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

through its dialogue and animation subsystems. The **dialogue generator** uses a language model fine-tuned for conversational flow and domain-specific expression. The generated text can be rendered as both written and spoken output.

For voice output, the system employs a **text-to-speech (TTS)** module that synthesizes natural, expressive speech aligned with the emotional tone detected from the user's input. The 3D avatar synchronizes lip movements and facial expres- sions with the generated audio using blendshape animation techniques. Gesture and posture control modules provide non- verbal cues such as nodding, pointing, or shrugging, improving the realism of interaction and maintaining user engagement.

The dialogue layer is also capable of grounding its responses in external knowledge sources through retrieval augmentation or database queries. This ensures factual reliability and con- sistency, especially in educational and medical contexts where misinformation can be critical.

A. Tool and Automation Connectors

The automation layer extends MeluBot's reasoning capabil- ities beyond conversation. It allows the agent to perform real- world tasks by invoking APIs, managing databases, or exe- cuting workflow sequences. This functionality is implemented through connectors to orchestration platforms such as **n8n**, which support the creation of complex logic flows using a visual interface.

Each action identified by the reasoning module is converted into a structured event following the schema:

⟨intent, parameters, context⟩,

which is sent to the automation service. The workflow engine interprets the event, executes predefined tasks (e.g., schedul- ing, report generation, email notification), and returns the results to the dialogue manager. This bidirectional commu- nication loop transforms MeluBot into an autonomous agent capable of completing multi-step tasks without explicit user guidance.

Such integration also ensures **scalability** and **domain in- dependence**. For instance, educational workflows may con- nect to Learning Management Systems (LMS), while medical workflows may interact with Electronic Health Record (EHR) systems. Because n8n is low-code, new workflows can be added without modifying MeluBot's core, significantly reduce- ing development overhead and encouraging domain-specific customization.

State and Memory Module

MeluBot maintains an internal memory architecture that captures both short-term dialogue state and long-term user profiles. The **episodic memory** stores contextual information from recent interactions (e.g., last query, current topic, detected emotion), while the **semantic memory** aggregates structured knowledge about the user, including preferences, goals, and historical activity.

This dual-memory framework supports continuity and per- sonalization across sessions. For example, in a tutoring sce- nario, MeluBot can recall which exercises a student has completed and adjust the difficulty level accordingly. In a healthcare context, it can retrieve previous symptom reports or check the evolution of a condition based on uploaded images.

To maintain data security and compliance, memory storage is encrypted, and personal identifiers are anonymized or re- placed with hashed references when interfacing with external workflows.

Workflow Integration via n8n

The automation backbone, implemented through **n8n**, acts as the operational bridge between MeluBot and the external environment. When the reasoning module determines that an action must be executed—such as "schedule an appointment," "generate a patient summary," or "create a progress report"—it emits a structured JSON payload that is transmitted to the corresponding workflow.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

Each workflow may include multiple nodes representing API calls, logical conditions, data transformations, or external triggers. For instance, in a healthcare scenario, a workflow could query a medical database, send a summary email to a practitioner, and log the event in a secure server. The results are returned to MeluBot in real time, allowing it to confirm completion and continue the conversation seamlessly.

This architecture enforces a clean separation of concerns: the **multimodal reasoning engine** remains focused on per- ception and understanding, while the **workflow layer** han- dles domain-specific execution. Such decoupling enhances modularity, supports reusability across sectors, and simplifies debugging. Moreover, it allows different instances of MeluBot to share the same automation infrastructure, facilitating de- ployment at scale across educational institutions or clinical networks.

USE-CASE SCENARIOS

Educational Tutoring

A student interacts with MeluBot in a virtual classroom. They can ask a question in speech or text, share an image (e.g., a photo of their handwritten work or a diagram). MeluBot fuses their query and visual context, provides explanations, draws sketches or annotations on the image, and triggers workflows such as grading tasks, sending reports to instructors, or scheduling follow-up lessons.

Additionally, MeluBot can monitor student progress over sessions, adapt difficulty, and proactively suggest exercises. It can automate reminders, feedback emails, or generate sum- mary PDFs of sessions.

Healthcare / Medical Assistance

In a remote consultation scenario, a patient can send an image (e.g., skin lesion, wound photo), describe symptoms via voice, and answer follow-up textual questions. MeluBot can process multimodal input, retrieve relevant medical knowl- edge (via API or plugin), and provide guidance, differential diagnosis suggestions, or escalate to a clinician. It can trigger workflows for appointment booking, sending reminders, or integrating with EHR (Electronic Health Record) systems.

In more monitored settings, MeluBot can periodically re- quest images (e.g., wound progression) or voice check-ins and flag anomalies via automated workflows.

DISCUSSION:

OPPORTUNITIES & CHALLENGES

Opportunities

1. **Rich interaction**: multimodal fusion enables deeper con- textual understanding than unimodal systems.
2. **Scalable automation**: by integrating workflow orches- tration, MeluBot can automate end-to-end processes, in- creasing efficiency in both education and health.
3. **Personalization**: memory, multimodal history, and adap- tive strategies support tailored experiences.
4. **Bridging modalities**: e.g., referencing visual content in conversation, correcting misinterpretations using redun- dant cues.

Challenges

1. **Latency & compute**: real-time fusion and reasoning across modalities can impose heavy resource demands, especially in low-latency conversational settings.
2. **Modality alignment & missing data**: modalities may be asynchronous or missing (e.g., user speaks but sends no image). The system must degrade gracefully.
3. **Robustness & distractions**: multimodal agents may be susceptible to irrelevant environmental cues or adversarial noise [9].
4. **Privacy & security**: especially in healthcare, handling visual, audio, and textual personal data demands

strong encryption, anonymization, and compliance (e.g., HIPAA, GDPR).
5. **Interpretability & trust**: explainable multimodal deci- sions remain an open problem.
6. **Data scarcity & bias**: multimodal training data are harder to gather; biases in one modality can propagate.

Future Directions

1. **Continual & online learning**: adapting the model over time from real interactions.
2. **Multi-agent collaboration**: multiple MeluBots or sub- agents collaborating in classrooms or hospitals [8].
3. **Few-shot adaptation & domain transfer**: applying the model in new settings with minimal data [10].
4. **Embodied / AR interfaces**: integrating mixed reality or gesture inputs along with voice/text/image fusion.
5. **Rigorous evaluation benchmarks**: combining user stud- ies, robustness tests, adversarial scenarios, and modality ablation analyses.

CONCLUSION

We have presented **MeluBot**, a multimodal AI agent ar- chitecture that integrates text, voice, and image inputs with workflow automation, targeting applications in education and healthcare. We believe this approach points **towards** a new generation of interactive, context-aware systems capable of deep, seamless human-AI collaboration. Many open challenges remain—especially in latency, robustness, and trust—but the potential impact in socially relevant domains is significant.

ACKNOWLEDGMENTS.

This work is part of the SeaFortress initiative.

REFERENCES

1. J. Xie, Z. Chen, R. Zhang, X. Wan, and G. Li, "Large multimodal agents:A survey," 2024. [Online]. Available: https://arxiv.org/abs/2402.15116
2. Z. Durante, Q. Huang, N. Wake, R. Gong, J. S. Park, B. Sarkar,R. Taori, Y. Noda, D. Terzopoulos, Y. Choi, K. Ikeuchi, H. Vo,
3. L. Fei-Fei, and J. Gao, "Agent ai: Surveying the horizons of multimodal interaction," 2024. [Online]. Available: https://arxiv.org/abs/2401.03568
4. L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo,L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas, "Integrated multimodal artificial intelligence framework for healthcare applications," *NPJ Digit. Med.*, vol. 5, no. 1, p. 149, Sep. 2022.
5. F. Krones, U. Marikkar, G. Parsons, A. Szmul, and A. Mahdi, "Review of multimodal machine learning approaches in healthcare," 2024. [Online]. Available: https://arxiv.org/abs/2402.02460
6. K. Saab and J. Freyberg, ""amie gains vision: A research ai agent for multimodal diagnostic dialogue"," Blog post, Google Research, May 2025, accessed: YYYY- MM-DD. [Online]. Available: https://research.google/blog/ amie-gains-vision-a-research-ai-agent-for-multi-modal-diagnostic-dialogue/
7. Z. Gao, B. Zhang, P. Li, X. Ma, T. Yuan, Y. Fan, Y. Wu,Y. Jia, S.-C. Zhu, and Q. Li, "Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage," 2025. [Online]. Available: https://arxiv.org/abs/2412.15606
8. L. Chen, Y. Zhang, S. Ren, H. Zhao, Z. Cai, Y. Wang, P. Wang, T. Liu, and B. Chang, "Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond," 2023. [Online]. Available: https://arxiv.org/abs/2310.02071
9. H. Yao, R. Zhang, J. Huang, J. Zhang, Y. Wang, B. Fang, R. Zhu,Y. Jing, S. Liu, G. Li, and D. Tao, "A survey on agentic multimodal large language models," Oct. 2025, version v1; accessed: YYYY-MM-DD. [Online]. Available: https://arxiv.org/abs/2510.10991

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

10. X. Ma, Y. Wang, Y. Yao, T. Yuan, A. Zhang, Z. Zhang, and Zhao, "Caution for the environment: Multimodal llm agents are susceptible to environmental distractions," 2025. [Online]. Available: https://arxiv.org/abs/2408.02544
11. G. Verma, R. Kaur, N. Srishankar, Z. Zeng, T. Balch, andM. Veloso, "Adaptagent: Adapting multimodal web agents with few- shot learning from human demonstrations," 2024. [Online]. Available: https://arxiv.org/abs/2411.13451