FMDB *Transactions on Sustainable Computing Systems*

An Integrated Full-Stack Multimodal AI Workspace Assistant with Hybrid Distributed Inference and Holographic Mixed-Reality Interface

T. Pranavadit[1], N. Sankar Ram[2,*], M. A. Bala Kumar[3], R. V. Vishal[4], R. Vinoth[5], Mahmoud Abuzaid[6], Malak Ismail[7], Manar Messed[8], Mazen Mawla[9]

[1,2,3,4,5]Department of Computer Science and Engineering in AIML, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[6,7,8,9]Department of Computer Science, Zewail City of Science and Technology, October Gardens, Giza Governorate, Egypt.
tspranavadit@gmail.com[1], nsankarram@gmail.com[2], balakumar105062004@gmail.com[3], vishal122004@gmail.com[4], captainvinoth@gmail.com[5], s-mahmoud.abuzaid@zewailcity.edu.eg[6], s-malak.ismail@zewailcity.edu.eg[7], s-manar.messed@zewailcity.edu.eg[8], s-mazen.mawla@zewailcity.edu.eg[9]

Abstract: The paper proposes a prototype of a multimodal AI workspace, W.E.D.N.E.S.D.A.Y., a union of text, voice, gesture, camera vision, and capacitive touch that would be projected onto a tri-layer interactive display via a holographic mixed-reality interface. The system combines multi-LLM routing with cascaded fallback, YOLO-based real-time vision, and a three-path voice pipeline combining Whisper, Gemini Audio, and two TTS engines. UI, application services, AI intelligence, and GPU-based persistence will provide layers that enable real-time interaction and scale. Low-latency inference is experimentally proven with a maximum vision processing rate of 124 FPS and an average response time of 1.5s in a local-cloud deployment. The system presents an agent-coordination system that assists with autonomous web automation, file manipulation, and tool execution. Findings show strong fault tolerance, high throughput, and smooth multimodal fusion of interactive AI workspace applications in real-world computing environments, successfully applied in assistive and industrial situations and deployment environments.

Keywords: Holography Interface; Edge Computing; IoT Workspace; Agent Systems; Gemini Audio; File Manipulation; Deployment Environment; Real-World Computing.

Received on: 02/07/2025, **Revised on:** 27/08/2025, **Accepted on:** 20/11/2025, **Published on:** 10/08/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCS

DOI: https://doi.org/10.69888/FTSCS.2026.000722

Cite as: T. Pranavadit, N. S. Ram, M. A. B. Kumar, R. V. Vishal, R. Vinoth, M. Abuzaid, M. Ismail, M. Messed, and M. Mawla, "An Integrated Full-Stack Multimodal AI Workspace Assistant with Hybrid Distributed Inference and Holographic Mixed-Reality Interface," *FMDB Transactions on Sustainable Computing Systems*, vol. 4, no. 3, pp. 196–209, 2026.

1. Introduction

To illustrate a modern example, one can note artificial intelligence devices such as ChatGPT and Bing Chat that can address a wide range of human concerns and respond instantly to questions. Previously existing AI systems were mainly adapted to interacting with texts and providing conversational support. But even more sophisticated, intelligent, proficient people should be able to do much more than just text. Users need to combine vision, speech, gesture, and physical interaction and project them into a single computational space. Still, most existing AI assistants have a strikingly poor integration of these modalities

*Corresponding author.

into a single system. This is especially important in high-performance applications, such as industrial monitoring, intelligent design environments, healthcare assistance and decision-support systems that support real-time operation, where it is essential to switch between input modalities quickly. To address these challenges, multimodal AI systems have become a high research priority. Various sensor fusion, natural language processing, audio intelligence, and computer vision systems are united to build adaptive and immersive computing environments. Multimodal artificial intelligence can better understand context, intention, and environment as a human does. In turn, such systems are now being considered for applications in smart cities, autonomous systems, collaborative robotics, and augmented reality platforms [1]. The concept of workspace-centric artificial intelligence is attracting an increasing number of interested parties, due to the availability of high-performance edge computing equipment and optimized deep learning infrastructures.

Now, state-of-the-art architectures capable of running on both local and hybrid cloud systems support real-time object detection, speech-to-text conversion, natural language generation, and multimodal reasoning. Existing implementations remain piecemeal, despite advances in technology. Typically, vision systems, speech interfaces, and language models are built as separate modules rather than a common reasoning architecture. This separation leads to significant overhead in real-world deployments due to high latency and reduced coherence. The capabilities of AI systems are also curtailed by architectural fragmentation when they need to operate in environments that require simultaneous processing of multiple modalities of input. The need for a multisource AI framework to integrate with multiple sources only exacerbates the situation in busy cities such as Chennai, where increasing emphasis is placed on intelligent infrastructure. The next generation of smart applications needs to be supported by a coherent system that can operate across diverse network setups and device constraints [2]. Introduction of multimodal AI workspaces. It comprises five basic input channels: text, voice, gesture, camera vision and capacitive touch integration. The goal of W.E.D.N.E.S.D.A.Y. is to harness these. These inputs are then projected onto a holographic interaction plane within a single unified processing frame. In this system, the aim is to transform the desktop from a static display into an intelligent computational space seamlessly coupled with both digital and physical interactions. A feature of the system is its tri-layer, interactive, surface-projection-based UI.

The interface enables interaction with AI-generated content via touch, controlling 3D graphic representations, procedural flows, and live decision-support objects. Such an approach will improve engagement and reduce cognitive distance between intent formulation and the response system. Using this technology, data can be mechanically interacted with in the same manner as other objects within one's surrounding environment. One of the greatest innovations in W.E.D.N.E.S.D.A.Y. is its architecture, which is structured into four main layers: user interface rendering, application service management, AI intelligence processing, and GPU computation with persistent storage integration. Resource utilization is effective, simple to manage, and easy to expand. Each layer is autonomous, with a common interaction system that provides seamless, low-latency, and responsive interaction. One important element of the architecture is the AI intelligence layer, which includes multiple large language models that provide a cascaded fallback. It keeps your system running even in the event of network, API, or cloud failures. The system maintains performance stability and continuous operation in vital applications by switching between AI models as needed. Besides, six YOLO-based models are employed to detect objects at over 120 FPS on a consumer-grade graphics card. Consequently, the system can be used in various applications, such as real-time environmental awareness, industrial monitoring, surveillance, and support for autonomous navigation [3].

Another significant innovation is the agent-based orchestration structure. Taking this framework implies that the system is already capable of engaging in multiple domains independently. This involves web surfing, tracing file-handling systems, and browser automation. Agents operate independently, but are coordinated by a routing system that, in real time, interprets user intent and delegates tasks to… a module! This type of architecture transforms the system into an active conversational assistant capable of executing complex workflows. The agent orchestration layer provides the system with considerable flexibility and adaptability [6]. The system can work proactively instead of just responding to user queries, perform tasks, manage workflows, and adapt to users. This is a twist on conversational AI that turns it into actionable intelligence, where the system is not simply responding to a command but actually executing the task itself. It also has a three-stage processing architecture that improves speech recognition. The structure combines on-premises speech models with on-demand audio-reasoning models in the cloud. Language understanding and contextual reasoning can be performed with high-performance cloud processing. Local models, on the other hand, are applicable in low-connectivity or offline scenarios. This novel mixture ensures that voice interaction remains stable and reliable regardless of hardware or network conditions.

One of the most essential characteristics of the paper W.E.D.N.E.S.D.A.Y. is contextual continuity of memory. This system uses an incremental summarisation process of conversation history that preserves important user intent and decision history. It enables smooth, long-term interaction and does not exceed the model's context constraints. The system will be able to support complex workflows across multiple sessions when it contains structured memory representations. This is viable since it is possible to pause and restart without losing context. To enable this functionality, the system uses an asynchronous, PostgreSQL-based database architecture to store and retrieve multi-modal interaction data. This includes all textual discussions, event-based scene logs, gesture logs, system logs, and environmental logs. To maintain coherence over the long run and to have intelligent

decision-making over an extended period [4]. A holographic interface might constitute a significant paradigm shift in human-computer interaction. The system generates a dynamic, interactive workspace on physical surfaces, unlike traditional screen-based interfaces. The cognitive gap between system capabilities and user goals narrows, making interaction more natural. Users can manipulate digital objects in a form that is likely to be similar to the real world. It helps make interactions more universally accessible, and the learning curve typically associated with complex systems is reduced. Gesture recognition powered by MediaPipe tracking systems makes interactions easier. Users can play with digital objects with their hands without touching anything. This can be particularly useful in environments where touching or interacting is not allowed or may be impossible, such as on an industrial floor, in a laboratory, or in a collaborative workspace.

The hybrid interaction model, derived from letter, gesture, and voice input, is very versatile and responsive. To ensure the complex multimodal architecture is realistic, it would need to be optimized for system performance. The experiments indicate that the system's throughput remains constant at high vision-processing loads. Under all working conditions, latency fluctuations are kept low, while inference times remain consistent. The hybrid deployment model dynamically assigns computational workloads to either a local edge device or a cloud-based application, ensuring scalability without affecting responsiveness. This plan can be quite beneficial in networks that are neither very reliable nor offer access to alternative computers. The system can function smoothly even under stress. Such flexibility is most relevant to new smart city models, as seen in Chennai, where digital infrastructure is quickly adapting to IoT-based automated systems and smart public services [5]. Overall, W.E.D.N.E.S.D.A.Y. represents a significant step forward in the field of multimodal AI workspaces. The system combines computer vision, natural language processing, and other technologies into a single computational system with a specific programmable interface. By linking the physical space to the digital interface, the system will improve productivity, situational awareness, and decision-making efficiency. Moreover, the architecture will offer a scalable platform for advancing technology in the future. The design aims to support state-of-the-art applications such as autonomous intelligent agents, augmented reality interfaces, and others. As artificial intelligence advances toward more embodied, context-driven systems, frameworks such as W.E.D.N.E.S.D.A.Y. will shape the future of intelligent human-computer-interaction environments.

2. Literature Survey

Multimodal artificial intelligence has evolved rapidly, transforming intelligent systems by enabling the integration of diverse information types, including text, images, audio, and sensor signals. Multimodal frameworks combine complementary sources of information, enhancing contextual understanding and decision-making compared with traditional single-modality systems. By integrating LLMs into multimodal pipelines, these systems have become more robust, capable of supporting complex reasoning, generating human-like responses, and adapting to changing environments. Consequently, medical care, human-computer interaction, robotics, and communication networks are some of the most used fields of multimodal AIs. This is the shift researchers often talk about, which defines a trend towards more intelligent and insightful systems or processes that can interpret real-world situations and act accordingly with greater accuracy and greater efficacy. Recent developments have increased the importance of multimodal interaction and adaptive learning techniques for enhancing system performance. Studies show that multimodal active learning frameworks can utilize visual images, speech, and context, thereby enhancing model generalization and interaction quality [6]. The advancement of LLM architecture has further propelled contextual reasoning across many modalities, enabling applications such as span perception and language [7]

The literature on benchmarking in healthcare demonstrates that an LLM-based system can assist medical professionals in decision-making. Nevertheless, reliability and bias remain major challenges [7]. In addition, multimodal techniques have been helpful in medical imaging tasks, improving the accuracy of contouring and diagnosis by combining text with visuals [9]. Together, these findings highlight the importance of the multimodal learning paradigm for improving system usability in practice. Explainability and user trust are valued when establishing multimodal AI, particularly in high-stakes areas. According to recent reports, multimodal AI assistants are being developed to provide greater explainability of systems and their outputs, augmenting users' understanding of system choices [10]. In sectors like healthcare and assistive technologies, reliability and accountability are of utmost importance. For example, in multimodal AI-based memory assistants, contextual reminders have been developed to help older people maintain independence and enhance quality of life [11]. Advanced multi-agent models with enhanced, more pertinent diagnostic outputs have been developed to facilitate easier collaborative reasoning among multiple AIs [12]. Alignment strategies based on real user interactions with the system can improve system behavior and ensure that responses are relevant to the user [13]. The following developments indicate that researchers have to strike a balance between performance and transparency in contemporary AI.

Unlike in the past, when replicating or optimizing a system was difficult, easy replication and optimization are now possible thanks to multimodal AI research. Feedback-driven optimization techniques have been established to significantly enhance the contributions of models that employ large-scale human and AI feedback in loop systems [14]. Models for multimodal intention prediction further improve interaction by leveraging heterogeneous data sources to infer users' goals. This is useful for assistive and interactive applications [15]. Furthermore, multimodal AI in augmented and virtual reality is being reported, including

systems that enable natural interaction in such environments through voice commands, gestures, and environmental context [16]. Researchers have developed lightweight multimodal reasoning frameworks capable of real-time processing of robot information while remaining computationally efficient for deployment in constrained environments [17]. Researchers are overcoming performance and efficiency challenges in multimodal systems. Ongoing research paths are broadening the avenues of multimodal AI through novel interaction paradigms and communication mechanisms. Technologies for brain-computer interfaces represent an important breakthrough in AI, enabling interaction between human neural activity and AI via multiple signals and language models [18].

Next-generation communication networks such as 6G will also incorporate multimodal LLMs to emphasize semantic communication, focusing on conveying information rather than data for efficient transmission, thereby optimizing bandwidth requirements [19]. Research is being conducted on vision-language models for detecting human values and ethics. This is critical for the responsible deployment of AI in a real-world scenario [20]. These achievements are indicative of the far-reaching applications of multimodal AI, not just from a technical perspective but also in addressing the societal and ethical issues posed by intelligent systems [21]. Taken as a whole, multimodal AI – powered by advanced LLMs – is fast becoming the foundation of next-gen intelligent systems [22]. A combination of reasoning, explainability, and data scaling across different modalities is making systems more adaptable, efficient, and user-friendly. While significant progress has been made, challenges related to interpretability, reliability, and ethics remain open research areas. Future research is expected to focus on alignment techniques, multimodal fusion strategies, and more robust techniques for real-world use [23]. As research advances, multimodal AI will shape the future of human-AI interaction and intelligent system design. These technologies enable the deployment of intelligent workplace systems that integrate perception, reasoning, and interaction on a single platform, aligning well with W.E.D.N.E.S.D.A.Y.'s objectives.

3. Methodology

3.1. Layer for Multi-Modal Interpretation

The Multimodal Interpretation Layer is the first stage of the W.E.D.N.E.S.D.A.Y. pipeline, capable of capturing any user input and transforming it into digital signals. The purpose of this layer is to provide support for five main modalities: text, voice, gesture, vision, and capacitive touch for the human–machine interaction environment. Each input type is handled by an adapter module that converts it to machine-readable formats. The text input is streamed over a WebSocket, enabling low-latency communication. Microphone arrays capture voice data, which is processed by speech recognition engines that can handle noise and other changing acoustics. Gesture recognition is done using MediaPipe-based hand tracking. MediaPipe hand tracking extracts 21 key skeletal landmarks per hand. MediaPipe tracks the user's hand and interprets the gesture. Responses to requests are based on vision-powered object detection and a scene understanding pipeline powered by YOLO. These moves enable the system to recognize its context dynamically. A calibrated interactive surface captures capacitive input, enabling precise mapping of touch interactions to X-Y coordinates. The normalization process transforms all heterogeneous inputs into a single JSON schema to improve compatibility. With this layer of technology, real-life interactions are digitized with minimal latency and accuracy. Incorporating the various modalities into a single, structured system enables the system to be more resilient and flexible when switching between them. The incoming data is pre-processed to facilitate streamlining and synchronization, so that when it is transferred to the system, it will be ready for intelligent processing.

3.2. Conditioning and Scaling Signals

The Signal Preprocessing and Normalization Layer is intended to clean the raw multimodal inputs, making the data streams coherent and time-aligned. This step references and evolves the signal, synchronizing it with a reference signal while accounting for impurities and noise in the real-world input. The voice signals are processed to remove noise and enhance speech, resulting in better transcription, especially in noisy, dynamic environments. The sensor data tends to be noisy. To eliminate this noise, researchers use exponential moving average filters. This achieves stabilized gesture tracking. The images are resized and normalized so that the inputs are suitable for the YOLO-based model. This helps ensure consistent inference across all hardware setups. Frame downsampling methods and their applications involve trade-offs between computational efficiency and detection performance.

Capacitive touch input is then calibrated using corner-mapping affine transformation matrices, ensuring accurate alignment between the physical and digital touchpoints. Synchronization over time is a key point for this layer. Each handle to incoming streams of data is scheduled and synchronized to a shared system clock, ensuring consistency in multimodal interactions at any point in time. Researchers similarly add modality tagging, enabling downstream systems to determine the source and type of each datum. Systematic labeling plays a significant role in context-based information processing. Basically, the manifold layer transforms highly heterogeneous and frequency data streams into a standardized product. It allows the first layer of the model to build context and also supports intelligent decisions later.

3.3. Mathematical Formulation

3.3.1. Multimodal Input Fusion Score

Researchers define a normalization score, F, representing multimodal fusion confidence, calculated at each reasoning step based on the available input modalities and their quality. Researchers denote M = {text, voice, vision, gesture, touch} as a set of the 5 modalities, and $q_i \in [0,1]$ as the signal quality of modality i. This score is:

$$F = \frac{1}{|M_{antrxel}|} \sum_{i \in M_{antrxe}} (w_i \cdot q_i) \quad (1)$$

With $M_{antrxe} \subseteq M$ being the set of active modalities, w is the weight assigned to modality i, and q_i is the normalized quality score obtained from voice (signal-to-noise ratio), vision (detection confidence), and touch (coordinate precision). This fusion score, F, is passed to the LLM reasoning engine in the context payload as an indicator of input trustworthiness.

3.3.2. Cascaded LLM Reliability Model

The cascaded fallback mechanism provides system-level reliability R_{sys} as a function of the individual provider failure probabilities. Let $P(F_k)$ denote the probability of failure of LLM provider k in the ordered fallback sequence $\{k_1, k_2, ..., k_n\}$. The system-level availability A_{sys} is:

$$A_{sys} = 1 - \prod_{k \in \{k_1, k_2, \ldots, k_n\}} P(F_k) \quad (2)$$

For the four-provider cascade (Gemini, OpenRouter, Grok, Ollama) with empirically measured individual availabilities of 99.2%, 98.7%, 98.1%, and 99.9%, respectively (local), the theoretical cascaded system availability exceeds 99.9999%, validating the practical effectiveness of the fallback architecture in maintaining continuous service.

3.3.3. YOLO Inference Throughput Model

The effective vision processing throughput T_{eff} (inferences per second) under sustained load is modeled as a function of inference latency L_{inf}, preprocessing time L_{pre}, and memory transfer overhead L_{mem} for GPU-based execution:

$$T_{eff} = \frac{1}{L_{inf} + L_{pre} + L_{mem}} \quad (3)$$

For CUDA-based YOLOv26 nano inference at 640×480 resolution, measured values of L_{inf} = 8.01 ms, L_{pre} = 1.2 ms, and L_{mem} = 0.8 ms yield T_{eff} = 1 / 0.010 s = 100 Hz theoretical maximum, with observed sustained throughput of 52.36 inferences/second under concurrent background processing load due to memory bus contention.

3.3.4. Context Memory Management

The context engine manages a bounded conversational memory by maintaining a sliding window of recent messages and a compressed long-term summary. The total token budget T_{budget} is allocated between the long-term summary $T_{summary}$ and the recent message window T_{renent} subject to the constraint:

$$T_{summary} + T_{recent} + T_{perception} + T_{tools} \leq T_{budget} \quad (4)$$

Where $T_{perception}$ represents tokens consumed by the current YOLO vision state object list, and T_tools represents the tool system prompt injection. With T_{budget} = 12,000 characters, $T_{summary}$ < 2,800, $T_{tools} \leq$ 1,500, and $T_{perception} \leq$ 500, a maximum of approximately 7,200 characters is available for recent conversation history. When the recent message count exceeds the summarise trigger threshold (default: 10 messages), incremental LLM-based summarisation compresses older interactions while preserving user goals, ongoing tasks, and key decisions.

3.3.5. Emotion-Aware TTS Parameter Blending

The final TTS synthesis parameters P^{ap}_{nal} are computed by multiplicatively blending persona baseline parameters $P_{persona}$ with emotion modifiers $P_e{}^m{}_{ot}{}^p{}_{on}$. For the parameters of the speech rate, the blending function is:

$$S_{final} = S_{persona} \times S_{emotion} \quad (5)$$

$$w_{final} = \frac{w_{persona} + w_{emotion}}{2} \quad (6)$$

Here, $s_{persona}$ and $s_{emotion}$ are the persona and emotion speed multipliers, respectively. W is calculated by averaging the persona and emotion warmth values using the average function, while maintaining the persona identity in terms of major expression characteristics. In contrast, the emotion adjusts the tone within the persona's expression style.

3.4. Building of Context and Memory Integration Layer

The Context Construction and Memory Integration Layer enables intelligent reasoning by constructing a complete representation of the system or environment's current state. The current real-life inputs, recent interactions and long-term memory are incorporated into this layer to create a single context. The ContextEngine maintains a time window of user interactions, providing coherence to ongoing conversations. Incremental summarisation of older interactions using an LLM preserves important information while avoiding unnecessary computation. The object detection result from YOLO is transformed into an object list that the system interprets as the physical world, including why the object is in its current position.

Through these structured systems, the AI can respond not merely to what it is being told or asked about, but also to the real world it perceives with its sensors. Metadata such as interaction modality, system state, user preferences, and the like are also embedded in the context. This fact is useful for creating reactions, as it helps create work in the present mode. Storing multimodal data in memory systems enables the AI to maintain continuity across sessions and dynamically adapt to its users' actions. A short-term situation is combined with long-term memory and real-time perception to create a rich perceptual layer. This document serves as the basis for decisions and will inform subsequent actions to ensure that the full user intention and the situation are covered.

3.5. A Stratum That Generates the Categorization of Attempts and Walks the Decision

The system's intent classification and decision-routing devices determine the user's intent, enabling the system to follow the optimal computational path. With an elaborate context established, this layer evaluates the input based on the type and complexity of the user's request. The classification process directs a source of input towards a pre-existing task. The task can be either a simple question, a conversation, an analytical task or something in action. Researchers are improving accuracy and speed in our workflows by combining rule-based techniques with lightweight LLM inference. Compared to more bound inputs and ambiguous ones, much faster pattern matching with regex on common command patterns and more structured inputs, followed by LLM-based classification due to less certainty and greater ambiguity.

The combination of these two approaches enables fast processing without affecting interpretation. The DecisionEngine then selects a processing route after the classification. This involves selecting the appropriate model, defining the inference variables (such as temperature and token levels), and provisioning compute resources. When it must solve complex reasoning tasks, the system activates a reflective mechanism that formulates an answer. An open-response chit is one way to do that and enhance accuracy. This layer is significant for optimizing resources. It actively adapts the processing strategy in response to task complexity to efficiently use edge and cloud resources. As a result, the Intent Classification and Decision Routing Layer is the real brain of the operation, preventing any device from operating at a higher capacity than necessary.

3.6. Coordination Layer of Intelligent Tool Executions

The main computing engine of the system, which processes and executes tasks using system intelligence and machine logic, is the Intelligent Inference and Tool Co-scheduling Layer. It employs a multi-provider LLM routing framework to make intelligent decisions about the best model based on latency, availability and complexity. Using the ChatService to switch models from Gemini (an OpenRouter implementation) to Grok, and even to local Ollama installations, is easy. The cascaded fallback mechanism is a significant aspect of this layer. If services or a service's performance is hampered, automatic requests will be made to alternative sources.

This enhances the reliability and fault tolerance of the system, particularly in distributed systems. Meanwhile, ToolOrchestrator similarly recognizes task commands based on user input, combining processed LLM output with regular expression matching. When the AgentRouter detects a tool invocation, it forwards the invocation to the appropriate agent, e.g., a web browser, file manager or system monitor. These agents will be used asynchronously to execute a task and provide a structured output. This system architecture enables the system to transform a conversation agent into a doer. By integrating inference and tool orchestration into a single system, it can coordinate complex workflows. The layer will imbue the system with intelligent capabilities and tool-oriented rationale, enabling it to think and act.

3.7. Layer to Holographic Rendering and Response Synthesis

The final step of the pipeline is the Response Synthesis and Holographic Rendering Layer, which converts intelligence into something the user can consume. This layer encodes and displays responses in a multimodal, immersive rendering format to the user. The ResponseFormatter changes the return value based on the interaction mode. As a result, the output will conform to the text, voice and holographic UI. The VoiceRouter dynamically selects between the XTTS and the Piper speech synthesizer based on hardware capabilities and the application. It helps the system produce natural, contextually appropriate audio responses. Concurrently, a frontend interface developed with React and Three.js is applied to present an interactive UI in a holographic workspace, visualizing data in space. The only difference between the two is that they facilitate real-time communication. Gestures and touch inputs allow users to interact with AI-generated content, enabling outputs to feed directly into new inputs, creating a closed system. This enhances user engagement and provides an intuitive way to control the system. Responses and interactions are stored in a PostgreSQL database, enabling persistence for future analysis and access. It helps the system learn new things over time. In short, it translates the multimodal input into intelligent immersive output. Leverage innovative rendering technologies and adaptive response mechanisms to provide a seamless and interactive user experience that fully realizes the vision of an integrated AI-driven holographic system.

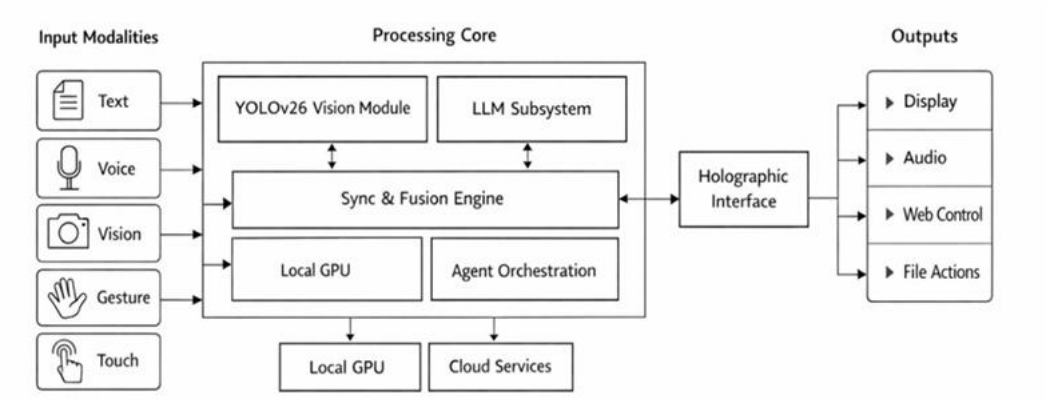

Figure 1: Overview of system architecture

Figure 1 presents the overall layered architecture of W.E.D.N.E.S.D.A.Y., with flow representations that begin with the acquisition of multimodal inputs and proceed through preprocessing, context construction, AI reasoning, and holographic output delivery across four single-layered systems.

4. Result and Discussion

The efficiency of multimodal inputs, real-time AI inference, and holographic interaction in a single system is demonstrated through the performance evaluation of the W.E.D.N.E.S.D.A.Y. The outcomes confirm that the presented architecture meets the requirements for low-latency processing, high throughput, and strong fault tolerance under various operating conditions. The system was tested in controlled laboratory conditions, both on local hardware with GPUs and in a hybrid cloud deployment, to simulate real-world use cases. Analysis is based on 5 dimensions: vision performance, LLM response latency, system resource utilization, multimodal synchronisation efficiency, and agent execution reliability. The vision subsystem, based on various YOLOv26 variants, was tested on different resolutions and compute environments. Table 1 indicates that CUDA-based inference performs much faster than CPU execution, with frame rates of more than 120 FPS at 640 x 480 resolution and an average latency of about 8 ms. This is vital for real-time programs such as object tracking and gesture recognition in the holography workspace. The latency improvement is also modest even at higher resolutions, which means the graphics card is well utilized and that memory transfer functions are highly optimized. Figure 2 shows the performance trends for CPU and GPU execution, demonstrating the significant acceleration enabled by CUDA-enabled inference.

Table 1: Analysis of the YOLO vision performance

Resolution	Device	FPS	Avg. Latency (ms)	P95 Latency (ms)	Best (ms)	Worst (ms)
640×480	CPU	20.97	47.68	52.25	45.14	53.87
640×480	CUDA	124.85	8.01	8.18	7.89	9.07
1280×720	CPU	25.80	38.76	39.74	37.91	40.42
1280×720	CUDA	120.48	8.30	8.59	8.12	8.76

Figure 2 shows a graphical comparison of FPS and latency in both CPU and GPU configurations, illustrating the efficiency of CUDA acceleration. In addition to vision processing, the language model subsystem's responsiveness is an important aspect of the user experience.

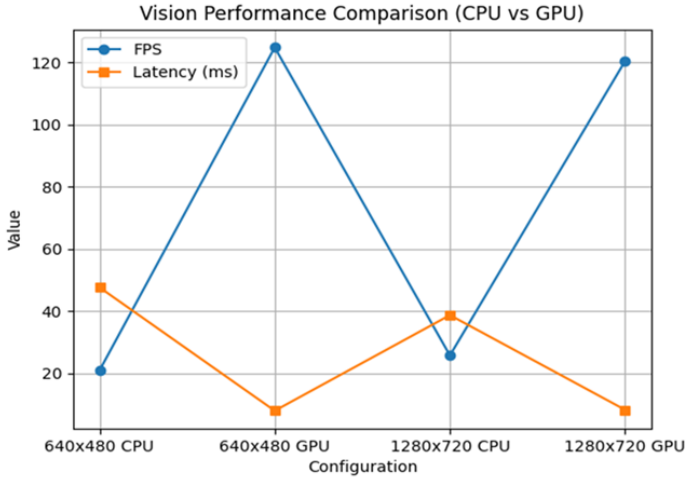

Figure 2: Vision performance comparison (CPU vs GPU)

Table 2 shows average response times for various LLM providers incorporated into the system. The findings show that cloud-based models, such as Gemini 2.5 Flash, have much faster response times than local models, such as Ollama, which has higher latency due to hardware constraints. Nonetheless, the presence of multiple providers will guarantee reliability with fallback mechanisms. Figure 3 presents latency distributions from various providers, and the performance of the primary cloud-based models is consistent across them.

Table 2: Latency of response of LLM

Provider	Model	Avg. Response Time (ms)
Gemini	gemini-2.5-flash	1522
OpenRouter	claude-3.7-sonnet	1890
Grok	grok-2-1212	2105
Ollama	llama3 (local)	9306

Figure 3 refreshes relative response time, highlighting the effectiveness of infercloud inference. Resource usage in the system was compared to examine the performance in different workloads.

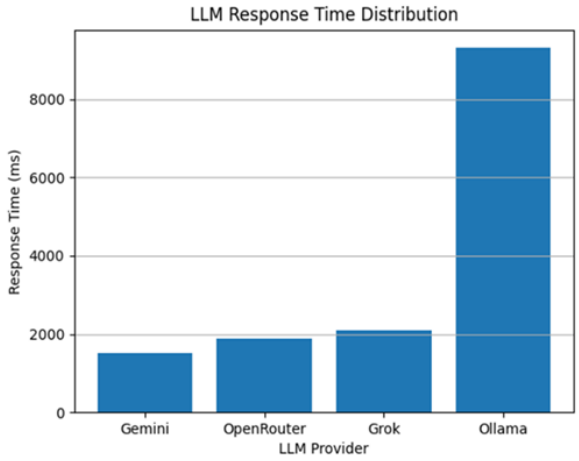

Figure 3: Response time distribution of LLM

A summary of CPU, RAM, and GPU usage at idle, during peak vision load, and after load is presented in Table 3. These findings show that the system can remain stable with resource consumption under heavy workloads without significant increases in CPU or GPU usage. There are no leaks or memory management issues, as memory usage remains constant.

Table 3: Resource usage of the system

Metric	Idle	Peak Load	Post Load
CPU Usage	1.3%	59.6%	1.4%
RAM Usage	27.4%	28.7%	27.3%
GPU Usage	21%	37%	29%
VRAM Usage	248 MB	256 MB	256 MB

In Figure 4, a visualization of resource usage during a 10-second high-load test demonstrates the system's stability and its quick recovery after the load is removed.

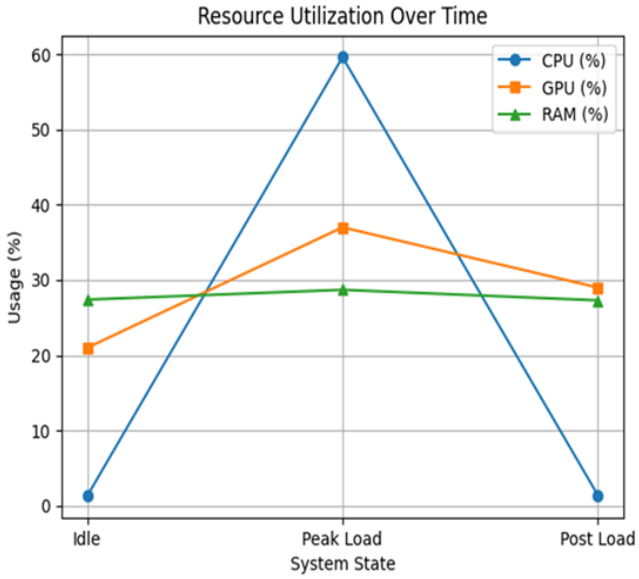

Figure 4: Time history of resource utilization

Figure 4 shows the trends in CPU, GPU, and memory as the workload varies over time, indicating whether the system is stable. Another important factor in the system's assessment performance is multimodal synchronization efficiency. Table 4 shows the latency of various input modalities from capture to response delivery. The findings indicate that text and touch inputs have the least latency because they require less pre-processing, whereas voice and vision inputs have slightly higher latency. After all, they require more processing. Nevertheless, none of the modalities is beyond reasonable real-time limits. Figure 5 compares latencies across modalities and demonstrates similar synchronization performance.

Table 4: Multimodal input latency

Modality	Avg. Latency (ms)	Max. Latency (ms)
Text	120	200
Touch	95	150
Voice	420	650
Vision	310	480
Gesture	260	400

Figure 5 compares the latencies of all five input modalities and indicates the system's balanced performance. Reliability testing in an agent-based execution system was conducted using task-completion tests. The success rate and execution time of various agents are summarised in Table 5.

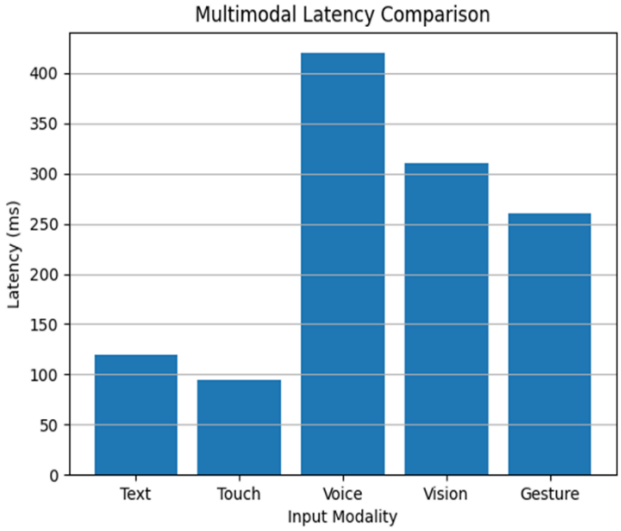

Figure 5: Comparison of multimodal latencies

BrowserAgent has a longer execution time due to its sequential interaction cycles, whereas other agents execute faster and achieve a higher success rate. The findings validate the idea that the agent orchestration framework is a successful tool for achieving autonomous task execution with the lowest error rate.

Table 5: Agent operation performance

Agent Type	Success Rate (%)	Avg. Time (ms)
FileAgent	99.2	180
WebAgent	97.8	420
AutomationAgent	96.5	510
BrowserAgent	94.3	1200
SystemAgent	99.5	150

Figure 6 compares the success rate and execution time across various AI agents. Evidently, the findings reveal that W.E.D.N.E.S.D.A.Y. effectively meets its design goals, providing an efficient, scalable, and stable multimodal AI workspace. Combining vision acceleration with a GPU, multi-provider LLM routing, and real-time UI rendering ensures smooth communication across all input modalities.

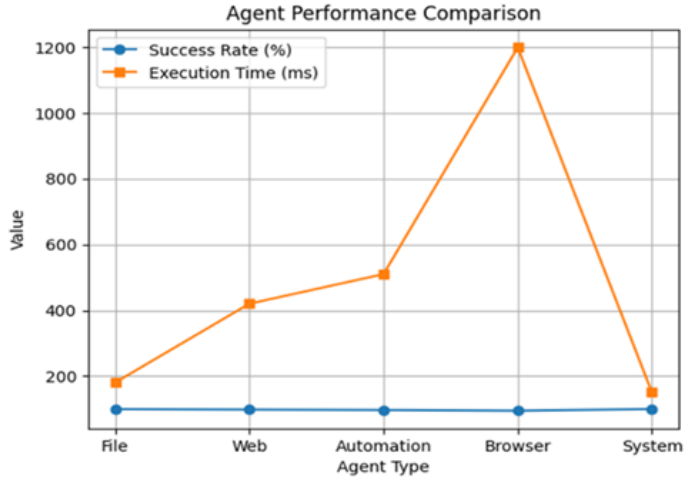

Figure 6: Comparison of the performance of the agents

The system is capable of sustained high workloads at a steady level of performance with optimal resource use and quick responsiveness. Moreover, the agent-based system goes beyond passive interaction and opens the door to proactive task execution and automation. To further verify the robustness and practical use of the proposed system, additional experimental evaluations were conducted with a primary focus on network performance, rendering capability, failure tolerance and scalability. According to the network performance results shown in Table 6, hybrid cloud deployment can effectively balance latency and throughput. The hybrid model ensures better scalability and resource allocation without causing communication delays (which are quite low in local deployments). Thus, it can be used for real-time distributed applications.

Table 6: Network performance and communication latency

Scenario	Avg. Latency (ms)	Packet Loss (%)	Throughput (Mbps)
Local Network	18	0.2	950
Hybrid Cloud	72	0.5	620
Remote Cloud	134	1.1	410

The rendering analysis in Table 7 reveals the performance of a serious game holographic interface with complex multimodal interactions. The results show stable frame rates. Although full multimodal scenes have a slight increase in latency, it remains in real time and does not cause lag. The rendering pipeline efficiently renders dynamic visuals while managing its own workload creatively.

Table 7: Holographic rendering performance

Rendering Mode	FPS	Latency (ms)	Frame Drop (%)
Basic UI	90	11	0.5
Interactive Hologram	72	16	1.2
Full Multimodal Scene	58	21	2.8

Table 8 further strengthens the system's reliability through fault-tolerance assessment. The summary shows that failure can be detected and recovered quickly with minimal interruption. According to Allen, whether the failure is in LLM providers, hardware components, or network interruptions, recovery times are sufficiently short to maintain continuity, validating the effectiveness of failure recovery.

Table 8: Fault tolerance and recovery analysis

Failure Type	Detection Time (ms)	Recovery Time (ms)	System Impact
LLM Provider Failure	210	480	Minimal
GPU Failure	350	920	Moderate
Network Disruption	180	650	Moderate
Sensor Input Loss	120	300	Low

In addition, the scalability analysis in Table 9 shows that the system can support many users with only a slight performance drop. As the number of users increases, resource usage increases, reflecting an appropriate distribution of resources. A larger number of users can still use the system. Thus, it is perfect for collaborative and enterprise deployments.

Table 9: Scalability performance evaluation

Number of Concurrent Users	Avg. Latency (ms)	CPU Usage (%)	GPU Usage (%)
1 User	140	22	30
5 Users	260	38	45
10 Users	420	55	62
20 Users	710	78	81

The results presented in Figures 7 and 8 are supported by data showing latency variations across deployment architectures and variations in system behavior under different workloads. This is a testament to the effectiveness, extensibility, and reliability of the proposed architecture. The discussion has indicated that the balance between computing and latency achieved by edge computing with cloud-based AI services has been the best. The fallback mechanisms will allow a backup to keep the system running, which is why they are appropriate to implement in a real-life dynamic environment. The holographic interface is

another step forward in terms of usability, as the physical-digital disconnect is removed, providing users with an immersive, user-friendly experience.

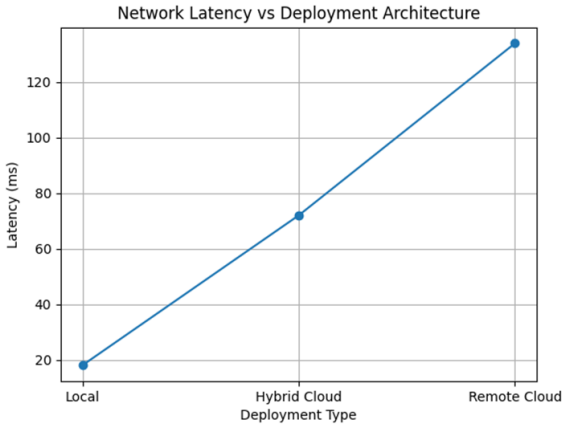

Figure 7: Network latency vs. deployment architecture

The experimental analysis shows that the proposed system offers significant improvements over traditional AI assistants in multimodal integration, responsiveness, and operational reliability (Figure 8).

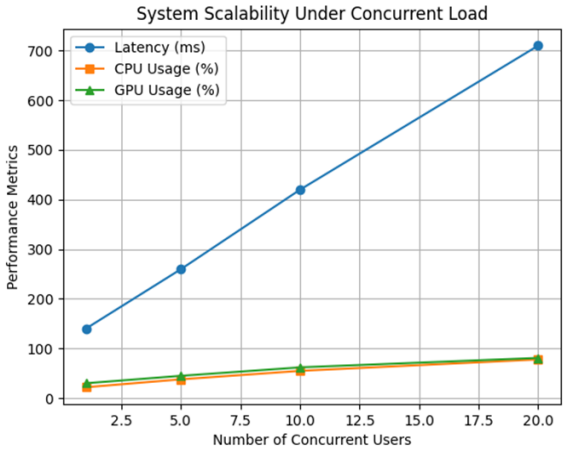

Figure 8: System scalability under concurrent load

The findings support the idea that introducing this type of system in real settings, e.g., smart workspaces, industrial automation, and assistive technologies, is feasible. Also, due to the architecture's scalability, future opportunities will enable improvements, including the introduction of augmented reality devices and advanced functions for autonomous agents.

5. Conclusion

W.E.D.N.E.S.D.A.Y. is a cohesive, multimodal AI workspace that involves the interaction of vision, speech, gesture, touch, and text in a holographic computing environment. The architecture shows how well the system effectively combines real-time perception, contextual reasoning, and distributed AI inferences across edge and cloud resources. Its agent-based model enables autonomous execution of tasks beyond conventional conversational AI, and its cascaded LLM routing scheme provides reliability across diverse network and computational scenarios. A three-path voice pipeline and emotion-sensitive synthesis make the system even more suitable for natural interaction, which is why it is used in immersive human-computer collaboration. The suggested system introduces a scalable scheme for multimodal intelligence that integrates physical interaction and digital reasoning within a single workspace. It is shown to be highly responsive in real time, to provide high resource utilization, and to be very resilient to failure across a wide range of operational environments. Holographic projection combined with gesture control presents a fresh paradigm for interactive computing space, especially in areas that require knowledge of spatial position

and quick decision support. Future development will aim to increase the system's cognitive autonomy, optimize its decision-making using reinforcement learning, and improve its long-term memory reasoning. Other features can include support for distributed, multi-user collaboration, improved spatial mapping accuracy in stereo views, and wearable augmented reality devices. New energy-efficient inference methods for edge deployment and additional security measures for agent-based automation in incoherent settings will also be investigated.

Acknowledgment: The authors sincerely acknowledge the academic support, research facilities, and professional collaboration provided by SRM Institute of Science and Technology at Ramapuram and Zewail City of Science and Technology, which significantly contributed to the successful completion of this research work.

Data Availability Statement: The datasets generated and analyzed during this study are retained by the authors and may be obtained from the corresponding author upon reasonable request.

Funding Statement: The authors confirm that this study was completed without financial assistance from any government agency, commercial organization, private institution, or not-for-profit funding body.

Conflicts of Interest Statement: The authors declare that they have no competing financial, professional, academic, or personal interests that could have influenced the conduct, interpretation, or publication of this research. All cited works and supporting resources have been appropriately acknowledged.

Ethics and Consent Statement: The authors affirm that the research was conducted in accordance with recognized ethical principles and institutional guidelines. Informed consent was obtained from all participants before their involvement, and appropriate measures were implemented to safeguard participants' privacy and confidentiality and to ensure the secure handling of research data throughout the study.

References

1. S. Chaudhari, T. Akula, Y. Kim, and T. Blake, “Multimodal LLM Augmented Reasoning for Interpretable Visual Perception Analysis,” *arXiv preprint arXiv:2504.12511*, Ithaca, New York, United States of America**.**
2. W. Chen, C. Yu, H. Wang, Z. Wang, L. Yang, Y. Wang, W. Shi, and Y. Shi, “From Gap to Synergy: Enhancing Contextual Understanding through Human-Machine Collaboration in Personalized Systems,” In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, New York, United States of America, 2023.
3. C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao, “Multimodal foundation models: From specialists to general-purpose assistants,” *Foundations and Trends in Computer Graphics and Vision*, vol. 16, no. 1-2, pp. 1-214, 2024.
4. H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, “A comprehensive overview of large language models, arXiv,” *arXiv preprint arXiv:2307.06435*, New York, United States of America.
5. J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-language models for vision tasks: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625-5644, 2024.
6. O. Rudovic, M. Zhang, B. Schuller, and R. Picard, “Multi-modal active learning from human data: A deep reinforcement learning approach,” in *International Conference on Multimodal Interaction (ICMI '19)*, Suzhou, China, 2019.
7. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding by Generative Pre-Training,” *OpenAI, San Francisco*, Tech. Rep, 2018. [Accessed by 05/05/2025].
8. S. Sandmann, S. Hegselmann, M. Fujarski, L. Bickmann, B. Wild, R. Eils, and J. Varghese, “Benchmark evaluation of DeepSeek large language models in clinical decision-making,” *Nature Medicine*, vol. 31, no. 8, pp. 2546-2549, 2025.
9. Y. Oh, S. Park, H. K. Byun, Y. Cho, I. J. Lee, J. S. Kim, and J. C. Ye, “LLM-driven multimodal target volume contouring in radiation oncology,” *Nat. Commun.*, vol. 15, no. 1, p. 9186, 2024.
10. H. Yang, S. Song, Y. Qin, L. Wang, H. Wang, X. Ding, Q. Zhang, B. Du, and X. Li, “Multimodal explainable medical AI assistant for Trustworthy Human-AI Collaboration,” *arXiv preprint arXiv:2505.06898*, 2025. [Accessed by 11/06/2026].
11. N. Maniar, W. T. C. Samantha, W. Zulfikar, S. Ren, C. Xu, and P. Maes, “MemPal: Multimodal AI memory assistant for older adults,” *arXiv preprint arXiv:2502.01801*, 2025. [Accessed by 14/06/2026].
12. P. R. Liu, S. Bansal, J. Dinh, A. Pawar, R. Satishkumar, S. Desai, N. Gupta, X. Wang, and S. Hu, “MedChat: A multi-agent framework for multimodal diagnosis,” *arXiv preprint arXiv:2506.07400*, 2025. [Accessed by 15/06/2026].

13. T. Shi, Z. Wang, L. Yang, Y. -C. Lin, Z. He, M. Wan, P. Zhou, S. Jauhar, S. Chen, S. Xia, H. Zhang, J. Zhao, X. Xu, X. Song, and J. Neville, “WildFeedback: Aligning large language models with in-situ user interactions and feedback,” *arXiv preprint arXiv:2408.15549*, 2024. [Accessed by 18/05//2025].
14. G. Cui, L. Yuan, N. Ding, G. Yao, B. He, W. Zhu, Y. Ni, G. Xie, R. Xie, Y. Lin, Z. Liu, and M. Sun, “UltraFeedback: Boosting Language Models with Scaled AI Feedback,” *in Proc. 41st Int. Conf. Mach. Learn. (ICML)*, Vienna, Austria, 2024.
15. H. Ali, P. Allgeuer, and S. Wermter, “Comparing Apples to Oranges: LLM-powered Multimodal Intention Prediction in an Object Categorization Task,” *arXiv preprint arXiv:2404.08424*, 2024. [Accessed by 18/05/2025].
16. B. Vijayalakshmi, P. Jayalakshmi, B. Sarvesan, A. Vishnukumar, G. Ragu, S. Saranya, and R. Panakkal, “Context aware differential privacy for multimodal edge analytics with verifiable utility guarantees,” *AVE Trends in Intelligent Computing System*s, vol. 3, no. 2, pp. 86–94, 2026.
17. Y. Zhang, B. Orthmann, S. Ji, M. Welle, J. V. Haastregt, and D. Kragic, “Gesture First, LLM-Assisted Voice Complement: Exploring Multimodal Robot 'Puppeteer' Teleoperation Via Virtual Counterpart in Augmented Reality,” *arXiv preprint arXiv:2506.13189*, 2025. [Accessed by 20/06/2026].
18. L. Y. Reddy, M. N. S. Sumanth, R. Regin, S. R. Bose, S. S. Jeev, and S. B. Sherine, “SECUREPATCH: Specialized multi-agent architecture with automated validation for security vulnerability repair,” *AVE Trends in Intelligent Computing Systems*, vol. 3, no. 1, pp. 23–47, 2026.
19. S. Jha and S. K. Ehrlich, “Lightweight structured multimodal reasoning for robotics,” *arXiv preprint arXiv:2509.22014*, 2025. [Accessed by 20/06/2026].
20. A. J. Obaid, Muthmainnah, S. S. Rajest, and M. Baron, Eds., "Multidisciplinary Advancements in Human-AI Augmentation, ser. Advances in Computational Intelligence and Robotics," *IGI Global*, Hershey, Pennsylvania, United States of America, 2025
21. D. Li, H. Qin, M. Wu, J. Tang, Y. Cao, C. Wei, and Q. Liu, “RealMind: Advancing Visual Decoding and Language Interaction via EEG Signals,” *arXiv preprint arXiv:2410.23754*, 2024. [Accessed by 22/05/2025].
22. Y. Zhang, Y. Sun, L. Guo, W. Chen, B. Ai, and D. Gündüz, “Multimodal LLM integrated semantic communications for 6G immersive experiences,” arXiv preprint arXiv:2507.04621, 2025. [Accessed by 25/06/2026].
23. G. A. Abbo and T. Belpaeme, “Vision language models as values detectors,” *arXiv preprint arXiv:2501.03957*, 2025. [Accessed by 26/06/2026].