Toward Multimodal, Memory-Augmented Agents: Just-in-Time Interfaces for eXtended Reality[⋆]

Luca Cordioli[1,*]

[1]Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy
Supervisor: Maristella Matera

Abstract

This doctoral research investigates multimodal, memory-augmented agents for Extended Reality (XR) to address current deficiencies in integrating generative AI, such as Large Language Models (LLMs), for improving the interaction through multimodality, personalization, and adaptivity. The research wants to investigate how to synthesize just-in-time, context-aware interfaces and integrate persistent memory systems. The primary objective is to identify models and technological architectures enabling the development of proactive, embodied XR assistants capable of mitigating cognitive load and enhancing user interaction through dynamic user interface synthesis, memory architectures, and anticipatory agent behaviors.

Keywords

Large Language Models, Multimodal Interaction, Memory-Augmented Agents, Just-in-Time Interfaces, Extended Reality, Human-Computer Interaction, Proactive Agents

1. Introduction

Over the past few years, Large Language Models (LLMs) have rapidly evolved till reaching multimodal capabilities, with models that accept both image and text inputs and achieve human-level performance on professional benchmarks [1]. In parallel, the Extended Reality (XR) ecosystem is expanding rapidly, enabling spatial, embodied, and perceptual interaction. These technological advancements are enabling new forms of interaction by enhancing the ability to understand user intents, interpret usage contexts, and adapt interaction paradigms accordingly. Research in Human-Computer Interaction (HCI) has long argued for interaction paradigms presenting only the controls required in situ. Thanks to the new AI models' capability, the concept of *ephemeral user interfaces*, composed of UI elements "intentionally created to last for a limited time only" and tailored to their physical or digital milieu, now captures this aspiration [2]. Nonetheless, most LLM systems are still limited to fixed chat window interactions.

Recent proposals suggest evolving traditional language assistants into *agentic systems* capable of perceiving environments and performing actions through dynamic interfaces, such as Rabbit's GUI-automating Large Action Model [3] and OpenAI's web-browsing Operator. The Model Context Protocol further facilitates this shift by standardizing inter-agent memory and tool interoperability [4].

The convergence of advanced AI, XR technologies, and action-oriented agents opens new opportunities for adaptive, just-in-time interfaces that cab be centered on human needs while automating tasks. My research explores this space by developing proactive, multimodal XR agents that interpret user input, generate context-aware ephemeral interfaces, and personalize interactions through structured long-term memory. This approach aims to anticipate user intent, adapt to context, and reduce cognitive load for more natural interactions.

This paper discusses the motivations and the main steps of this research, and is organized as follows: Section 2 discusses the motivations and the objectives of my research. Section 3 illustrates the main related work, and Section 4 proposes research directions to address current lacks. Section 5 finally draws the conclusions.

CHItaly 2025: Technologies and Methodologies of Human-Computer Interaction in the Third Millenium, Doctoral Consortium, 6-10 October 2025, Salerno, Italy

luca.cordioli@polimi.it (L. Cordioli)
0009-0007-1988-7990 (L. Cordioli)

2. Motivations and Objectives

This research positions XR as a hub for multimodal interaction, integrating gestures, voice, and spatial context to enable personalized, adaptive interfaces beyond purely verbal interaction, as discussed below.

Limitations of Verbal-Only Interaction. Current interactions with LLMs predominantly rely on text-based conversational interfaces. Although intuitive for basic information retrieval tasks, relying solely on textual or verbal communication introduces significant cognitive load by forcing users to serialize complex, multimodal intentions into linear text [5]. This inadequacy is emphasized in XR scenarios, where traditional input modalities (e.g., text or speech) often fail to support effective interaction due to spatial, cognitive, and contextual challenges [6].

Decades of research in HCI have consistently demonstrated that multimodal interfaces, combining speech, gesture, spatial visualization, and tactile feedback, substantially reduce cognitive load, lower error rates, and accelerate task completion compared to unimodal interactions [7].

Shallow Personalization and Lack of Adaptivity. Existing LLM deployments exhibit limited personalization, typically treating interactions as isolated episodes without meaningful longitudinal adaptation. Current models lack stable and structured long-term memory, hindering persistent user profiling and contextual awareness [8].

In XR, deep adaptivity is essential. Multimodal inputs enable rich user-agent interaction, yet current efforts—like cognitive load–aware interfaces [9] and adaptive training systems [10]—lack long-term preference learning. Embedding structured, user-controlled memory remains a key research challenge.

Restricted Interaction Flexibility in XR Environments. Current XR systems struggle with integrating and semantically enriching multimodal inputs. While modalities like gesture, speech, and gaze are common, they remain limited to rigid vocabularies and shallow interpretations. Surveys highlight that collaboration in AR/VR/MR still relies on inflexible paradigms [11]. Prototypes such as LLMR show promise by using language models to generate XR scenes, but remain weak in real-world use due to unpredictability, limited precision, and performance constraints [12]. Unlocking XR's full potential requires multimodal frameworks that fuse inputs in real-time and adapt to user context and intent.

These gaps call for rethinking LLM interactions: adaptive multimodal interfaces, dynamically shaped by agentic LLMs and grounded in persistent, user-controlled memory, are key to building truly human-centered AI in XR.

2.1. Research Objectives

My project pursues three tightly coupled goals. First, I will formalize *ephemeral interface synthesis* as a constrained generation problem in which an AI agent assembles just-in-time multimodal widgets, building on early evidence that language models can already reason about user intent to shape task-specific UIs [13]. Second, I will endow the agent with a *privacy-aware, persistent memory* substrate that combines vector retrieval with symbolic policies, extending recent long-term-memory architectures such as MemoryBank and LongMem that demonstrate continual recall and update across sessions [14, 15]. Third, I will integrate these capabilities into an embodied, proactive assistant that leverages spatial, visual, and verbal cues native to XR to *anticipate user needs and autonomously execute actions*, drawing on design principles from proactive context-aware chatbots [16].

3. Related Work

LLM-Generated Interfaces. Recent research has explored leveraging LLMs to dynamically generate user interfaces tailored to user intentions and context. The Bespoke system utilizes LLM agents to produce just-in-time interfaces by directly reasoning about user intent, enhancing task-specific interactions [13]. Similarly, the Large Language User Interface (LAUI) paradigm leverages LLMs to proactively suggest adaptive interface elements, anticipating user needs and dynamically adjusting interactions [17]. OpenAI's Operator illustrates practical applications, where an LLM autonomously navigates existing web interfaces to accomplish user-defined tasks, demonstrating both feasibility and robustness in real-world scenarios. ReactGenie complements these approaches by leveraging LLMs to interpret multimodal (voice + touch) user commands and map them to developer-defined UI components and state abstractions, enabling rich, context-aware interactions with minimal development effort [18]. Most recently, Cao et al. introduced a model-driven pipeline for generative and malleable user interfaces, where task-specific data models generated via LLMs serve as the foundation for dynamically constructed and user-customizable UIs [19].

These systems provide important contributions to prove the feasibility of generating intent-based interfaces felxibly. However, they remain limited to desktop and web-based interactions, lack continuous real-time adaptations, and do not address immersive spatial affordances or long-term user personalization, all of which my research aims to fill.

Memory Systems for LLMs. Persistent memory systems are vital for long-term personalization and adaptivity in LLM interactions. MemoryBank implements structured long-term memory, allowing LLMs to recall and utilize relevant user-specific and contextual information [14]. Similarly, Jones et al. [16] demonstrate the importance of contextual and persistent user information in a proactive, goal-oriented chatbot. Their system continuously integrates environmental and personal context (e.g., location, time, weather, user goals) to proactively recommend personalized actions, highlighting how persistent and adaptive memory can improve user engagement and effectiveness in long-term goal pursuit.

Despite their advances, these memory systems primarily handle textual data, lack multimodal integration, and offer limited user control over memory retention policies. My research extends these approaches to include structured multimodal memory with explicit user control in XR environments.

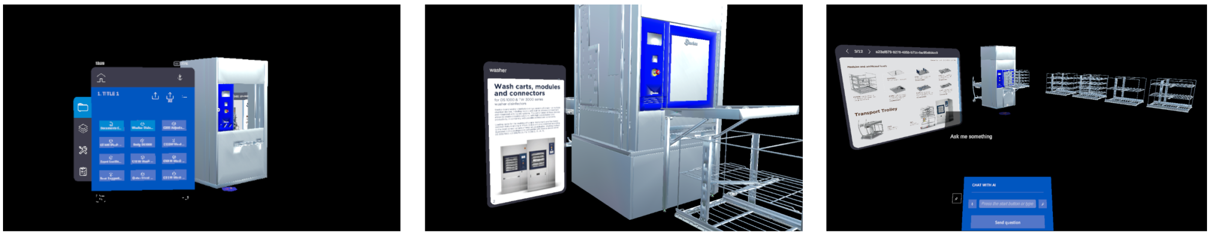
Interaction in XR Environments and Multimodal Agents. Integrating LLM-based agents into XR environments significantly enhances multimodal interaction capabilities. Systems like LLMR illustrate how LLMs can interpret natural language commands to dynamically generate and manipulate complex virtual scenes, improving intuitiveness and accessibility in 3D content creation [12]. XaiR employs multimodal capabilities, combining vision and language processing to provide real-time contextual guidance in augmented reality (AR) tasks, markedly improving task performance and efficiency [20].

However, most systems rely on predefined tasks, lack persistent multimodal memory, and seldom include proactive behaviors. My approach, instead, aims to combine ephemeral interfaces, structured memory, and anticipation in immersive XR.

4. Towards Just-in-Time Multimodal Interfaces

Our proposed approach aims to overcome the limitations of existing text-based and basic personalized interactions by developing *adaptive agentic systems that dynamically generate just-in-time multimodal interfaces.* The core concept integrates adaptive interfaces, personalized contextual memory, and proactive anticipation to deliver a deeply personalized and context-aware user experience, particularly tailored for XR environments. The envisioned AI agent will use multimodal inputs to synthesize context-aware interfaces in real time, adapting to tasks and environments. Interaction devices like smart glasses offer an ideal platform for continuous sensing and interaction.

Grounded in HCI, the research will leverage user-centered design methods, using iterative prototyping with XR technologies. The focus will be on developing and validating new multimodal interaction mechanisms, including real-time semantic fusion of voice, gesture, and visual cues. To this end, the research will adopt a research-through-design approach [21], with extensive user research to identify challenges and needs in multimodal interaction, followed by empirical studies to assess usability, interaction effectiveness, and cognitive load of the devised solutions.



1. Menu-based interaction 2. Infographic-based interaction 3. AI-driven interaction

Figure 1: Overview of the three types of interactions.

The research will adopt in particular simulated environments to test fine-grained mechanisms in anticipatory agent behaviors in controlled, context-rich scenarios, enabling systematic evaluation of relevant interaction patterns. This methodology will support the development of interaction paradigms that are responsive but still unobtrusive. It will also help identify meaningful scenarios in which the new technology can help address relevant user needs.

Currently, we are conducting a first user study to validate initial design choices developed in a prototype that integrates multimodal LLM interactions within immersive XR environments [22]. The prototype allows interactions via gestures, voice, and text, providing users with immediate, contextual responses within industrial training and maintenance scenarios. The user study, in particular, aims to evaluate the effectiveness of different XR interactions, as illustrated in Figure 1:.

1. **Menu-based interaction:** users manually navigate a menu to search for 3D models and documentation.
2. **Infographic-based interaction:** relevant 3D models are pre-loaded in the XR space, and users can access detailed infographics displaying characteristics by interacting directly with the models.
3. **AI-driven interaction:** an intelligent agent retrieves and contextualizes information from documentation, highlights relevant 3D models within the environment, and answers context-specific questions related to these models.

Early results show that the AI assistant improves task efficiency and reduces cognitive load by operating directly within the 3D space and offering contextual support. Even if preliminary, these outcomes suggest that AI integration into the XR environment — beyond traditional chatbot-style interaction — offers opportunities to enhance user experience in complex tasks, paving the way for more immersive and adaptive interaction paradigms. These results also highlight interesting research directions for advancing the definition of multimodal interaction paradigms along three main technological directions:

- **Multimodal interfaces generated by LLMs**. Using AI to dynamically synthesize multimodal interfaces, based on both explicit user requests and inferred needs, can help generate ephemeral graphical overlays. If integrated into the interactive environment via XR devices, like smart glasses, this technique can prioritize visual interaction over text to reduce cognitive load.
- **Memory and contextual customization.** This direction explores persistent, structured memory for deep personalization. By combining graph-based retrieval with symbolic reasoning, agents can track and recall contextual data—visual inputs, conversations, behaviors—enabling tailored, evolving interactions with increasing precision over time.
- **Proactivity and anticipation in XR environments.** This third direction focuses on proactive and anticipatory behaviors in XR. By analyzing multimodal inputs, visual cues, gestures, spatial data, and conversations, the agent can suggest interfaces or actions in real time. The design of proactive, anticipatory agents ca be grounded in cognitive science, which identifies expectation, goal modeling, and anticipatory representations as key drivers of adaptive and emotionally resonant interactions [23].

5. Conclusion

The research illustrated in this paper aims to introduce an innovative paradigm for adaptive multimodal interaction by integrating intelligent agents into XR environments. The project is currently in its early stages, adopting a user-centered approach to systematically identify user needs and understand interaction challenges in complex multimodal XR contexts.

Given the limited body of research on LLM-based multimodal interaction in XR, the initial focus is on characterizing the properties of the emerging interaction style described. Achieving this goal requires extensive user studies and the formulation of new conceptual models and design principles.

Future work will involve iterative empirical investigations to collect data and insights that will drive the refinement of the design framework. The research aims to develop generalized models and actionable design principles that can guide the creation of intuitive, adaptive user experiences across a wide range of contexts and user needs.

Declaration on Generative AI

The author used tools such as OpenAI's GPT and Grammarly to assist in refining the text's grammar and clarity. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] OpenAI, J. Achiam, S. Adler, et al., GPT-4 Technical Report, 2024. URL: http://arxiv.org/abs/2303.08774. doi:10.48550/arXiv.2303.08774, arXiv:2303.08774 [cs].

[2] T. Döring, A. Sylvester, A. Schmidt, Ephemeral user interfaces: valuing the aesthetics of interface components that do not last, Interactions 20 (2013) 32–37. URL: https://dl.acm.org/doi/10.1145/2486227.2486235. doi:10.1145/2486227.2486235.

[3] R. Inc., A peek into rabbit's progress with lam playground, 2024. URL: https://www.rabbit.tech/research/a-peek-into-rabbit-s-progress-with-LAM-playground, accessed: 2025-05-29.

[4] X. Hou, Y. Zhao, S. Wang, H. Wang, Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions, 2025. URL: http://arxiv.org/abs/2503.23278. doi:10.48550/arXiv.2503.23278, arXiv:2503.23278 [cs].

[5] S. Oviatt, Ten myths of multimodal interaction, Commun. ACM 42 (1999) 74–81. URL: https://doi.org/10.1145/319382.319398. doi:10.1145/319382.319398.

[6] M. Billinghurst, A. Clark, G. Lee, A survey of augmented reality, Foundations and Trends® in Human–Computer Interaction 8 (2015) 73–272. URL: http://dx.doi.org/10.1561/1100000049. doi:10.1561/1100000049.

[7] S. Oviatt, P. R. Cohen, Aims and Advantages of Multimodal Interfaces, in: The Paradigm Shift to Multimodality in Contemporary Computer Interfaces, Springer International Publishing, Cham, 2015, pp. 17–25. URL: https://link.springer.com/10.1007/978-3-031-02213-5_3. doi:10.1007/978-3-031-02213-5_3, series Title: Synthesis Lectures on Human-Centered Informatics.

[8] L. Shan, S. Luo, Z. Zhu, Y. Yuan, Y. Wu, Cognitive Memory in Large Language Models, 2025. URL: http://arxiv.org/abs/2504.02441. doi:10.48550/arXiv.2504.02441, arXiv:2504.02441 [cs].

[9] D. Lindlbauer, A. M. Feit, O. Hilliges, Context-Aware Online Adaptation of Mixed Reality Interfaces, in: Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology, ACM, New Orleans LA USA, 2019, pp. 147–160. URL: https://dl.acm.org/doi/10.1145/3332165.3347945. doi:10.1145/3332165.3347945.

[10] R. Bödding, S. A. Schriek, G. W. Maier, A systematic review and meta-analysis of mixed reality in vocational education and training: examining behavioral, cognitive, and affective training outcomes and possible moderators, Virtual Reality 29 (2025) 44. URL: https://link.springer.com/10.1007/s10055-025-01118-z. doi:10.1007/s10055-025-01118-z.

[11] A. Schäfer, G. Reis, D. Stricker, A Survey on Synchronous Augmented, Virtual, andMixed Reality Remote Collaboration Systems, ACM Computing Surveys 55 (2023) 1–27. URL: https://dl.acm.org/doi/10.1145/3533376. doi:10.1145/3533376.

[12] F. De La Torre, C. M. Fang, H. Huang, A. Banburski-Fahey, J. Amores Fernandez, J. Lanier, LLMR: Real-time Prompting of Interactive Worlds using Large Language Models, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM, Honolulu HI USA, 2024, pp. 1–22. URL: https://dl.acm.org/doi/10.1145/3613904.3642579. doi:10.1145/3613904.3642579.

[13] P. Nandy, S. O. Adalgeirsson, A. K. Sinha, T. Kraljic, M. Cleron, L. Shi, A. Singh, A. Chaudhary, A. Ganti, C. A. Melancon, S. Zhang, D. Robishaw, H. Ciurdar, J. Secor, K. A. Robertsen, K. Climer, M. Le, M. Venkatesan, P. Chi, P. Li, P. F. McDermott, R. Shim, S. Onsan, S. Vaishnav, S. Guamán, Bespoke: Using LLM agents to generate just-in-time interfaces by reasoning about user intent, in: Companion Proceedings of the 26th International Conference on Multimodal Interaction, ACM, San Jose Costa Rica, 2024, pp. 78–81. URL: https://dl.acm.org/doi/10.1145/3686215.3688372. doi:10.1145/3686215.3688372.

[14] W. Zhong, L. Guo, Q. Gao, H. Ye, Y. Wang, MemoryBank: Enhancing Large Language Models with Long-Term Memory, 2023. URL: http://arxiv.org/abs/2305.10250. doi:10.48550/arXiv.2305.10250, arXiv:2305.10250 [cs].

[15] W. Wang, L. Dong, H. Cheng, X. Liu, X. Yan, J. Gao, F. Wei, Augmenting Language Models with Long-Term Memory, 2023. URL: http://arxiv.org/abs/2306.07174. doi:10.48550/arXiv.2306.07174, arXiv:2306.07174 [cs].

[16] B. Jones, Y. Xu, Q. Li, S. Scherer, Designing a Proactive Context-Aware AI Chatbot for People's Long-Term Goals, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, ACM, Honolulu HI USA, 2024, pp. 1–7. URL: https://dl.acm.org/doi/10.1145/3613905.3650912. doi:10.1145/3613905.3650912.

[17] S. M. Wasti, K. Q. Pu, A. Neshati, Large Language User Interfaces: Voice Interactive User Interfaces powered by LLMs, 2024. URL: http://arxiv.org/abs/2402.07938. doi:10.48550/arXiv.2402.07938, arXiv:2402.07938 [cs].

[18] J. J. Yang, Y. Shi, Y. Zhang, K. Li, D. W. Rosli, A. Jain, S. Zhang, T. Li, J. A. Landay, M. S. Lam, ReactGenie: A Development Framework for Complex Multimodal Interactions Using Large Language Models, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–23. URL: http://arxiv.org/abs/2306.09649. doi:10.1145/3613904.3642517, arXiv:2306.09649 [cs].

[19] Y. Cao, P. Jiang, H. Xia, Generative and Malleable User Interfaces with Generative and Evolving Task-Driven Data Model, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, 2025, pp. 1–20. URL: https://dl.acm.org/doi/10.1145/3706598.3713285. doi:10.1145/3706598.3713285.

[20] S. Srinidhi, E. Lu, A. Rowe, XaiR: An XR Platform that Integrates Large Language Models with the Physical World, in: 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, Bellevue, WA, USA, 2024, pp. 759–767. URL: https://ieeexplore.ieee.org/document/10765376/. doi:10.1109/ISMAR62088.2024.00091.

[21] J. Zimmerman, J. Forlizzi, S. Evenson, Research through design as a method for interaction design research in hci, in: Proceedings of the SIGCHI conference on Human factors in computing systems, 2007, pp. 493–502.

[22] L. Cordioli, M. Valoriani, M. Matera, Integrating large language models into extended reality environments for enhanced user experiences, Submitted for publication (2025). Manuscript submitted for publication.

[23] M. Miceli, C. Castelfranchi, A. Ortony, Expectancy and emotion, Series in affective science, Oxford University Press, New York, NY, 2015.