# Intelligent Augmented Reality System for Multi-Tasking Guidance and Support with Seamless AI-HCI Integration

Yan-Ming Chiou[1], Bob Price[1] , Suibi Che-Chuan Weng[1,2] and Charles Ortiz[1]

[1] Palo Alto Research Center (PARC), SRI International, 3333 Coyote Hill Rd, Palo Alto, CA, USA
[2] ATLAS Institute, University of Colorado Boulder, CO, USA

## Abstract

Autonomous Multimodal Ingestion for Goal Oriented Support (AMIGOS) system, an innovative convergence of Artificial Intelligence, Augmented Reality, and Human-Computer Interaction technologies. AMIGOS integrates AI-AR technology to deliver a multi-tasking, perceptive reasoning assistant with an intuitive user interface that supports users contextually in their tasks. AMIGOS system uniquely combines deep learning and symbolic AI methods within a unified framework, enhancing the functionality and robustness of AI components. It features an advanced AI architecture encompassing a vision system, task reasoning, user modeling, and language processing capabilities, all integrated to support a task-based guidance system. This system provides extensive assistance through a multi-modal approach in an AR personal assistance system. AMIGOS employs a microphone and a head-mounted stereo camera to perceive and understand the user's environment, delivering contextually accurate and responsive assistance.

## Keywords

Augmented Reality, Explainable AI, Human-Computer Interaction

## 1. Introduction

In recent years, the realm of Artificial Intelligence (AI) has seen exponential growth, particularly in the areas of Large Language Models (LLMs) such as OpenAI GPT[1], Google Gemini[2], and Anthropic Claude[3], as well as Vision Language Models (VLMs) such as OpenAI GPT4-Vision[4] and Google Gemini Visions[2]. Advancements in LLM and VLM have significantly boosted its integration into daily human activities, with users increasingly engaging with LLMs for visual question answering, obtaining advice, and decision-making support. In parallel, the evolution of commercial Augmented Reality (AR) glasses, Ray-Ban Meta smart glasses, has enabled the integration of the users' viewpoint with multi-modal data, facilitating the query of powerful AI models on the cloud. However, despite these technological strides, current AR glasses and similar devices have not yet achieved effective multi-tasking capabilities, domain-specific step-by-step guidance, or the ability to monitor and track user actions and progress comprehensively.

Addressing these gaps, this paper introduces AMIGOS, our innovative system that acts as a perceptive reasoning assistant for physical tasks on AR devices. AMIGOS excels in offering Multi-Tasking Guidance and Support, seamlessly integrating AI with Human-Computer Interaction (HCI) across vision, language, and guidance components within the AI-AR system. We have meticulously developed various key UI modules in our AR system to facilitate multitasking in spatial computing environments, catering to users regardless of their proficiency in this domain. The system is tailored to offer timely and need-based assistance, ensuring users successfully accomplish their designated tasks. Moreover, our system incorporates the principles of eXplainable Artificial Intelligence (XAI), crucial for allowing AR users to comprehend, engage with, and trust AI systems. This integration aligns with the recent advancements in XAIR, which

[1]Joint Proceedings of the ACM IUI Workshops 2024, March 18-21, 2024, Greenville, South Carolina, USA
yan-ming.chiou@sri.com (Y.M. Chiou); bprice@sri.com (B. Price); suibi.weng@colorado.edu (C.C. Weng);charles.ortiz@sri.com (C. Ortiz)

focus on embedding context-aware intelligence and XAI in AR systems for everyday use, thereby enhancing user experiences [5][6]. With the context-aware intelligence in AR systems, it is helping the user understand what the system can perceive and what criteria the system is looking for reduce frustration and trust because the user can adapt their actions to work effectively with the AI system. If without this understanding, the system may be waiting for a condition that never occurs (perhaps the user occluded the scene with their hand during a key action). Such comprehension is fundamental to cultivating effective collaboration and trust between humans and AI systems. This approach is in harmony with the findings from Meta Reality Lab's XAIR framework[5], further underscoring the importance of transparency and intelligibility in AI-AR interactions.

## 2. Related Work

### 2.1. XAI in AR

Integrating AI with AR is crucial for enhancing user experiences, making them more intelligent and intuitive. AI's ability to analyze the environment, objects, conversations, and even emotions adds rich context, crucial for AR. This integration allows AR to offer personalized, responsive interactions, making it more effective and immersive, thus significantly broadening its applicability and appeal across various user groups. However, recent studies have shown that trust in the AI plays an important role in AI, AR, and human-interaction. Meta Reality Lab's XAIR [5] framework was developed to offer XR designers guidance on the timing, nature, and methods of delivering AI output explanations in AR environments. This framework notably includes practical use case scenarios, providing insights into content presentation for individuals with varying levels of AI literacy. Building on this foundation, Lu and et al. [7] expanded the scope to investigate the influence of various AI system and user factors on the user experience in AI-assisted spatial tasks. Their research revealed a complex interplay involving AI performance, initiation, the user's conscientiousness, and prior trust in AI. Specifically, they uncovered a significant interaction between the user's level of conscientiousness as measured by Big 5 personality instrument and distrust in the AI's performance, presumably because conscientious people want to know the system is correct and won’t accept a black box generated answer.

### 2.2. Large Language Models in AR

The recent surge in LLMs has led to their application across various fields like physics, medicine, education, and engineering. AR, while not immediately apparent, is also part of this Generative AI revolution, embracing innovations like text-to-image, text-to-3D models, and the use of LLMs, thereby expanding its capabilities and research scope. The LLMR framework [8], an extension of the GPT-4 Large Language Model, incorporates specialized modules/agents to enhance coding in C# and reasoning for building interactive 3D virtual reality experiences in Unity Game Engine. Additionally, researchers from Meta Reality Lab [9] emphasizes the need for next-generation AR/VR assistants to process multimodal contexts from diverse sources, using LLM models to integrate situational and conversational contexts. This approach highlights the importance of integration AI for advanced contextual understanding in the development of effective AR/VR system.

### 2.3. AR as Assistant and Multi-task in AR

The role of AR as a task assistant has been a subject of extensive research for years. Researchers are investigating a wide array of strategies for designing AR experiences, including the development of user-friendly interfaces, utilizing additional sensors for tutorial recording, employing computer vision for task transition detection, and evolving from single-task to multi-task support systems. In a recent study, Feiyu and Yan [10] utilized VR-simulated AR user studies

to explore three different UI transition mechanisms in AR interfaces, varying in levels of automation and user control. Their findings indicated a preference for semi-automated systems over fully automated ones, as users favored having more flexibility in making modifications. InstruMentAR [11] combines gestural information and pressure sensor data from hand-worn devices to track user manipulations on control panels. It employs pressure sensors on fingers and a decision tree classifier to accurately identify and register each specific user action within the hardware UI. Stanescu and et al. [12] using object detection models for assisting augmented reality tutorials involving complex assemblies with numerous parts (Lego and Origami). For the multi-tasks support, Oh and et al. [13] identified the problem of lack of efficient multi-tasking user interfaces. Typically, users have to repeatedly open and close applications to switch tasks. To address this, they developed a system that integrates smartphones with HoloLens 2, enabling users to multitask effectively with AR HMDs. This system utilizes spatial anchored information and introduces a concept to streamline multi-tasking in AR environments. Nevertheless, prior research has not yielded an all-in-one system capable of supporting multi-tasking on a single AR device with a robust AI guidance system. In the following section of this paper, we introduce the AMIGOS system which supports a wide range of tasks, enhancing both the speed and accuracy of task execution.

## 3. System Design

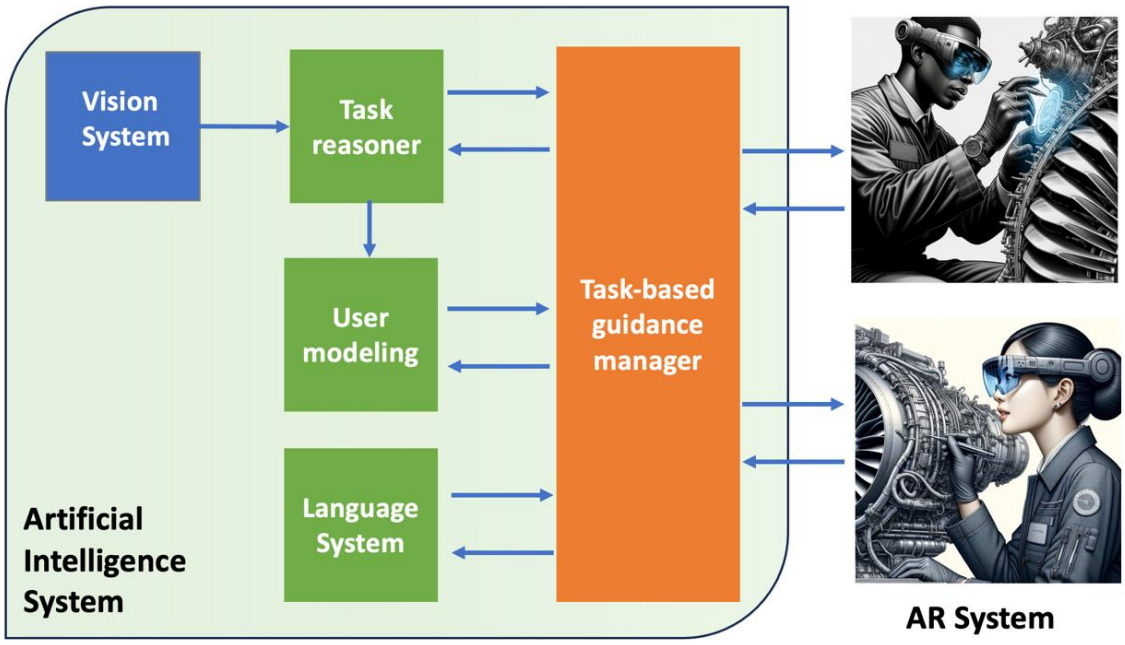


Figure 1: Architecture of AMIGOS System (Figures are illustrated by GPT4 [1])

In this work, we have devised a system of tools for creating augmented reality assistants we call **Autonomous Multimodal Ingestion for Goal Oriented Support (AMIGOS)**, an innovative dual-component framework that integrates an AI System with an AR System. The **AI** component of AMIGOS is characterized by its hybrid approach, merging deep learning with symbolic methodologies. This system is notable for its comprehensive integration of a variety of cognitive modules, marking a departure from the conventional AI development model that typically segregates the advancement of individual modality. As illustrated in Figure 1, AMIGOS includes a vision system, task reasoning capabilities, user modeling, and language processing systems. These components collectively strengthen the task-based guidance manager, offering extensive support to the AR personal assistance system through multi-modal guidance fashion. In a significant advancement over current AI assistants on the market, AMIGOS technology is designed to perceive the user's environment as they do, integrating with a microphone and a head-

mounted stereo camera (RGB and depth cameras). This setup enables the system to respond to the user's queries based on their immediate situational context, using devices such as AR headsets to provide precise, contextually relevant instructions.

### 3.1 AI-Powered Support for Augmented Reality System


Figure 2: AMIGOS AR Multi-Task Supporting System and the supporting domains (cooking, piloting/pre-flight check, medicine, and engine maintenance, Figures are illustrated by GPT4 [1])

Unlike traditional AR systems that offer limited, single-task operations, AMIGOS is designed to support multitasking for the different profile of users. This ensures users from different skill can efficiently manage multiple tasks without losing productivity. Our primary design goal is an intuitive interface that enhances engagement with AI technology. A unique feature of AMIGOS is its capability to accurately interpret the user's environment. This is achieved through an advanced perceptual model that comprehends the user's current context, including interactions with objects and various actions. The system effectively processes user commands using cutting-edge speech-to-text technology and advanced LLMs for clear communication. AMIGOS further improves the user experience by providing contextual visual aids like text, images, and 3D animations on the AR interface for guidance. Additionally, a text-to-speech model is used to deliver responses in a natural, human-like manner, greatly enhancing user experience.

In the integration of the AI vision system with human interaction, our primary emphasis has been on the deployment of AI-driven solutions for the precise detection and recognition of objects within the user's physical environment. In the cooking domain, the system recognizes key objects such as a coffee mug or a measuring cup. This is complemented by computer vision algorithms tailored to effectively map and organize the user's operational space. Our system's inherently versatile multi-tasking design enables it to accurately identify the relevant locations of objects and define the working area for assigned tasks. If the user is making tea and coffee, the system might recognize that there are two distinct cups and that each of these tasks has its own place on the counter. This capability allows the system to adaptively modify its guidance based on different tasks or recipes, contingent upon the user's workspace or the location of the objects in question. It also allows the system to update the correct recipe when an action is taken (e.g., filling the coffee cup for the coffee recipe vs. filling the tea cup for the tea recipe). These objects may vary in nature, ranging from ingredients essential for cooking recipes to tools required for engine maintenance tasks, demonstrating the system's wide-ranging applicability and adaptability. Not only for switching the tasks and guidance, but the user can use AI's capabilities to find and locate the objects on the messy workspace.

Another fundamental element of our research involves the improvement of AI language systems for human interaction, with a significant focus on the development of Voice User Interface (VUI) technologies. This development is centered on finetuning and prompt tuning LLMs to enable interactions that closely mimic human conversation. One challenge we address is

the inherent limitation of conventional off-the-shelf LLMs (GPT3 and GPT4), which, while trained on extensive web-based corpora, are prone to producing 'hallucinations' or misleading responses. This phenomenon can adversely impact the user's trust in the AMIGOS system. To mitigate this, our language system employs a Retrieval-Augmented Generation (RAG)-based approach for LLMs. We utilize specific domain materials, such as cooking recipes and engine maintenance manuals, to construct a vector database. This database enables the RAG system to accurately retrieve information that is relevant and domain-specific, thereby enhancing the reliability and contextual relevance of the system's responses.

### 3.2 Integrating AI capabilities with HCI on AR User Interface

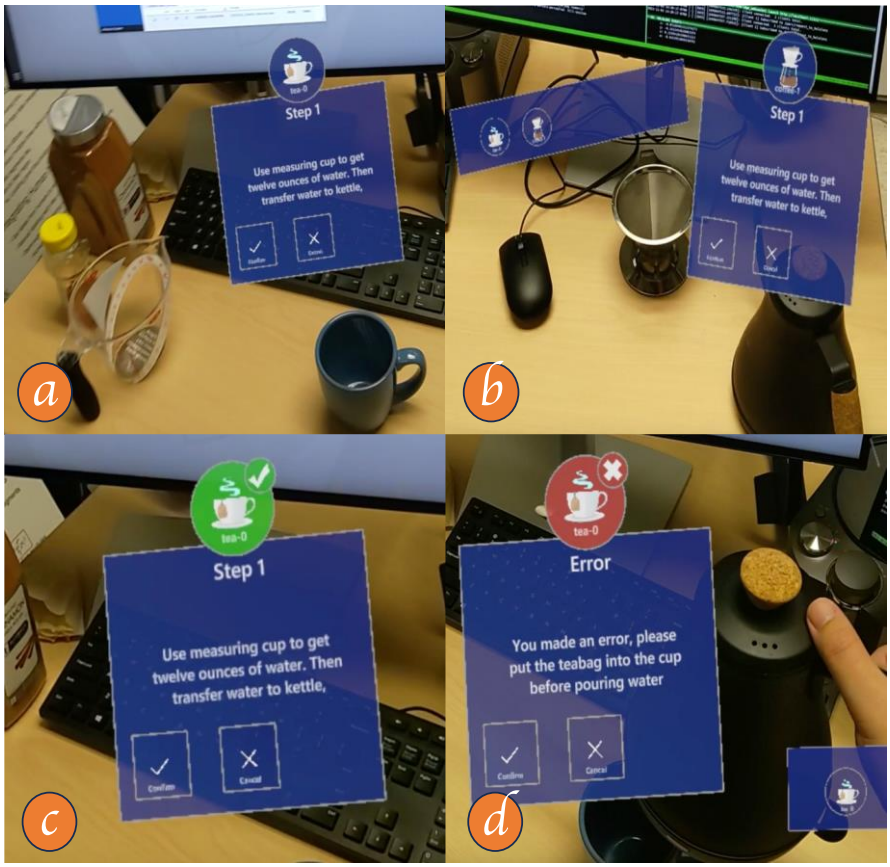


Figure 3 AMIGOS AR UI for coffee and tea tasks (a) a task panel anchored to a first object of the current task and task panel displaying relevant information based on the current task; (b) a task bar designed to monitor multiple tasks simultaneously; (c) an icon featuring a green background and a checkmark animation, used to indicate that the current step has been successfully completed and confirmed by AI; (d) an icon with a red background and a cross animation, signifying that the current step has an error and has been detected by the AI.

Our integrated AI-Human-Computer Interaction (HCI) system encompasses a comprehensive design, with the AR component being crucial yet not the only focus. Our research investigates enhancements in **spatial window positioning** to avoid UI elements obscuring the user's workspace, along with the development of **advanced multi-task tracking** features for monitoring the status of multiple simultaneous tasks. Additionally, we emphasize the provision of context-aware, intelligent widget support and the creation of dynamically adaptable templates. These templates are designed to scale efficiently across numerous tasks, significantly reducing the need for manual UI customization for each task. In the subsequent paragraphs, we will detail the design decisions that contribute to the efficacy of this robust system.

Figure 2, illustrates the user interface where three tasks are assigned: baking cookies, making coffee, and engine maintenance. Upon initiating a task, the user's visual field displays a task panel, identifiable by an icon and anchored to the first crucial object, establishing an obvious starting point. When the user selects a coffee making task, the instructions for making coffee might be anchored to the coffee cup in view. Each panel is equipped with helpful widgets, such as a timer set by the AI based on task instructions or 3D animations to clarify complex procedures. When a task necessitates a waiting period, or the user decides to switch tasks, the system smartly condenses the active panel into a small icon, thus conserving visual space and minimizing cognitive load to improve focus on upcoming tasks. This process repeats as the user navigates between tasks, ensuring undistracted concentration. The system also maintains continuous communication with a server to smoothly manage task and context transitions, providing timely, relevant support. The AMIGOS AI system, recognizing the specific task in progress (as depicted in Figure 2 with engine maintenance), proactively tracks the user's progress and supports hands-free interaction. In additional, to further enhance user efficiency in monitoring and managing multiple tasks, an intelligent task bar has been integrated into the AR experience. This feature, as depicted in Figure 2, is strategically positioned on the left side of the view and is designed to assist users in tracking various tasks concurrently. The task bar displays tasks in two distinct states: an active state, represented by full-color icons, and an inactive state, indicated by non-filled color icons. This visualization enables users to easily discern which tasks the AI system is actively supporting and tracking, and which ones are currently idle. In scenarios where a task is requires urgent attention, the system deploys multi-modal alerts. These alerts combine auditory signals emitted from the AR glasses with visual cues, such as red animations, ensuring that the user's attention is effectively drawn to the pertinent task; moreover, the alerts are designed with accessibility feature for color blindness. These features allow users to inquire and receive guidance about their present task, streamlining the experience by eliminating unnecessary visual elements and distraction.

To illustrate our design with a practical example, we present a scenario using the actual user interface, as depicted in Figure 3. In this instance, the user is tasked with simultaneously making a cup of coffee and a cup of tea. When initiating the tea task, shown in Figure 3a, the task panel anchors to the mug, indicating the starting point. Each task panel features a descriptive icon and a written and spoken summary of the current step's requirements, conveyed both visually on the panel and audibly through the AR glasses' speaker. Once the water has been poured on the teabag, the system automatically sets a timer for the tea to steep. The user switches to the coffee task (Figure 3b), the task bar remains accessible, assisting the user in monitoring all ongoing tasks so the tea will not be forgotten and go cold before serving. To enhance AI-Human collaboration and clarify the task's progress, as presented in Figure 3c and 3d, we incorporated animations and symbols (checkmark and cross) on the icons. These elements communicate the AI's decisions, and in case of errors, the system provides explanations, helping the user understand why a step might not have been correctly executed according to the AI's assessment. In this case, it explains that the user has forgotten the teabag so the user knows how to fix the situation to continue. This approach bridges the gap between AI decision-making and human understanding, facilitating a more intuitive and collaborative interaction.

In AMIGOS system, we have integrated XAI that go beyond the traditional AI-HCI interface. The inclusion of XAI is crucial, particularly as end-users increasingly interact with AI-generated result and outcomes. Through XAI, our system makes the behaviors of intelligent AR systems transparent, helping to clarify any confusion or unexpected results from AI actions. The AMIGOS assistant actively offers feedback to inform users of their goals and provides real-time responses to queries posed to AI agents during moments of uncertainty in a task, confusion over instructions, or queries about the system's progression. For instance, the system will mention that it has seen the teabag over the cup. This informs the user about the kind of evidence the system is using to conclude the user completed the step "added teabag to cup" even though the system might not be able to actually observe the teabag in the cup from the current viewpoint. If the system makes a mistake, this explicit communication about the form of its inferences makes its behavior understandable instead of random and therefore preserves the user trust in the system's

conclusions. The user can also use this information to tailor his or her future behavior to increase accuracy by ensuring visibility of key actions or objects. Leveraging VUI technology, the system grants users' effortless access to in-depth knowledge, assisting them in addressing unforeseen issues effectively and maintaining a seamless and intuitive interaction between the user and the AI. Similarly, when a cooking recipe requires the preparation of five pinwheels and the user has only made four, our system would alert the user by saying, "I see four pinwheels on the plate, but I am looking for five pinwheels on the plate." The user knows why the system judges the recipe incomplete and therefore knows how to fix the problem. Such XAI-driven design ensures that users understand the AI's requirements, recognize the nature of their mistakes, and comprehend the AI's assessments of their success or failure in each task.

## 4. Evaluation

The AMIGOS system was evaluated by an independent government contractor MIT Lincoln Labs on a set of test tasks under a variety of tasks conditions (e.g., lighting, clutter, user expertise). In the evaluation report [14] the evaluator measured the usability of the AMIGOS system and found that participants generally found it easy to use with minimal instruction and were pleased with its ability to accurately recognize steps and provide timely audio instructions. The report highlighted that the audio instructions were less distracting than reading from information panels displayed on the AR system. The integration of audio instructions and XAI principles into the AMIGOS system has markedly enhanced participants' comfort and ease of interaction, facilitating smoother collaboration with the AI. Also, Participants appreciated AMIGOS system's hands-free assistance and the audio indicators that informed them of the next steps and the system status. These features, including comprehensive step lists, were valuable when known to participants, but there was a suggestion for making them more prominent and accessible from the start. AMIGOS system's integration directly into the user's workspace, particularly offering audio assistance without user input, significantly aided participants' understanding.

## 5. Conclusion

In summary, this paper has detailed the development of AMIGOS, an innovative AI-AR system. AMIGOS is uniquely capable of comprehending the user's workspace, intentions, and objects within their everyday environment for helping user's physical task. We have designed an intelligent AR user interface that dynamically adapts to spatial contexts and responds to user needs through advanced vision and language systems, while also facilitating multi-tasking capabilities. This work illustrates the integration of multiple robust AI systems into a comprehensive AI stack and engineering the stack to connected to the AR UI for seamless AI-HCI integration. This integration is pivotal in achieving dynamic synchronization between the AI system and the user, ensuring a seamless alignment of tasks and actions, providing explanation in improving the understanding and trust between the AI and the user, thus significantly enhancing the efficiency and effectiveness of human-AI collaboration.

## Acknowledgements

This research was funded by DARPA research project "Perceptually-enabled Task Guidance (PTG)" and the contract number is HR001122C0009.

## References

[1] OpenAI, "OpenAI GPT." [Online]. Available: https://platform.openai.com/docs/models/overview

[2] Google, "Google Gemini." [Online]. Available: https://deepmind.google/technologies/gemini/#introduction
[3] Anthropic, "Anthropic Claude." [Online]. Available: https://www.anthropic.com/index/introducing-claude
[4] OpenAI, "GPT-4 Vision." [Online]. Available: https://platform.openai.com/docs/guides/vision
[5] X. Xu *et al.*, "XAIR: A Framework of Explainable AI in Augmented Reality," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Apr. 2023, pp. 1–30. doi: 10.1145/3544548.3581500.
[6] M. Abrash, "Creating the Future: Augmented Reality, the next Human-Machine Interface," in *2021 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2021, pp. 1–11. doi: 10.1109/IEDM19574.2021.9720526.
[7] F. Lu, Y. Xu, X. Xu, B. Jones, and L. Malamed, "Exploring the Impact of User and System Factors on Human-AI Interactions in Head-Worn Displays," in *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2023, pp. 109–118. doi: 10.1109/ISMAR59233.2023.00025.
[8] F. De La Torre, C. M. Fang, H. Huang, A. Banburski-Fahey, J. A. Fernandez, and J. Lanier, "LLMR: Real-time Prompting of Interactive Worlds using Large Language Models." arXiv, Dec. 18, 2023. doi: 10.48550/arXiv.2309.12276.
[9] X. L. Dong, S. Moon, Y. E. Xu, K. Malik, and Z. Yu, "Towards Next-Generation Intelligent Assistants Leveraging LLM Techniques," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, in KDD '23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 5792–5793. doi: 10.1145/3580305.3599572.
[10] F. Lu and Y. Xu, "Exploring Spatial UI Transition Mechanisms with Head-Worn Augmented Reality," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, in CHI '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–16. doi: 10.1145/3491102.3517723.
[11] Z. Liu *et al.*, "InstruMentAR: Auto-Generation of Augmented Reality Tutorials for Operating Digital Instruments Through Recording Embodied Demonstration," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–17. doi: 10.1145/3544548.3581442.
[12] A. Stanescu, P. Mohr, M. Kozinski, S. Mori, D. Schmalstieg, and D. Kalkofen, "State-Aware Configuration Detection for Augmented Reality Step-by-Step Tutorials," in *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2023, pp. 157–166. doi: 10.1109/ISMAR59233.2023.00030.
[13] S. Y. Oh, B. Yoon, and W. Woo, "AR-HMD Multitask Viewing System Concept with a Supporting Handheld Viewport for Multiple Spatially-Anchored Workspaces," in *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Singapore, Singapore: IEEE, Oct. 2022, pp. 812–813. doi: 10.1109/ISMAR-Adjunct57072.2022.00175.
[14] M. DeAngelus, J. Alekseyev, M. Timm, and V. Mancuso, "DARPA PTG TA1 System Evaluation Report (Month 12)." MASSACHUSETTS INSTITUTE OF TECHNOLOGY LINCOLN LABORATORY, Jan. 27, 2023.