Demo abstract: An XR Platform that Integrates Large Language Models with the Physical World

Sruti Srinidhi
ssrinidh@andrew.cmu.edu
Carnegie Mellon University

Edward Lu
elu2@andrew.cmu.edu
Carnegie Mellon University

Akul Singh
akuls@andrew.cmu.edu
Carnegie Mellon University

Saisha Kartik
skartik@andrew.cmu.edu
Carnegie Mellon University

Audi Lin
audil@andrew.cmu.edu
Carnegie Mellon University

Tarana Laroia
tlaroia@andrew.cmu.edu
Carnegie Mellon University

Anthony Rowe
agr@andrew.cmu.edu
Carnegie Mellon University
Bosch Research

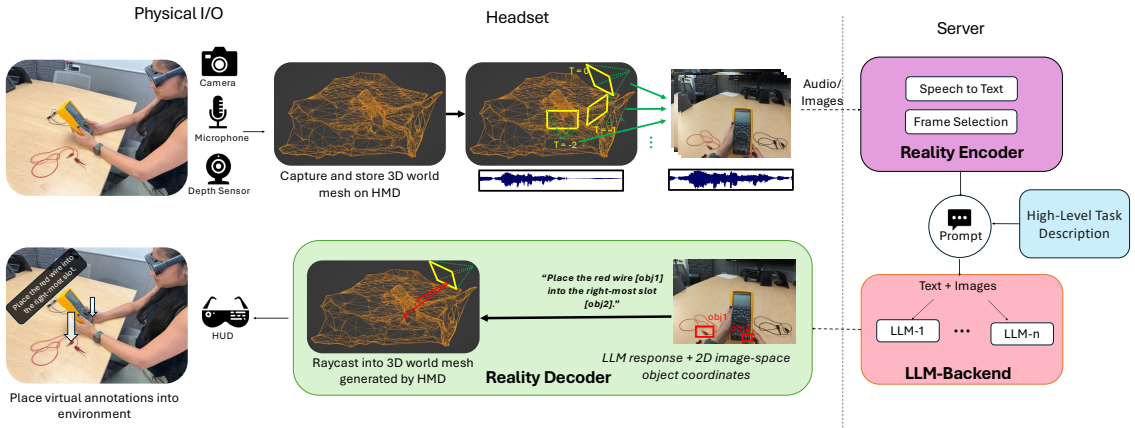

Figure 1: XaiR system architecture and dataflow – This demo showcases an Augmented Reality (AR) system that guides users through a task using virtual 3D objects anchored in the environment. First, a 3D world mesh reconstruction lives natively on Magic Leap 2 AR headset, which streams audio and images tagged with camera pose to a server. A "Reality Encoder" processes this data and, along with a high-level task description, creates a prompt sent to an "LLM-Backend." Here, the prompt is sent to multiple multimodal Large Language Models (LLMs) for inference—one generating a textual response, the other generating 2D bounding box coordinates of relevant objects. These responses are then sent to the headset. A "Reality Decoder" converts 2D coordinates to 3D by raycasting into the reconstructed mesh. Using these 3D coordinates, AR objects overlay in the user's environment for task guidance.

ABSTRACT

As Artificial Intelligence (AI) and eXtended Reality (XR) evolve, integrating them effectively remains a challenge. Although multimodal large language models (MLLMs) offer powerful reasoning over text and images, they lack an inherent understanding of 3D space. Additionally, XR headsets are resource-constrained and cannot run these models locally. To address this gap, we introduce *XaiR*, a system that integrates MLLMs with XR to enable AI-driven spatial reasoning and interaction. XaiR employs a client-server architecture in which an XR headset (client) captures spatial data, generates 2D snapshots of the 3D environment, and renders augmented reality (AR) content, while a remote server runs multiple parallel MLLMs to generate contextually aware responses. Our demo showcases an XR cognitive assistant application that guides a user through a series of instructions. Deployed on a mobile AR headset, our system dynamically interprets user actions, tracks task progress in real time, and provides textual feedback and AR-guided assistance.

SenSys '25, May 6–9, 2025, Irvine, CA, USA

ACM ISBN 979-8-4007-1479-5/25/05.
https://doi.org/10.1145/3715014.3724366

CCS CONCEPTS

• **Human-centered computing** → **Mixed / augmented reality**; • **Computing methodologies** → **Spatial and physical reasoning**; • **Computer systems organization** → **Embedded and cyber-physical systems**;

KEYWORDS

Extended Reality, Large Language Models, Artificial Intelligence

ACM Reference Format:
Sruti Srinidhi, Edward Lu, Akul Singh, Saisha Kartik, Audi Lin, Tarana Laroia, and Anthony Rowe. 2025. Demo abstract: An XR Platform that Integrates Large Language Models with the Physical World. In *The 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3715014.3724366

1 INTRODUCTION

Integrating eXtended Reality (XR) with Artificial Intelligence (AI) opens new possibilities for intelligent, immersive experiences but presents significant challenges. XR headsets capture rich spatial data yet lack the power to run large AI models. Meanwhile, Multimodal Large Language Models (MLLMs) excel at processing text and images [5, 6] but struggle to reason about 3D environments. Bridging this gap is essential to enable AI systems to not only understand but also interact with the physical world meaningfully.

We introduce XaiR (pronounced "x-air") [4], a platform that enhances XR experiences by integrating spatial context with MLLM-powered reasoning. Unlike traditional AI applications on head-mounted devices (HMDs) focused on verbal queries [1, 2], XaiR incorporates a live 3D world mesh, allowing MLLMs to interpret surroundings and generate AR responses in the physical world. This enables more immersive interactions where MLLMs provide context-aware assistance beyond simple question-answering.

To overcome XR headsets' memory limitations, XaiR employs a client-server architecture [1, 3]. The headset (client) captures spatial data, 3D world mesh, and renders AR content, while the server handles computationally intensive MLLM inference with the ability to run multiple parallel models. After processing images, speech, and text, XaiR maps responses into the physical world via raycasting within the headset's world mesh reconstruction, ensuring responses are anchored in the user's environment.

We demonstrate our system through a cognitive assistant application that guides users through physical-world tasks using AR. This application supports tasks requiring sequential instruction, such as setting up appliances or assembling furniture. By leveraging XaiR's ability to capture, reason, and respond to the physical world, the assistant provides real-time, interactive guidance anchored to the user's environment.

2 SYSTEM OVERVIEW

Our cognitive assistant application, built on XaiR, provides real-time AR guidance for hands-on tasks like assembling furniture or setting up appliances. The system consists of three components—*Reality Encoder*, *LLM-backend*, and *Reality Decoder* as seen in Figure 1—distributed between a Magic Leap 2 headset and a server with two RTX 3090 Ti GPUs. The headset captures user input (audio, images, and spatial data), while the server processes this data with multimodal AI models to generate real-time AR instructions.

2.1 The Reality Encoder

Reality Encoder on the headset converts user speech and egocentric images into prompts that the LLM-Backend can process. Each image, when captured, is marked with a timestamp and camera pose, which will be useful later when creating AR content based on the images. To track progress in sequential tasks, two images are included in each prompt—one from the current moment and one from an earlier step—along with transcribed speech and a *high-level task description*. The *high-level task description* adds information like the current instruction the user is following and the overall instruction set. This allows the system to understand the user's state and environment for more contextual feedback from the MLLMs.

2.2 LLM-Backend

LLM-backend processes prompts from the Reality Encoder, running multiple MLLMs in parallel to generate real-time assistance. In our cognitive assistant, GPT-4V [6] provides reasoning and feedback, while Ferret [5] identifies objects relevant to the instruction. Ferret generates 2D bounding boxes, helping the system track task-related objects, while GPT-4V analyzes user actions and generates textual guidance. The two models are queried simultaneously, and the server merges their outputs to produce accurate, context-aware responses. Running the models in parallel ensures that our inference time is limited by the slowest MLLM, which is typically GPT-4V.

2.3 The Reality Decoder

Reality Decoder translates LLM-Backend responses into AR overlays. By raycasting from the camera pose of the captured image, it converts Ferret's 2D object locations into 3D coordinates, allowing AR arrows to point to key objects in the user's environment. The assistant also displays step-by-step instructions and feedback as text in AR, helping users understand and correct their actions in real time. The headset persistently tracks virtual objects, keeping AR overlays in place without needing continuous reprocessing.

3 DEMONSTRATION

Our demonstration will showcase the cognitive assistant that guides users in assembling circuits using electronic snap kits. Wearing a Magic Leap 2 headset, the user will interact with real-world components while the assistant provides step-by-step AR guidance. The assistant will recognize user actions, overlay AR annotations on relevant components, and display textual instructions to ensure the user follows the correct sequence. The assistant will also continuously analyze actions and provide real-time feedback on whether each step was completed correctly and automatically advance to the next step. The user can also ask a question by speaking to the assistant, which will provide context-aware responses. This setup highlights XaiR's ability to capture and interpret physical scenes, reason about them using MLLMs, and provide intelligent, interactive AR responses.

REFERENCES

[1] [n. d.]. Meta Ray-Ban Glasses Multimodal AI. https://about.fb.com/news/2023/09/new-ray-ban-meta-smart-glasses/. Accessed: Jan 2, 2024.
[2] 2024. Project Astra Google Deepmind. https://deepmind.google/technologies/gemini/project-astra/. Online. Accessed: July 2024.
[3] Jorge Askur Vazquez Fernandez, Jae Joong Lee, Santiago Andrés Serrano Vacca, Alejandra Magana, Bedrich Benes, and Voicu Popescu. 2024. Hands-Free VR. arXiv:2402.15083 [cs.HC] https://arxiv.org/abs/2402.15083
[4] Sruti Srinidhi, Edward Lu, and Anthony Rowe. 2024. XaiR: An XR Platform that Integrates Large Language Models with the Physical World . In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 759–767. https://doi.ieeecomputersociety.org/10.1109/ISMAR62088.2024.00091
[5] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. arXiv:2310.07704 [cs.CV] https://arxiv.org/abs/2310.07704
[6] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. GPT-4V(ision) as a Generalist Evaluator for Vision-Language Tasks. arXiv:2311.01361 [cs.CV] https://arxiv.org/abs/2311.01361