Vision-Language Guided Object Localization in Mixed Reality

Han Xi
ETH Zurich
hxi@ethz.ch

Ard Kastrati
ETH Zurich
akastrati@ethz.ch

Dushan Vasilevski
Magic Leap
dvasilevski@magicleap.com

Roger Wattenhofer
ETH Zurich
wattenhofer@ethz.ch

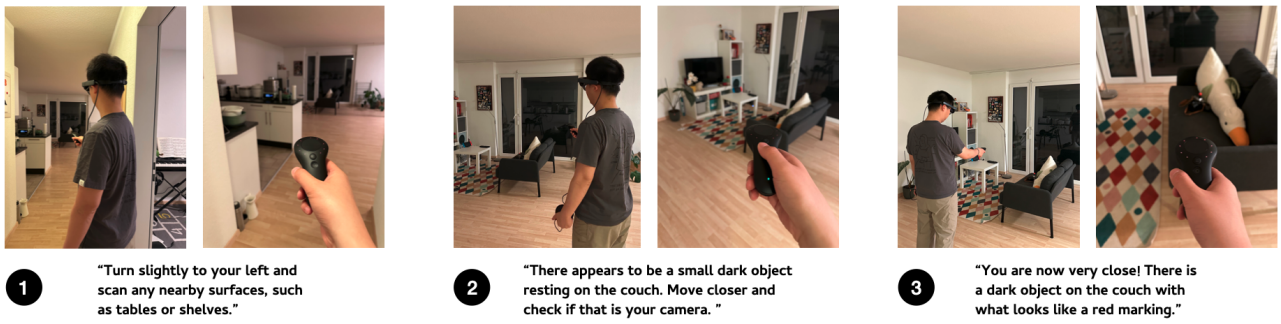

Figure 1. Step-by-step visual and textual guidance from our system to help the user locate a camera.

Abstract

We present a pipeline for real-time guided object localization on head-mounted devices (HMD), with a particular focus on object-finding tasks. Our approach leverages Vision-Language Models (VLMs), Large Language Models (LLMs), and open-vocabulary object detectors to interpret user queries, identify target objects in mixed reality environments, and provide intuitive guidance to the wearer. We integrate these components into the Magic Leap 2 (ML2) headset. In doing so, we demonstrate how combining real-time object detection with VLM-based guidance can expedite the process of locating, tracking, and retrieving items, addressing a variety of real-world scenarios such as finding misplaced tools or navigating dynamic workspaces. The extensibility of our system opens new avenues for future mixed reality applications.

1. Introduction

Recent breakthroughs in deep learning have brought LLMs and VLMs to the forefront of research, enabling new capabilities in multimodal understanding and generation. When integrated with Mixed Reality (MR) devices, they unlock innovative, context-aware experiences that seamlessly blend the digital and physical worlds. Recent work [3, 4, 7, 9, 10] has highlighted the potential of combining MR with LLMs and VLMs, creating systems capable of providing immersive guidance and contextually relevant information.

In everyday life, finding small or misplaced objects is often challenging, leading to inefficiencies and frustration. By harnessing the perceptual precision of VLMs and the reasoning capabilities of LLMs, it becomes feasible to develop systems that not only detect objects in real-time, but also provide intuitive and user-friendly visual hints to guide users to their targets.

Motivated by the practical need to simplify the object-finding process, this paper introduces a simple pipeline that integrates vision-language guidance with MR for real-time, guided object localization. Our approach interprets user queries using advanced deep learning models and overlays clear, dynamic visual cues within the user's environment, thereby easing the task of locating lost or misplaced items.

Prior work on the ML2 headset [11] has shown that exhaustive environment scanning and unified model training can yield rich spatial representations, which benefit tasks like navigation and scene understanding. However, these approaches depend on an offline preprocessing stage and coarse, large-scale scene fusion to achieve spatial awareness. TAGGAR [8], a general-purpose task-guidance pipeline driven by a GroundingDINO [5] detector, offers broad applicability but suffers from inference latencies without a high-end GPU. In contrast, our pipeline operates entirely in real time, processing incoming visual and language inputs on the fly without requiring a global 3D reconstruction. This allows for greater flexibility in dynamic

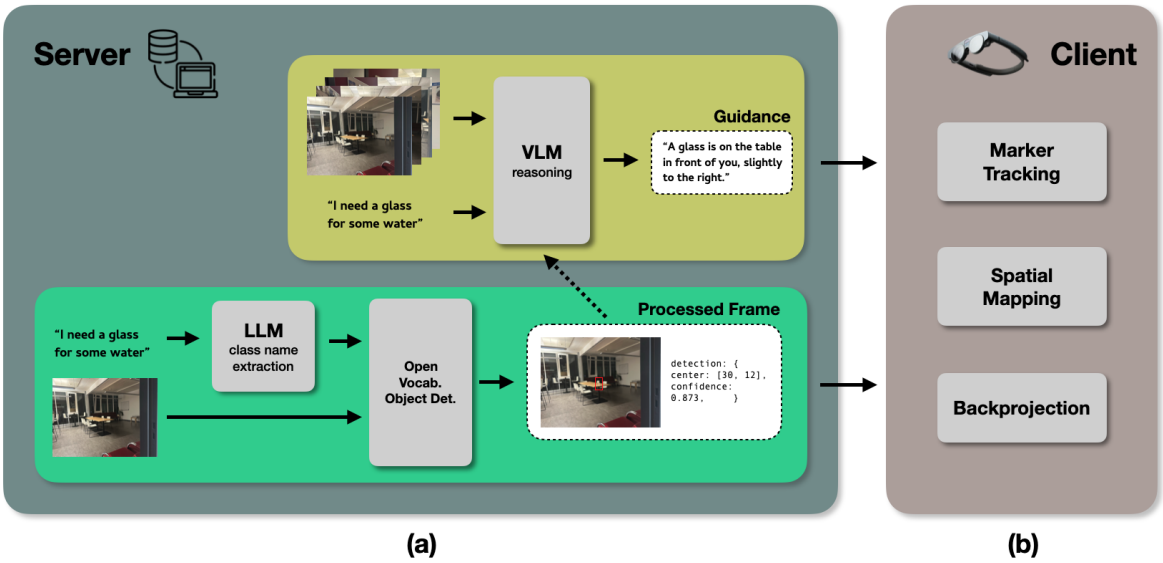

Figure 2. **Detailed end-to-end pipeline**. (a) The server–hosted agent processes incoming video frames and user prompt, generates both textual and visual guidance, then streams annotated frames and guidance data to the client. (b) The client on Magic Leap 2 applies backprojection and leverages built-in subsystems to place 3D pins and render mixed-reality cues.

or previously unseen settings and enables accurate recognition of smaller or obscured objects that may be overlooked by coarse scene-level fusion methods.

2. Method

We propose an agent-based system that interprets user commands, locates objects in real time, and provides intuitive visual guidance. The overall architecture of the agent is illustrated in Figure 2 (a). The agent is responsible for interacting with and guiding the user in locating objects. It is designed with (i) VLM for chat-based interaction and reasoning, (ii) an object detection model for visual localization, and (iii) an LLM to process minor textual information. Through this combined setup, the agent can effectively perform both reasoning and object detection while offering textual and visual guidance.

The agent's input consists of a user prompt t for the task, and a list of frames $\mathcal{F}$. Visual guidance g_{vision} is provided in the form of visual cues, which are generated frame-by-frame via a two-step process:

1. Detecting the object the user is searching for and placing a visual cue in 3D space.
2. Generating guidance (in textual form[1]) to help the user find the visual cue.

In the first step, we initially generate 2D bounding boxes r for a frame $f \in \mathcal{F}$ with respect to the user task given in natural language by using an open-vocabulary object detection model $\mathcal{D}$. The list of target objects $\mathcal{C} = \{c_1, c_2, ..., c_n\}$ is extracted from the prompt using an additional LLM (see Figure 3). This yields a list of detection results r, each of which includes bounding boxes and confidence scores. After a 2D bounding box is detected, its center is backprojected into 3D space to determine the precise location for placing the cue. More precisely, let $r_{i,center}$ be the center of the i-th bounding box, then we can perform backprojection via raycasting. Assuming a spatial mesh $\mathcal{M}$ of the environment, we cast a sphere along the ray originating from the device camera and passing through the center of the bounding box on a virtual projection plane in the user's depth direction, and then calculate the intersection point between this sphere and the mesh. We then anchor these cues to the computed 3D coordinates using a cue placement function ϕ.

$$r = \mathcal{D}(LLM(t), f) \tag{1}$$

$$g_{vision} = \phi(\text{Backprojection}(\text{r}_{\text{i,center}}, \mathcal{M})) \tag{2}$$

In the second step, for textual guidance g_{text}, we employ a VLM that takes as input the user task t and the list of image frames $\mathcal{F}$. We optionally provide object detection results $\mathcal{R}$ as an additional textual input to the VLM and obtain the textual guidance g_{text}:

$$g_{text} = VLM(\text{concat}(\text{t}, \mathcal{R}), \mathcal{F}) \tag{3}$$

The resulting textual and visual guidance form two asynchronous data streams, enabling real-time feedback within a HMD setup through a client-server framework. By integrating these streams, users benefit from immersive and

[1]We plan to extend the setup to support additional modalities, such as voice.

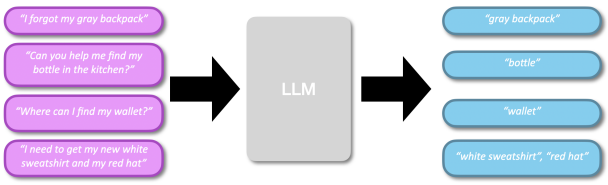

Figure 3. The LLM processes the user's prompt and outputs a list of target objects for the detection pipeline.

interactive feedback that combines contextual language instructions with dynamic on-screen cues.

3. Implementation

To enable the interaction between the user and the agent, we adopt a client-server approach and deploy the agent on the server. Hosting the agent server-side allows for dynamic updating of detection class names through client-server communication, making the system more adaptable to different user requirements. In addition, a server-based deployment ensures that the agent remains a unified entity, avoiding the complexities associated with distributed execution. The server component runs on an Apple M3 Pro with 18GB of RAM, while the client application is developed in Unity and deployed on the ML2 headset. As shown in Figure 4, during an application session, the user, wearing the HMD, prompts the agent with a specific object-finding task while actively moving through the environment. The captured visual information and processed results are constantly exchanged between the client and the server.

Figure 4. Implementation of client (Magic Leap Device) and server setup for our application.

3.1. Agent

Our agent integrates state-of-the-art deep learning models, specifically employing GPT4o-mini [6] as the VLM, Gemini as the LLM [2], and YOLOWorld [1] as the open-vocabulary object detector.

3.2. Server

The server receives continuous visual inputs from the client's camera stream, processes these frames using the integrated agent, and generates corresponding textual guidance along with annotated frames. To handle effective

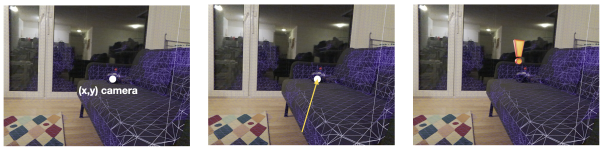

Figure 5. The visual cue (object pin) can be instantiated in space via backprojection and spherecast.

scene and object memorization both within individual sessions and across multiple sessions, we implemented an object storage database to persistently save detection results. This allows for the retrieval of previously detected objects through semantic text embeddings and similarity searches. Additionally, we employ data structures to store frames and detected objects during each session, supporting dynamic and context-aware guidance.

3.3. Client

The client periodically transmits captured frames to the server and receives processed frames along with guidance information in return. The client application, running directly on the ML2 device, leverages its integrated OpenXR subsystems within Unity to efficiently handle sensor inputs and spatial tracking. Using the ML2 meshing subsystem, we implement backprojection. The cue placement function is implemented to place a visual pin at the detected object's estimated world position. This position is determined by the intersection point between the spatial mesh and the spherecast ray originating from the user's camera. Our pin visualization approach is illustrated in Figure 5.

4. Use Case

We demonstrate the performance of our application by conducting experiments in an indoor environment, specifically in an apartment setting[2]. This allows us to assess the robustness of the system under typical household conditions, before extending the experiments to more complex or dynamic environments.

Object Detection shows how our pipeline can accurately perform object detection in 2D and back-project their bounding-box centers to recover reasonably accurate 3D position. The system is able to track moving objects in real time, ensuring robust guidance even when the targets are shifting or the wearer moves through the environment.

Localization and Reasoning demonstrates how our system combines VLM with the object detector to provide synchronized text and visual guidance in real time. Upon a

[2]A video demo can be found here

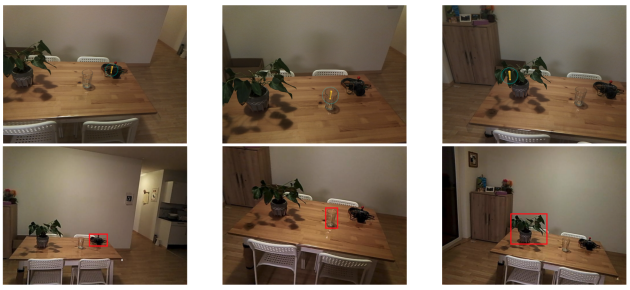

Figure 6. **Detecting objects with our system**. Top row: 3D pins marking object locations after backprojection onto the spatial mesh; bottom row: corresponding 2D bounding boxes detected in each camera frame.

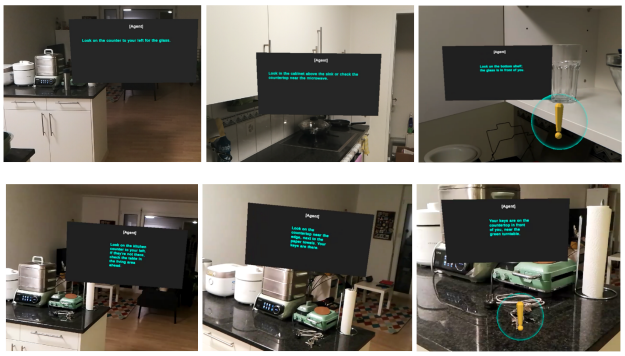

Figure 7. **Two examples of joint reasoning and localization**. In the top row, the user locates a glass by following step-by-step text prompts; in the bottom row, the system directs the user straight to their keys.

user's natural language query (for example *'Where can I find a glass?'*), the VLM interprets intent and generates stepwise instructions, while the detector scans each video frame for candidate objects. The pipeline can infer plausible object locations — even in previously unseen areas — and guide the wearer with common sense cues.

5. Conclusion

In this work, we presented a pipeline for agent-based object detection in mixed reality. Our pipeline is simple and effective, offering a working solution that can be easily extended for various applications.

The current system opens multiple directions for further development and integration into broader applications. Potential extensions include integrating voice interaction for hands-free object retrieval, enhancing MR-based guidance for navigation in complex environments, and enabling direct manipulation of virtual objects within mixed reality spaces. Moreover, the framework could be adapted for use in, for example, autonomous driving systems or as a daily assistant to improve task efficiency in various settings. Refine the search process by allowing the agent to adapt dynamically to the user's task — by continuously listening to the user's input and incorporating additional object attributes into the search criteria — is another promising direction that could further enhance the precision and usability of the system.

Overall, we believe that our contributions provide a step towards more intelligent and interactive mixed reality systems for real-time object localization, while also highlighting key challenges and opportunities for future work.

References

[1] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection, 2024. 3

[2] Google. Gemini: A family of highly capable multimodal models, 2024. 3

[3] Mikhail Konenkov, Artem Lykov, Daria Trinitatova, and Dzmitry Tsetserukou. Vr-gpt: Visual language model for intelligent virtual reality applications, 2024. 1

[4] Zhipeng Li, Christoph Gebhardt, Yves Inglin, Nicolas Steck, Paul Streli, and Christian Holz. Situationadapt: Contextual ui optimization in mixed reality with situation awareness via llm reasoning, 2024. 1

[5] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 1

[6] OpenAI. Gpt-4o system card, 2024. 3

[7] Vineet Parikh, Saif Mahmud, Devansh Agarwal, Ke Li, François Guimbretière, and Cheng Zhang. Echoguide: Active acoustic guidance for llm-based eating event analysis from egocentric videos, 2024. 1

[8] Daniel Stover and Doug Bowman. Taggar: General-purpose task guidance from natural language in augmented reality using vision-language models. In *Proceedings of the 2024 ACM Symposium on Spatial User Interaction*, New York, NY, USA, 2024. Association for Computing Machinery. 1

[9] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. Llmr: Real-time prompting of interactive worlds using large language models, 2024. 1

[10] Cindy Xu, Mengyu Chen, Pranav Deshpande, Elvir Azanli, Runqing Yang, and Joseph Ligman. Enabling data-driven and empathetic interactions: A context-aware 3d virtual agent in mixed reality for enhanced financial customer experience, 2024. 1

[11] Chengyuan Xu, Radha Kumaran, Noah Stier, Kangyou Yu, and Tobias Höllerer. Multimodal 3d fusion and in-situ learning for spatially aware ai, 2024. 1