Context-Aware Multimodal Retrieval and Adaptive UI Generation in Mixed Reality

Alireza Taheritajar* Augusta University
Jieqiong Zhao† Augusta University
Jason Orlosky‡ Augusta University

Abstract

Mixed reality (MR) systems have the potential to transform human-computer interaction by embedding digital content directly into the physical environment. However, existing approaches are often limited by static interfaces, weak contextual awareness, and fragmented integration of perception and reasoning, including state-of-the-art large language models (LLMs). In this work, we present a unified, multimodal MR framework that integrates real-time perception, visual retrieval, retrieval-augmented generation (RAG), and dynamic interface generation. Our system uses a head-mounted device to capture spatial information from the environment, while a backend performs class-agnostic object detection and semantic grounding using vision-language embeddings (SigLIP-2) and a vector database (Weaviate). Detected objects are associated with 3D locations and incorporated into user queries to support physically scene-aware reasoning. We further introduce a schema-driven approach for adaptive UI generation, where a language model produces structured, spatially grounded interfaces that dynamically respond to user intent and environmental context. We demonstrate and visually validate the efficacy of the system through several practical instruction-based use cases.

Index Terms: Context aware mixed reality, multimodal retrieval augmented generation, adaptive 3D user interfaces.

1 Introduction

Mixed reality (MR) has emerged as a promising paradigm for seamlessly integrating digital information with the physical environment, but despite rapid progress in hardware and tracking capabilities, current MR systems remain largely constrained by static interface designs and limited contextual awareness. At the same time, advances in large language models (LLMs) [15] and vision language models (VLMs) [25] have significantly improved multimodal reasoning and contextual understanding. However, their integration into MR systems remains fragmented. A key issue is that there is still a gap in creating generative, adaptive, and spatially aware user interfaces that can be built dynamically according to the physical environment and user intentions.

To address these limitations, we propose an end-to-end multimodal MR framework that unifies perception, retrieval, reasoning, and interface generation into a single pipeline. Our system operates with a head-mounted device (Meta Quest 3) as the front-end and a local workstation as the computational backend. It captures egocentric RGB input alongside spatial cues such as depth and camera pose, enabling efficient estimation of 3D object locations in real time. We use YOLOE [24] as a class-agnostic object detection module that extracts candidate regions of interest, which are subsequently refined through a lightweight post-processing pipeline to ensure robustness in unconstrained environments. Cropped object regions are embedded using a vision-language model (SigLIP-2 [22]) and queried against a vector database (Weaviate [27]) to infer object identities through similarity search. This decoupling of detection and naming allows the system to generalize beyond predefined classes and dynamically expand its knowledge base without retraining detection models.

Building upon this representation, we employ a retrieval-augmented generation (RAG) framework [8] that integrates visual context, user queries, and conversation history. Relevant documents are retrieved and combined with scene-aware object information to guide a large language model in producing grounded, context-aware responses. This design enables the system to move beyond simple visual recognition toward deeper semantic understanding and task-driven assistance.

A key contribution of this work is the introduction of a schema-driven interface generation mechanism in MR. Instead of rendering fixed UI layouts, a secondary LLM translates the generated response and scene context into a structured UI schema. This schema specifies interface components (e.g., text, images, buttons) along with their spatial placement in the 3D environment, and the resulting interfaces are dynamically instantiated and anchored to real-world objects, enabling interactive and context-aware MR experiences that evolve with user intent.

Furthermore, the system maintains session-based conversational memory, allowing users to engage in multi-round interactions that incorporate both past dialogue and newly observed visual context. This persistent memory supports more coherent and personalized interactions over time. In summary, our work advances the state of mixed reality systems by tightly integrating perception, multimodal retrieval, language reasoning, and adaptive interface generation.

2 Prior Work

We review prior work at the intersection of MR, multimodal perception, and interaction systems. We focus on key research areas that inform our approach, including object detection, vision-language modeling, retrieval-augmented generation, and spatial user interface design.

Object Detection in Resource-Constrained and MR Environments. Recent advances in object detection have enabled deployment on resource-constrained platforms such as mobile devices and MR headsets [11]. Early R-CNN-based methods [29] achieved strong accuracy but were computationally expensive, while single-stage detectors like YOLO [23] improved speed for real-time use [19]. In AR settings, additional challenges such as latency, temporal stability, and robustness under varying lighting conditions become critical [3]. Recent work also explores open-vocabulary and class-agnostic detection for dynamic environments [21], alongside vision-language models (VLMs) such as Florence-2 [28], though these often struggle in unconstrained MR scenarios [20]. As a result, hybrid approaches separating detection from semantic reasoning have gained traction [2, 12].

Vision-Language Models and Multimodal Embeddings. VLMs enable joint reasoning over images and text through shared representations. Models such as CLIP [18] introduced a unified embedding space for zero-shot classification and retrieval, while newer

*e-mail: ataheritajar@augusta.edu
†e-mail: jzhao@augusta.edu
‡e-mail: jorlosky@augusta.edu

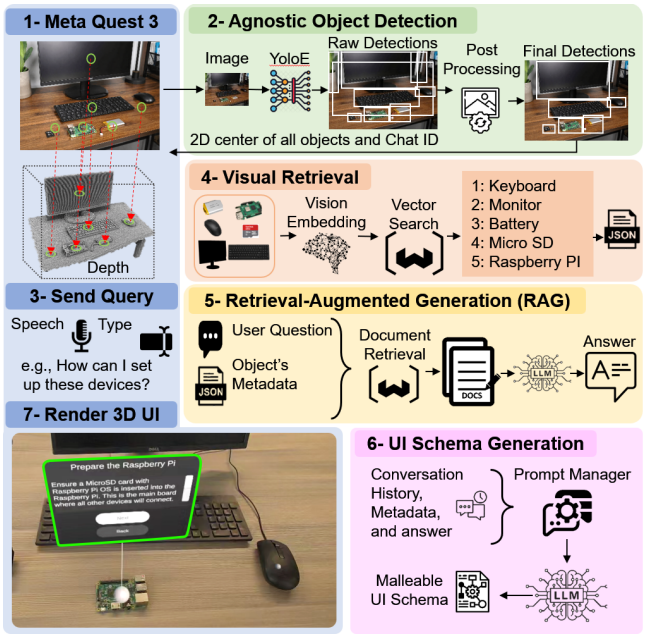

Figure 1: Our VRAG MR pipeline: egocentric images are captured on the MQ3 (1) and processed on the server using class-agnostic object detection and post-processing to obtain bounding boxes (2). 2D centers are returned to estimate 3D locations, which are combined with the user query and sent to the backend (3). Visual retrieval assigns semantic labels using vision-language embeddings and vector search (4). Labels, query and conversation history are used with RAG (5) to produce a response, and finally a generation stage converts the output into a structured UI (6) rendered in MR (7).

variants like SigLIP improve alignment and scalability for large-scale systems [14]. These methods are especially useful when explicit labels are unavailable, allowing image regions to be mapped into a semantic space for the retrieval of similar instances. This retrieval-based approach complements object detection by providing flexible semantic grounding without requiring retraining.

Retrieval-Augmented Generation (RAG). RAG [8] enhances large language models by grounding responses in external knowledge, improving factuality, interpretability, and adaptability. It has recently been extended to multimodal settings [13, 5], where both visual and textual inputs contribute to retrieval. These systems commonly rely on vector databases such as FAISS [6] and Weaviate [27] for efficient similarity search over high-dimensional embeddings. In AR applications, RAG enables context-aware interaction by linking perceptual inputs from the physical environment to external knowledge sources, supporting more informed and explainable assistance.

Spatial Computing and Mixed Reality Interfaces. Spatial computing integrates digital content into the physical world [26], requiring both accurate perception and meaningful placement of UI elements. Prior work has explored spatial anchoring [17], world-locked interfaces [9], and adaptive layouts to improve usability and reduce cognitive load [10, 7]. Traditional 2D UI paradigms often do not transfer well to MR environments, motivating context-aware and object-centric interfaces where elements are dynamically placed based on the scene [4]. These methods emphasize aligning digital content with relevant physical objects to improve understanding and interaction.

Limitations of Existing Approaches. Despite progress in object detection and the subsequent generation of in-situ instructions, key limitations remain. Most detection models are still closed-set, limiting open-world use. VLMs struggle with real-time constraints, noisy AR inputs, and domain-specific robustness. RAG systems are largely text-based with limited spatial and visual grounding. In addition, MR interfaces are mostly manually designed and not adaptive to changing scenes or user intent. Overall, these components are typically developed in isolation, leading to fragmented pipelines. Our approach addresses this gap by unifying class-agnostic detection, visual retrieval, multimodal RAG, and dynamic spatial UI generation in a lightweight, real-time system that leverages 3D context beyond current VLM-based methods.

3 Framework and Methodology

As shown in Figure 1, our system follows a distributed client-server architecture designed to support real-time visual understanding and dynamic context-aware interaction in MR. A *Meta Quest 3* headset functions as the front-end device, while a local workstation functions as the backend server. The server integrates multiple components, including a class-agnostic object detector (YOLOE), a vision-language embedding model (SigLIP 2), a vector database (Weaviate), a RESTful API (FastAPI), and a large language model (LLM) served via a local *SGLang [30]* runtime and optionally a cloud-based API such as Anthropic Claude [1]. A lightweight web interface is provided through an NGINX server for administrative document management.

This system is designed to operate in two main stages: (1) fast spatial perception and object localization, and (2) semantic understanding and response generation through a Visual Retrieval-Augmented Generation (VRAG) pipeline.

3.1 Object Detection and Post-Processing

Given an RGB frame captured by the MR device, the image is transmitted to the server via a `/detect` API endpoint. The YOLOE detector processes the image in a class-agnostic manner, producing a dense set of candidate bounding boxes. Since the objective is to capture all salient objects rather than rely on predefined categories, class labels are not directly used in downstream processing. Also, we tested Florence-2 instead of YOLOE, but it could not detect small objects in a more crowded scene.

To ensure robustness and reduce noise, a multi-stage post-processing pipeline is applied. First, detections are filtered using confidence and geometric constraints. Bounding boxes with confidence scores below a threshold (0.5) are discarded, along with boxes whose area is either too small or excessively large relative to the image. Additionally, extreme aspect ratios are suppressed to eliminate elongated artifacts often associated with background edges or slicing errors.

Next, Non-Maximum Suppression (NMS) [16] is applied in a class-wise manner using an Intersection-over-Union (IoU) threshold of 0.55 to remove redundant overlapping detections. A second pass further eliminates near-duplicate boxes across classes by enforcing a stricter IoU threshold (0.85), ensuring that each physical object is represented by a single bounding region. The final output of this stage consists solely of bounding box coordinates. To ensure the correctness of projecting the points to the actual object, only the centers of these boxes are then returned to the MR device.

3.2 3D Localization in Mixed Reality

Upon receiving the 2D bounding box centers, the Meta Quest 3 device computes their corresponding 3D positions using its internal spatial mapping capabilities. This design choice minimizes server-side computation and leverages on-device depth estimation and pose tracking for real-time performance.

Each detected object is thus associated with both 2D image-space coordinates and world-space 3D positions. These enriched object representations are then combined with the user query in natural language and forwarded to the backend via the `/vrag` endpoint.

3.3 Visual Retrieval and Object Naming

For semantic grounding, each detected object is cropped from the original image on the server. These cropped regions are encoded using the SigLIP 2 vision-language embedding model and queried against a Weaviate vector database.

The retrieval process identifies the most similar stored image for each crop, along with its associated metadata (e.g., textual description). This enables the system to assign approximate semantic labels to previously unseen objects without requiring explicit classification during detection. The result of this stage is a set of named objects, each linked to its 2D location, 3D position, and metadata.

3.4 Retrieval-Augmented Generation

The RAG module integrates visual context with textual knowledge to generate informed responses. Given the user query and the list of detected object names, the system retrieves relevant documents from the vector database. These documents, together with the current query and conversation history, are passed to the LLM.

The LLM generates a context-aware response that reflects both the physical scene and external knowledge sources. Importantly, conversation history is maintained through a session-based memory mechanism, enabling multi-round interactions. Visual retrieval artifacts are preserved separately to support debugging and visualization, while avoiding unnecessary prompt inflation.

3.5 Schema-Driven Spatial UI Generation

To bridge high-level semantic reasoning and user interaction, the system adopts a schema-driven approach for generating and rendering spatial user interfaces in mixed reality. Following the retrieval-augmented generation stage, a second LLM transforms the generated answer, scene context, and retrieved knowledge into a structured UI schema. This schema defines the interface layout, component hierarchy, and interaction flow as a sequence of steps, where each step represents a self-contained UI state.

Each step includes a unique identifier, a layout type (e.g., dialog), an optional spatial position, and an ordered list of UI components. These components specify both content and behavior, including text elements, images, and interactive buttons. Actions associated with interactive elements define navigation between steps as well as system-level operations such as closing or completing the interface. The sequencing and flow of UI elements are dynamically determined by the LLM based on the user query and retrieved instructional content.

A key feature of this design is spatial grounding. UI components can be associated with the 3D positions of detected objects, enabling the interface to be anchored directly within the physical environment. When spatial context is not required, the system supports user-centered UI placement relative to the viewer. This flexibility allows the interface to adapt to both object-centric and general interaction scenarios.

On the client side, the Meta Quest application parses the generated schema and dynamically instantiates UI elements using predefined prefabs (e.g., text blocks, images, and buttons). These elements are composed into panels and positioned according to their specified spatial attributes, enabling interactive and context-aware MR experiences.

Each interaction session is associated with a unique *chat ID*, shared between the `/detect` and `/vrag` endpoints. This mechanism enables context-preserving conversational memory, allowing users to iteratively refine queries, introduce new objects, and maintain continuity across multi-round interactions.

3.6 Design Considerations

A key design principle of the system is the separation of perception and reasoning. Fast, lightweight detection and localization are performed first to maintain real-time responsiveness, while more computationally intensive semantic reasoning is handled asynchronously on the server. Furthermore, by decoupling object detection from object naming, the system remains flexible and extensible, capable of adapting to new domains through updates to the retrieval database rather than retraining detection models.

Overall, this pipeline enables a seamless integration of spatial perception, visual retrieval, and reasoning, forming the foundation for interactive and context-aware mixed reality experiences.

4 Evaluation

We evaluate our framework by conducting a basic systems test of its end-to-end latency and response time as well as demonstrating visual results

4.1 System Performance

We evaluate the runtime performance of the proposed pipeline on a local workstation. The system is deployed on a Linux Ubuntu machine equipped with a *13th Gen Intel Core i9-13900KF CPU*, *64 GB of RAM*, and an *NVIDIA GeForce RTX 4090 GPU*. This configuration enables real-time object detection, embedding computation, and generating UI Schema with acceptable latency.

For the language reasoning and generation components, we evaluate both locally hosted and cloud-based LLMs. In the local deployment, we compared three state-of-the-art open source models: `Llama-3.2-1B-Instruct`, `Qwen2.5-3B-Instruct`, and `openai-gpt-oss-20b`. Among these, *Llama-3.2-1B-Instruct* provided the best overall performance in our setting, primarily due to more contextually relevant and consistent responses in our task setting. In addition, we benchmark a closed-source model, *Claude Opus 4.7 by Anthropic*, for higher-quality response generation in complex reasoning scenarios; however, this comes at the cost of significantly increased latency compared to local models.

This hybrid configuration enables the system to balance real-time interaction requirements with high-quality language understanding. Lightweight local models are primarily used for schema generation and fast inference, while larger closed-source models are optionally employed when higher response quality is required and latency constraints are relaxed.

Table 1: End-to-end system latency breakdown.

Component	Latency	
Object localization	65 ms	
Object detection (VRAG) (per object)	29 ms	
Document retrieval (per document)	8 ms	
Answer generation (Local LLM*)	385 ms	*
Answer generation (Cloud LLM**)	3900 ms	
UI schema generation (Local models*)	3000 ms	
UI schema generation (Cloud models**)	11600 ms	
UI rendering (per step)	11 ms	

`Llama-3.2-1B-Instruct`, ** *Claude Opus 4.7 (Anthropic)*

4.2 Qualitative Results: Spatial UI Generation

We evaluate the proposed system through qualitative analysis of end-to-end interaction scenarios in a practical MR scenario. Figure 2 illustrates a representative pipeline, demonstrating how the system transitions from perception to interactive UI generation. In this scenario, five objects are present in the scene: a *keyboard, monitor, mouse, battery pack,* and *Raspberry Pi*. In the first step, the system detects all the aforementioned objects by displaying an indicator at each object's center. The user then submits their query by typing *"How to set up these devices?"*. The user intentionally omits the names of any specific devices in order to evaluate the accuracy of the visual retrieval system in identifying objects from scene context alone.

Based on the query and scene context, the system generates interfaces in two forms: (i) user-centered (general) UIs independent of

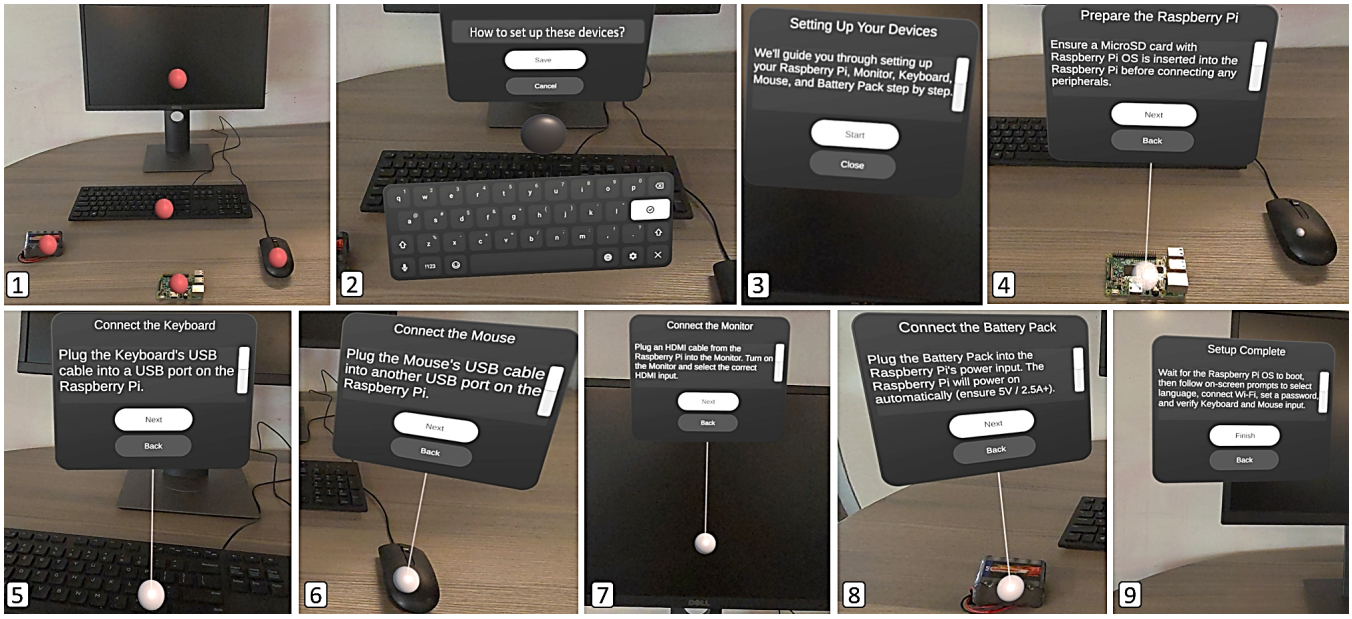

Figure 2: End-to-end example of the proposed mixed reality interaction pipeline. From left to right: (1) objects in the scene are detected and localized in 3D space, (2) the user provides a query through the interface, (3) a general user-centered UI is generated, (4-8) a spatially anchored UI is attached to each relevant object, which shows procedural instructions and interactive widgets such as buttons; and (9) the session concludes with a completion screen. This demonstrates both object-centric and user-centered UI generation within the same framework.

specific objects (steps 3 and 9 in Figure 2), and (ii) object-anchored UIs spatially aligned with relevant physical items (steps 4-8 in Figure 2). These interfaces are dynamically adapted to the task, with instructional content first presented generally and then refined into object specific guidance anchored directly to each device. This progression improves contextual relevance by signifying the relationship between digital instructions and associated physical targets.

The final stage demonstrates interactive capabilities, where users can engage with UI elements such as buttons to navigate steps or trigger actions. These interactions are integrated into the spatial environment, maintaining consistency between visual context and system response. Overall, the results highlight the system's ability to (1) ground semantic information in 3D space, (2) dynamically generate adaptive UI layouts, and (3) support interactive, multi-step task execution within a unified MR framework. A comprehensive quantitative evaluation, including latency analysis and user studies, is left for future work.

5 Discussion and Conclusion

This work presents a mixed reality pipeline that integrates real-time perception, retrieval-augmented reasoning, and dynamic UI generation to support interactive and context-aware XR experiences. The results demonstrate the feasibility of building an end-to-end system that connects object-level scene understanding with LLM-assisted reasoning and interface generation. In particular, the system shows that lightweight local perception models combined with LLM-assisted reasoning can support responsive and interactive MR applications under practical computational constraints.

Despite these promising results, several limitations remain. First, the system is constrained by local hardware resources, particularly GPU memory and computational throughput. This limitation becomes more obvious when multiple models are loaded concurrently, such as object detection, embedding generation, and LLM inference pipelines. Second, complex scene understanding remains a challenge. In cluttered or highly dynamic environments, object detection and tracking errors may propagate through the pipeline, directly affecting downstream reasoning and UI generation quality. While the current system performs adequately in controlled or moderately complex environments, robustness in unstructured scenes requires further evaluation.

In future work, we plan to incorporate more advanced VLMs and stronger object detection architectures to enhance scene understanding and semantic grounding. For example, videos can be used to demonstrate step-by-step procedures. Third, the current system is highly dependent on the quality and behavior of the underlying LLMs. Variations in model parameter size, training data, and reasoning capability significantly affect the consistency and relevance of LLM-generated outputs. To mitigate this dependency, we plan to explore multi-agent reasoning architectures, where specialized agents handle different subtasks such as perception interpretation, reasoning, UI planning, and verification. Additionally, we aim to introduce a local structured reasoning layer that reduces reliance on external model inference for intermediate decision-making.

Another limitation is the expressiveness of the generated UI. At present, UI generation is primarily schema-based and limited to predefined interaction primitives. However, we aim to extend this capability toward fully dynamic UI synthesis, where LLMs can generate complex spatial interfaces, including 3D objects such as arrows, signs, annotations, spatial highlights, and interactive game-like elements. This would enable more expressive interaction paradigms in MR environments and improve the system usability in practial applications.

Furthermore, the current system does not incorporate user profiling or context-aware personalization. We plan to integrate user and context profiling mechanisms that allow the system to adapt UI generation based on user preferences, behavior history, and task context. This would enable more personalized and adaptive MR experiences, where UI are not only scene-aware but also user-aware.

In conclusion, this work demonstrates the feasibility of an end-to-end mixed reality system that connects perception, reasoning, and dynamic interface generation using a hybrid local and cloud-based LLM architecture. While the current implementation shows promising results, it also highlights several important directions for future research. Overall, this work provides a foundation for more intelligent, adaptive, and context-aware mixed reality systems.

REFERENCES

[1] Anthropic. Claude. https://www.anthropic.com, 2024. Large language model. 2

[2] H. Bahri, D. Krčmařík, and J. Kočí. Accurate object detection system on HoloLens using YOLO algorithm. In *Proceedings of the International Conference on Control, Artificial Intelligence, Robotics & Optimization*, ICCAIRO 2019, pp. 219–224. IEEE, 2019. 1

[3] M. Benavent-Lledo, D. Mulero-Pérez, J. Garcia-Rodriguez, E. Martinez-Martin, and F. Vizcaya-Moreno. Holo4Care: A MR framework for assisting in activities of daily living by context-aware action recognition. *Multimedia Tools and Applications*, 84(22):24983–25007, 2025. 1

[4] Y. Cao, P. Jiang, and H. Xia. Generative and malleable user interfaces with generative and evolving task-driven data model. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '25, pp. 686:1–686:20, 2025. 2

[5] S. Ding and Y. Chen. RAG-VR: Leveraging retrieval-augmented generation for 3D question answering in vr environments. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, VRW '25, pp. 131–136. IEEE, 2025. 2

[6] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The faiss library. *IEEE Transactions on Big Data*, 2025. 2

[7] Y. Kim, D. Pradhan, D. Jadeja, and A. E. Kaufman. From speech-to-spatial: Grounding utterances on a live shared view with augmented reality. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*, VR '26, pp. 228–238. IEEE, 2026. 2

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, vol. 33 of *NeurIPS 2020*, pp. 9459–9474. Curran Associates, Red Hook, NY, Dec. 2020. 1, 2

[9] J. Lin, W. Sun, X. Zhang, J. Wang, P. Feng, D. Yu, and J. Zhang. SAMR: A spatial-augmented mixed reality method for enhancing vision-language models in 3D scene understanding. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, ISMAR 2025, pp. 857–866. IEEE, 2025. 2

[10] J. Lin, J. Wang, P. Feng, X. Zhang, D. Yu, and J. Zhang. AI-aided automated AR-assisted assembly instruction authoring and generation method. *Journal of Manufacturing Systems*, 83:405–423, 2025. 2

[11] M. Łysakowski, K. Żywanowski, A. Banaszczyk, M. R. Nowicki, P. Skrzypczyński, and S. K. Tadeja. Real-time onboard object detection for augmented reality: Enhancing head-mounted display with yolov8. In *Proceedings of the IEEE International Conference on Edge Computing and Communications*, EDGE '23, pp. 364–371. IEEE, 2023. 1

[12] A. Malta, M. Mendes, and T. Farinha. Augmented reality maintenance assistant using yolov5. *Applied Sciences*, 11(11):4758, 2021. 1

[13] A. Nagy, Y. Spyridis, and V. Argyriou. Cross-format retrieval-augmented generation in XR with LLMs for context-aware maintenance assistance. In *Proceedings of the IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality*, AIxVR 2025, pp. 355–361. IEEE, 2025. 2

[14] O. Nantha, B. Sathanarugsawait, and P. Praneetpolgrang. Enhanced CLEFT lip and palate classification using SigLIP 2: A comparative study with vision transformers and siamese networks. *Applied Sciences*, 15(9):4766, 2025. 2

[15] 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. *ACM Transactions on Intelligent Systems and Technology*, 16(5):106:1–106:72, 2025. 1

[16] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *Proceedings of the International Conference on Pattern Recognition*, vol. 3 of *ICPR '06*, pp. 850–855. IEEE, 2006. 2

[17] J. Orlosky, T. Höllerer, and B. Huynh. In-situ labeling for augmented reality language learning. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*, VR '19, pp. 1606–1611. IEEE, 2019. 2

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021. 1

[19] A. Taheri Tajar, A. Ramazani, and M. Mansoorizadeh. A lightweight tiny-yolov3 vehicle detection approach. *Journal of Real-Time Image Processing*, 18(6):2389–2401, 2021. 1

[20] A. Taheritajar, J. Benson, A. Gibson, B. Wilburn, J. Zhao, and J. Orlosky. Scalable object detection in mixed reality using incremental re-training and one-shot 3D annotation. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, ISMAR 2025, pp. 582–592. IEEE, 2025. 1

[21] A. Taheritajar, J. Zhao, and J. Orlosky. Augmented reality visual retrieval for object detection and corpus-guided content generation. In *Proceedings of the IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality*, AIxVR 2026, pp. 48–56. IEEE, 2026. 1

[22] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1

[23] A. Vijayakumar and S. Vairavasundaram. Yolo-based object detection models: A review and its applications. *Multimedia Tools and Applications*, 83(35):83535–83574, 2024. 1

[24] A. Wang, L. Liu, H. Chen, Z. Lin, J. Han, and G. Ding. YOLOE: Real-time seeing anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24591–24602, 2025. 1

[25] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, and H. Yang. Exploring the reasoning abilities of multimodal large language models (MLLMs): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024. 1

[26] Z.-M. Wang, M.-H. Rao, S.-H. Ye, W.-T. Song, and F. Lu. Towards spatial computing: Recent advances in multimodal natural interaction for extended reality headsets. *Frontiers of Computer Science*, 19(12):1912708:1–1912708:22, 2025. 2

[27] Weaviate B.V. Weaviate: Open source vector database. https://weaviate.io, 2024. 1, 2

[28] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024. 1

[29] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han. Oriented R-CNN for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3520–3529, 2021. 1

[30] L. Zheng, L. Yin, Z. Xie, C. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez, et al. SGLang: Efficient execution of structured language model programs. *Advances in Neural Information Processing Systems*, 37:62557–62583, 2024. 2