Generative Situated XR Assistance: Generating Context-Aware Guidance for the Physical World

Sruti Srinidhi
Electrical and Computer Engineering
Carnegie Mellon University, USA
ssrinidh@andrew.cmu.edu

David Lindlbauer
Human Computer Interaction Institute
Carnegie Mellon University, USA
davidlindlbauer@cmu.edu

Anthony Rowe
Electrical and Computer Engineering
Carnegie Mellon University and
Bosch Research, USA
agr@andrew.cmu.edu

Abstract

Current physical assistance relies on static, generic content that requires users to manually translate information onto their physical environments. We propose *Generative Situated XR Assistance*, a new paradigm for situated assistance that replaces pre-authored content with a continuous perception-generation loop. By shifting the focus from designing static assets to defining generative logic, these systems perceive a user's specific physical context and synthesize context-aware spatial content at runtime. We outline the computational principles of this closed-loop lifecycle and discuss the primary technical requirements and challenges in achieving this vision, specifically around detailed state inference, spatial verification, and long-term adaptation.

1 The Cognitive Gap in Physical Tasks

From industrial maintenance to collaborative fabrication, humans perform complex and embodied tasks daily. These tasks are inherently spatial, yet the assistance supporting them remain largely static, 2D, and generic. Current assistance frameworks rarely account for the specific state of a workspace, the exact variant of a hardware component, or the unique constraints of the user's immediate environment.

This creates a significant cognitive burden: the user must perform the "heavy lifting" of mentally mapping static digital data onto dynamic physical spaces, translating generic instructions to their specific environment, and filtering out irrelevant information[1, 5, 10, 13, 17]. This gap exists because static content is fundamentally decoupled from the live physical context, making the burden of contextualization, i.e. adapting the guidance to the moment, remains entirely on the user. This often leading to frustration and loss of task flow.

Historically, this challenge mirrors the evolution of digital information retrieval. Previously, users queried search engines and had to manually parse and synthesize information from a list of disparate sources to find an answer. Modern Large Language Models (LLMs) have transformed this experience by directly synthesizing a singular, context-specific answer from that vast body of data. We believe that situated physical assistance is at a similar inflection point. Just as LLMs moved us from searching to answering, generative XR assistance can move us from interpreting generic content to interacting with adaptive, situated guidance.

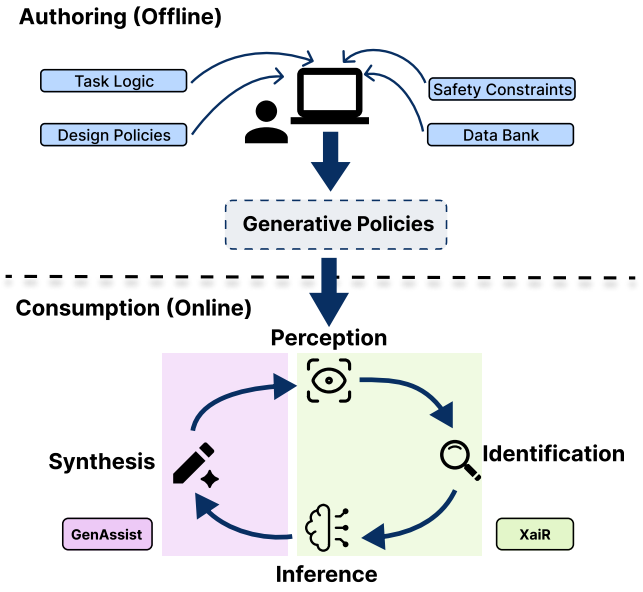

Figure 1: The Generative Situated XR Assistance Lifecycle. Unlike traditional documentation, this architecture operates as a continuous, closed-loop system. (Top) Offline Authoring: Experts define high-level generative policies, such as safety constraints and task logic, rather than static assets. (Bottom) Online Consumption: The system continuously cycles through perception, inference, and synthesis to provide context-aware guidance. This loop allows the interface to synchronize digital information with the dynamic physical environment in real time.

While prior research in situated assistance has demonstrated that projecting 3D annotations directly onto tasks significantly reduces errors and improves completion speed [7, 9], these systems remain notoriously brittle. Because they rely on pre-authored, fixed content, every permutation of a task and environment requires intensive manual design. This reliance on expert-led, hand-crafted assets creates an authoring bottleneck that makes traditional XR unscalable for the infinite variety of everyday physical tasks [2, 11].

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

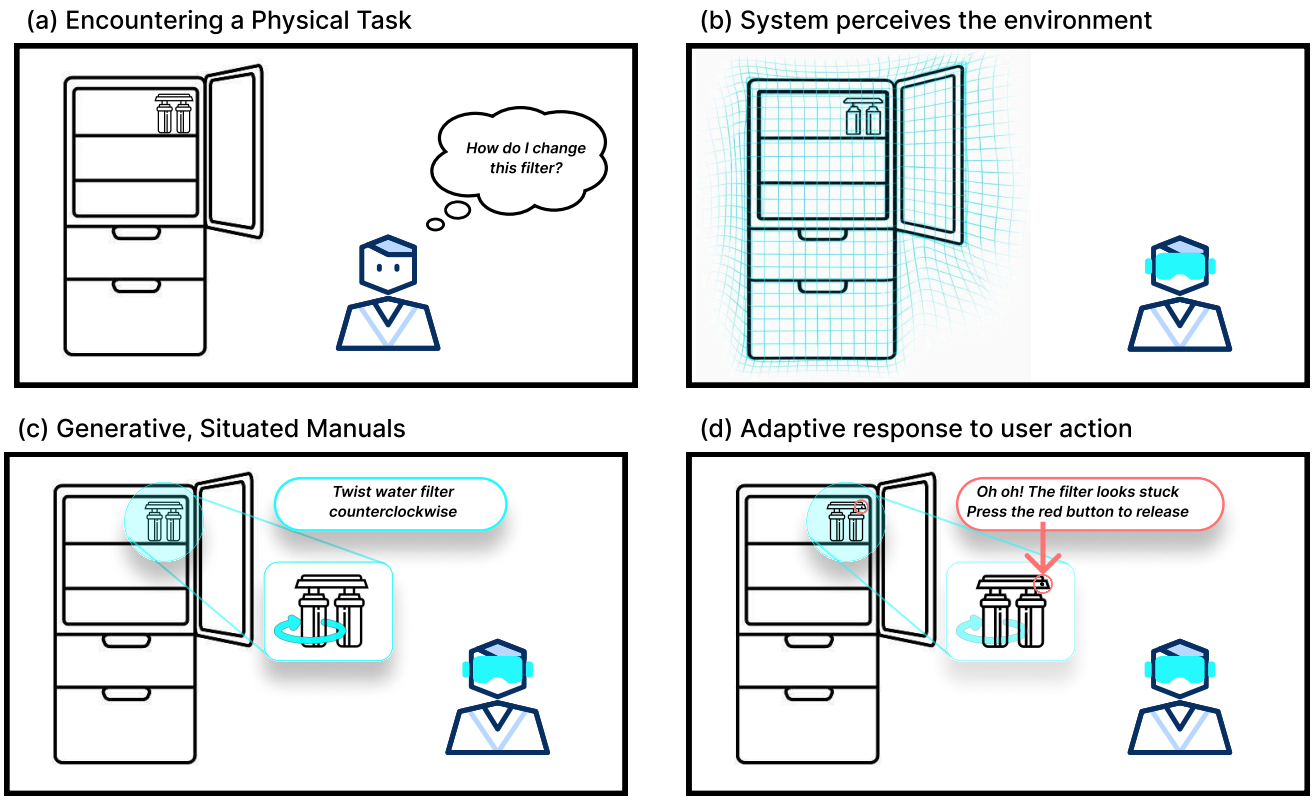

Figure 2: A four-stage interaction illustrating runtime-generated guidance. (a) A user encounters a physical task without context-specific support. (b) A generative XR system perceives and models the environment. (c) Spatially grounded instructions are generated in situ, attaching guidance to relevant components. (d) As the situation evolves, the system adapts its guidance to the user's actions.

Position Statement: We argue that recent advances in generative AI and XR enable a fundamentally different class of interface. We propose **Generative Situated XR Assistance** as seen in Figure 1. These systems perceive the environment, infer user task progress, and generate spatial content on the fly for *this* user, *this* space, and *this* moment. Unlike pre-authored content, these are runtime-generated XR interfaces that attach guidance directly to the physical world and adapt as situations change. By positioning these systems as a form of Generative User Interface (GenUI), we surface the technical requirements and responsibility challenges this shift introduces. This framing moves the field toward a paradigm where designers no longer author static content but instead define the generative logic for physical interactions.

2 Concept: From Authoring to Generation

We propose a shift from pre-authored, static content to systems that treat spatial information as a dynamic, runtime-generated interface. This approach situates generative UI (GenUI) in physical contexts, shifting the design paradigm from authoring static content to defining generative logic. This paradigm is applicable to a wide range of embodied tasks where digital information must be mapped to the world, from complex industrial assembly and laboratory protocols to collaborative remote assistance.

To illustrate this, we focus on the classic case of instruction manuals for physical tasks. This scenario remains a standard benchmark for situated assistance, because despite the digitization of most domains, much of the knowledge and instructions for physical tasks (like assembly and troubleshooting) remain represented by static documentation. These traditional systems represent a persistent gap between the dynamic nature of physical interaction and the rigid nature of printed or 2D digital data.

Consider the task of replacing a refrigerator water filter. In traditional instruction manuals, the user takes on the overhead of searching for the information needed for their task, such as finding the specific section on filter replacement. This process is rarely self-contained. Users often find themselves piecing together information from multiple sources. They might flip through a physical manual while they also search for videos or blogs to see how the task must be performed, especially if the current state of the appliance does not match what the manual depicts. The user essentially has to build the manual themselves by mentally mapping abstracted 2D diagrams or generic video angles onto their specific physical environment.

In contrast, a generative system begins by sensing and interpreting the appliance's physical state and uses this information to generate context appropriate guidance. As the user moves, the system generates viewpoint-dependent UI, such as a 3D arrow pointing at the water filter and rendering a "twist" animation aligned with the filter's orientation. The system can also respond to interaction errors or hesitation by generating contextual warnings or introducing supplemental sub-tasks, such as "Check for a release button," adapting guidance in real time to the observed physical state. A visualization of this concept is shown in Figure 2.

3 Lifecycle of Generative Physical XR Interfaces

To move beyond the limitations of pre-authored assistance, we must redefine the relationship between digital information and physical tasks. Traditionally, this process is linear and manual. An expert authors static instructions, and a user is left to interpret them. We propose a shift toward a continuous, closed-loop lifecycle that bridges the gap between expert logic and user needs. An overview of this lifecycle is shown in Figure 1.

3.1 Authoring: From Static Content to Generative Policies

In this new lifecycle, the role of the expert shifts from manually creating content to defining generative policies. Instead of authoring fixed assets for specific physical configurations, the expert provides high-level logic and safety constraints. This allows the system to scale across physical variants without manual redesign because the expert intent is translated into interfaces at runtime.

This shift requires the ability to synthesize complex XR content, such as animations or spatial markers, on the fly from high-level descriptions. These descriptions are not fixed. They stem from a real-time understanding of the environment and the current state of the task. The system must not only determine when to present information but must also be able to generate the spatial content required to communicate that information in the specific physical context. This makes the interface synthesis-based rather than retrieval-based.

3.2 Consumption: From General to Personal Guidance

For the end-user, the consumption of instructions becomes a personalized, adaptive experience. As the system operates through a continuous perception loop, it can identify the specific tools available or the unique layout of their workspace. Unlike a static documentation, the system synthesizes instructions that are executable in the user's current context. For instance, if a user encounters a jammed component during consumption, the system can infer this state and immediately generate a sub-task or a contextual warning (e.g., "The filter is stuck: press the red release button"). This move from general to personal guidance ensures the interface responds to the messy reality of physical work. By adapting the interface in real time, the system transitions from a passive reference to an active participant that manages the cognitive load of the user.

3.3 The Continuous Perception-Generation Loop

At the core of this lifecycle is a real-time system that replaces manual content mapping with a four-stage loop. This loop ensures that the information is not a static document but a living interface that responds to the physical environment.

(1) **Perception:** The system continuously perceives the physical environment through multi-modal sensor streams like audio, egocentric video and depth data. This stage goes beyond simple image capture by maintaining a constant awareness of the user's workspace and tool layout. It serves as the foundation for the loop by providing the raw data needed to understand the current context of the task and the user.
(2) **Identification:** In this step, the system identifies relevant components and their semantic states. This goes beyond simple object detection to understand the functional condition of the environment, such as whether a mechanism is engaged or a component is out of alignment. This stage maps raw visual data onto a known model or task graph to determine exactly what parts are present and what state they are in. By focusing on semantic states, the system can recognize hardware variations and specific sub-models that might not match a standard reference set. This allows the generative engine to adjust its guidance based on the unique physical reality of the objects in front of the user.
(3) **Inference:** In this stage, the system infers the user intent and progress by analyzing the sequence of observed states and actions. The system determines whether the user is successfully progressing through a workflow, if they have encountered a bottleneck, or if they are hesitating due to situational confusion. This intelligence is what allows the interface to shift from a rigid, pre-defined sequence to a dynamic assistant that knows when to offer help or adjust the information flow. By modeling the relationship between user action and physical state, the system can provide proactive support that aligns with the user actual pace and situational judgment.
(4) **Synthesis:** The system generates spatial XR elements, such as 3D markers, text labels, and animations, directly within the user's field of view. Unlike traditional XR that pulls from a library of existing assets, this stage uses generative logic to create context-aware information on the fly, so that the content displayed communicated the guidance as clearly as possible. These elements are rendered to align with the physical objects to ensure that the guidance is clear and actionable for the user's specific perspective.

3.4 Foundations and Lessons from Prior Systems

The proposed framework builds on our prior research, which has explored specific segments of the perception-generative loop, the development and evaluation of which has better helped inform this position.

XaiR [16] addresses the Perception, Identification, and Inference aspects of the loop. It integrates multimodal Large Language Models

(LLMs) with XR devices to provide context-aware assistance by processing multi-modal sensor streams. By leveraging egocentric audio, visual, and depth data, XaiR identifies user actions like picking up a tool or approaching a specific appliance and provides personalized feedback anchored as spatial arrows. This system demonstrates the power of state-aware assistance where the system acts as an observer that understands what is happening in the physical space and provides textual feedback on how to complete a task as well as creates 3D anchors to objects in the scene that are needed to complete the task. However, XaiR is limited to linear instruction following. It primarily checks for instruction completion, such as verifying if a user picked up the right tool or identifying if they have the wrong one, rather than a generative assistant that can restructure the task flow in real time.

In contrast, GenAssist (to appear) focuses on the Synthesis stage by investigating how high-level natural language prompts can be translated into interactive XR content at runtime. GenAssist moves beyond static markers by utilizing a generative engine to produce XR programs. This includes dynamic animations, user interactions, and adaptive spatial layouts. This enables a zero-authoring workflow where an expert can describe a task in plain text and the system handles the 3D asset generation and spatial anchoring. While GenAssist is a powerful text-to-XR pipeline, it cannot yet bridge the gap between physical state and XR content synthesis. It can create an animation of an arrow twisting if asked directly, but it cannot yet go from the observation that a screw needs to be loosened to the creation of that arrow animation.

By mapping these systems onto our proposed lifecycle, we see that the full potential of Generative Situated XR Assistance lies in bridging these two capabilities into a unified, closed-loop architecture. This integration would allow the system to move from simply seeing the environment or simply generating content to a state where the content itself is a direct, verified response to perceived physical changes.

4 Requirements and Implications

We identify the following requirements, challenges and implications for the next generation of situated assistance.

4.1 The Shift from States to Policies

Generative, situated interfaces challenge traditional instruction authoring practices that rely on specifying fixed interface states and transitions. When interfaces are generated at runtime based on environmental dynamics, designers must instead reason about higher-level intent, stylistic constraints, and safety boundaries. This shifts the creative process from drawing static assets to engineering the underlying logic that governs how those assets are synthesized. Consequently, this raises fundamental questions about how design knowledge is specified, communicated, and validated when the human designer never explicitly authors the final interface.

4.2 Agency and Responsibility

As authoring shifts to generative logic, system behavior becomes less predictable. In high-stakes physical tasks, the black box nature of generative AI is a liability. Users should not be forced to follow instructions they do not understand or trust. We must consider how users can maintain agency through mechanisms that allow them to inspect the system reasoning or override guidance when it conflicts with their situational judgment. This necessitates a new framework for accountability to clarify whether a failure arises from an error in perception, generative logic, or a user misinterpretation of intent.

4.3 Authoring: Personalization and Adaptation

A generative system must move beyond generic documentation to provide personalized assistance. This requires grounding generative logic in a dynamic user model that exists outside the immediate visual frame. Beyond sensing environmental information (e.g., what tools are available, and what objects the user is working with), a truly personalized assistant should be able to cater to a user's specific expertise and historical behavioral patterns. While context-aware XR has successfully used immediate sensing to show labels [8, 18, 19], we lack architectures that can synthesize long-term behavioral data into real-time spatial content.

By integrating long-term data, such as specific mechanical steps the user frequently overlooks or errors they tend to repeat, the system can shift its generative strategy from reactive labeling to proactive mentorship. The technical challenge lies in developing architectures that can retrieve and synthesize these personal histories into real-time spatial guidance, ensuring that the manual evolves at the same pace as the user's growing skill set.

4.4 Consumption: Continuous Inference and the Perception Gap

To provide personalized guidance, a system needs more than high-level object recognition. It must understand fine-grained state information, such as whether a component is partially installed or if a tool is being used incorrectly. Although scene understanding research has improved object detection [3, 12, 14, 15, 20] and data collection [4, 6], these efforts typically focus on identifying what is in the frame rather than what needs to happen next.

The primary challenge lies in bridging the gap between raw perception and instructional logic in real time. Knowing that a screw is halfway turned is only useful if the system can use that data to synthesize a contextually appropriate response. We must develop architectures that can convert *what is perceived* into a *generative action* for that specific moment.

4.5 Consumption: Verification and Spatial Reliability

In embodied XR, errors in generated guidance carry immediate physical consequences. This requires a higher standard of reliability than static or text-based assistance. We identify a dual layered risk in situated assistance. First, *semantic hallucinations* occur when the system generates logically incorrect or unsafe instructions [21]. Second, *spatial misalignment* occurs when a logically sound instruction is incorrectly grounded in the physical environment. While current vision systems can anchor to complete objects, they often fail to resolve the fine grained sub components like levers or buttons critical for task completion. We need new verification methods that verify both the logical intent and the spatial accuracy of an instruction before it is presented to the user. Bridging this gap requires new verification architectures that operate between perception and

display. Rather than relying solely on raw model outputs, these systems need internal sanity checks to ensure that a generated 3D annotation actually corresponds to the intended physical target within the current context. We must develop methods to verify both the logical intent and the spatial accuracy of an instruction before it is presented to the user.

5 Conclusion

We frame generative situated XR assistance as a form of Human-AI-UI interaction, where AI systems act as interface-generation agents that synthesize user interfaces from multimodal perception of the physical world. By reframing assistance as a dynamic lifecycle rather than a fixed artifact, we enable interfaces that are as fluid and context-aware as the physical tasks they support. Realizing this vision requires moving beyond the boundaries of traditional interface design. We must develop the frameworks and perception systems necessary to ensure that generative, personalized and embodied assistance remains safe, reliable, and fundamentally human-centered. This shift moves the focus from authoring content to defining the underlying generative logic that allows digital information to become an active participant in physical tasks.

References

[1] Andrea Bellucci, Alberto Ruiz, Paloma Díaz, and Igancio Aedo. 2018. Investigating augmented reality support for novice users in circuit prototyping. In *Proceedings of the 2018 international conference on advanced visual interfaces*. 1–5.

[2] Yining Cao, Peiling Jiang, and Haijun Xia. 2025. 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 (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 686, 20 pages. doi:10.1145/3706598.3713285

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294 [cs.CV] https://arxiv.org/abs/2104.14294

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. arXiv:1804.02748 [cs.CV] https://arxiv.org/abs/1804.02748

[5] Marius Fischer, Bernhard Fuerst, Sing Chun Lee, Javad Fotouhi, Severine Habert, Simon Weidert, Ekkehard Euler, Greg Osgood, and Nassir Navab. 2016. Preclinical usability study of multiple augmented reality concepts for K-wire placement. *International journal of computer assisted radiology and surgery* 11, 6 (2016), 1007–1014.

[6] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. arXiv:2110.07058 [cs.CV] https://arxiv.org/abs/2110.07058

[7] Steven J. Henderson and Steven Feiner. 2009. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *2009 8th IEEE International Symposium on Mixed and Augmented Reality*. 135–144. doi:10.1109/ISMAR.2009.5336486

[8] Gaoping Huang, Xun Qian, Tianyi Wang, Fagun Patel, Maitreya Sreeram, Yuanzhi Cao, Karthik Ramani, and Alexander J. Quinn. 2021. AdapTutAR: An Adaptive Tutoring System for Machine Tasks in Augmented Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 417, 15 pages. doi:10.1145/3411764.3445283

[9] Michael R. Marner, Andrew Irlitti, and Bruce H. Thomas. 2013. Improving procedural task performance with Augmented Reality annotations. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 39–48. doi:10.1109/ISMAR.2013.6671762

[10] Peter Mohr, David Mandl, Markus Tatzgern, Eduardo Veas, Dieter Schmalstieg, and Denis Kalkofen. 2017. Retargeting Video Tutorials Showing Tools With Surface Contact to Augmented Reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6547–6558. doi:10.1145/3025453.3025688

[11] Riccardo Palmarini, Iñigo Fernández Del Amo, Dedy Ariansyah, Samir Khan, John Ahmet Erkoyuncu, and Rajkumar Roy. 2023. Fast Augmented Reality Authoring: Fast Creation of AR Step-by-Step Procedures for Maintenance Operations. *IEEE Access* 11 (2023), 8407–8421. doi:10.1109/ACCESS.2023.3235871

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[13] Rafael Radkowski. 2015. Investigation of visual features for augmented reality assembly assistance. In *International conference on virtual, augmented and mixed reality*. Springer, 488–498.

[14] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714 [cs.CV] https://arxiv.org/abs/2408.00714

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640 [cs.CV] https://arxiv.org/abs/1506.02640

[16] 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. doi:10.1109/ISMAR62088.2024.00091

[17] Arthur Tang, Charles Owen, Frank Biocca, and Weimin Mou. 2003. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 73–80. doi:10.1145/642611.642626

[18] Arno Verstraete, Eva Geurts, and Maarten Wijnants. 2025. Does One-Size Training Fit All? Evaluating Adaptive Learning for VR Assembly Training. *Proc. ACM Hum.-Comput. Interact.* 9, 4, Article EICS010 (June 2025), 26 pages. doi:10.1145/3734187

[19] Giles Westerfield, Antonija Mitrovic, and Mark Billinghurst. 2015. Intelligent augmented reality training for motherboard assembly. *International Journal of Artificial Intelligence in Education* 25, 1 (2015), 157–172.

[20] 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

[21] Ada Yi Zhao, Aditya Gunturu, Ellen Yi-Luen Do, and Ryo Suzuki. 2025. Guided Reality: Generating Visually-Enriched AR Task Guidance with LLMs and Vision Models. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 146, 15 pages. doi:10.1145/3746059.3747784