Semantic-Augmented Reality: A Hybrid Robotic Framework Combining Edge AI and Vision Language Models for Dynamic Industrial Inspection

Ying-Jui Huang , Fu-Li Hsiao , Hsing-Wen Wang , Chih-Min Lo *

Posted Date: 24 December 2025

doi: 10.20944/preprints202512.2091.v1

Keywords: augmented reality; edge AI; vision language models; structured prompting; semantic-aware

 doi:10.20944/preprints202512.2091.v1

Article

Semantic-Augmented Reality: A Hybrid Robotic Framework Combining Edge AI and Vision Language Models for Dynamic Industrial Inspection

Ying-Jui Huang [1], Fu-Li Hsiao [1], Hsing-Wen Wang [2] and Chih-Min Lo [3,*]

[1] Institute of Photonics, National Changhua University of Education, Taiwan
[2] Department of Business Administration, National Changhua University of Education, Taiwan
[3] Department of Digital Multimedia Design, National Taipei University of Business, Taiwan
* Correspondence: cmlotw@ntub.edu.tw

Abstract

With the rise of Industry 4.0, Augmented Reality (AR) has become pivotal for human-robot collaboration. However, most industrial AR systems still rely on pre-defined tracked images or markers, limiting adaptability in unmodeled or dynamic environments. This paper proposes a novel **Interactive Semantic-Augmented Reality (ISAR)** framework that synergizes Edge AI and Cloud Vision-Language Models (VLMs). To ensure real-time performance, we implement a **Dual-Thread Asynchronous Architecture** on the robotic edge, decoupling video streaming from AI inference. We introduce a **Confidence-Based Triggering Mechanism**, where a cloud-based VLM is invoked only when edge detection confidence falls below a specific threshold. Instead of traditional image cropping, we employ a **Visual Prompting** strategy—overlaying bounding boxes on full-frame images—to preserve spatial context for accurate VLM semantic analysis. Finally, the generated insights are anchored to the physical world via **Screen-to-World Raycasting** without fiducial markers. This framework realizes a semantic-aware 'Intelligent Agent' that enhances **Human-in-the-Loop (HITL)** decision-making in complex industrial settings.

Keywords: augmented reality; edge AI; vision language models; structured prompting; semantic-aware

1. Introduction

1.1. Background and Motivation

The heart of Industry 4.0 is digitally-driven intelligence, with integrated robotics and AI signaling this shift. Intelligent teleoperated inspection robots are vital for reducing human presence in hazardous areas, lowering errors, and reaching confined spaces. Serving as perceptual interfaces between the physical and digital worlds, these robots are essential for detection tasks in extreme settings.

Modern inspection robots are equipped with vision modules for real-time object detection; advanced systems also integrate navigation, infrared ranging (distance measurement using infrared light), and multi-sensor fusion (combining data from multiple sensors). Although Edge Computing (local device data processing) and Large Vision Language Models (LVLMs, AI linking visual and linguistic data) are emerging in robotics, systems still depend on human interpretation and decision-making in complex, dynamic scenarios. A major limitation is that traditional teleoperation (remote machine control) relies on 2D video feedback, which lacks spatial context and depth. As a result, operators must mentally reconstruct the 3D environment, increasing cognitive load and reducing efficiency.

Consequently, AR has become a key Human-Machine Interface (HMI) to enhance operator perception. Recent studies show AR is crucial for human-virtual interaction, bridging task needs and tools through visualization [1]. However, despite advances in Robotic Autonomous Systems (RAS),

 doi:10.20944/preprints202512.2091.v1

most AR work still focuses on improving end-effector dexterity and precision [2]. Few frameworks use AR for semantic understanding and decision support in unstructured industrial environments.

1.2. Problem Statement

Despite advances in inspection robotics, three key challenges hinder their effectiveness in complex industrial applications:

The Semantic Gap in Computer Vision and Navigation. Traditional object detection algorithms, such as YOLO, can identify what and where an object is, but not how it is conditioned or why it is in a certain state. For example, a model may detect a valve but cannot diagnose issues like severe corrosion or abnormal pressure. Current navigation systems rely mainly on geometric mapping and obstacle avoidance, lacking high-level semantic awareness. This prevents them from interpreting contextual cues, such as a Toxic Hazard sign, to dynamically alter paths or avoid dangerous zones without physical barriers.

Limitations of AR Registration in Unstructured Environments: Industrial AR often depends on pre-set markers or maps for spatial tracking. These methods are fragile in dynamic or degraded environments, such as sewers, where sludge, smooth surfaces, or low light can cause feature extraction to fail. This loss of anchors renders marker-based AR ineffective.

Cognitive Overload and Limited Actionable Intelligence: Visual perception modules give operators detailed video feeds, but excess data can cause cognitive overload. Most AR interfaces act as passive displays, not intelligent assistants, lacking information filtering or actionable guidance for anomalies. Without effective decision support, operators may struggle to assess issues accurately, risking safety and efficiency.

1.3. Objectives and Proposed Approach

To address the challenges of semantic gaps and unstructured environments, this study proposes the **Interactive Semantic-Augmented Reality (ISAR)** framework. The specific approach is outlined as follows:

1. **Dual-Thread Asynchronous Perception:** To ensure low-latency performance on resource-constrained edge devices, we implement a **Dual-Thread Asynchronous Architecture**. This design decouples the video streaming task from the AI inference task, guaranteeing that the operator receives smooth FPV feedback regardless of the computational load from the object detection model.
2. **Confidence-Based Semantic Verification:** Instead of processing every frame via the cloud, we employ a **Confidence-Based Triggering Mechanism**. The system invokes the cloud-based VLM only when the edge detection confidence falls below a predefined threshold (e.g., 70%). To preserve spatial context, we utilize a **Visual Prompting** strategy—overlaying a bounding box on the full-frame image—rather than cropping the ROI, allowing the VLM to analyze the object within its environment.
3. **Markerless Spatial Registration:** To overcome the limitations of marker-based tracking, we employ **Screen-to-World Raycasting**. This technique maps the 2D detection coordinates to 3D spatial anchors, ensuring precise overlay of semantic information without pre-deployed markers.
4. **Human-in-the-Loop Control:** Finally, an **Action Manager** is introduced to close the decision loop. Operators can issue high-level commands based on AR-visualized semantic insights, establishing a robust **Human-in-the-Loop (HITL)** system.

 doi:10.20944/preprints202512.2091.v1

1.4. Main Contributions

The main contributions of this paper are summarized as follows:

1. **Proposal of an Asynchronous Hybrid Perception Architecture:** We propose a robust edge-cloud framework featuring a **Dual-Thread Asynchronous Architecture**. This design solves the latency bottleneck in real-time FPV transmission. Furthermore, by integrating a **Confidence-Based Triggering Mechanism**, we optimize computational resources, leveraging Cloud VLM intelligence only when Edge AI uncertainty is high.
2. **Development of Markerless Semantic AR Registration:** We develop a spatial registration technique combining deep learning coordinates with **Screen-to-World Raycasting**. This enables the precise anchoring of semantic insights in unmodeled, unstructured environments without fiducial markers.
3. **Integration of Visual Prompting for HITL Control:** We introduce a **Visual Prompting** strategy that retains full-frame context for VLM analysis, significantly enhancing the accuracy of semantic verification (e.g., correcting low-confidence detections). Coupled with our **Action Manager**, this forms a reliable **Semantic-Driven Human-in-the-Loop (HITL)** control system, transforming unstructured VLM outputs into actionable robot commands.

1.5. Paper Organization

The remainder of this paper is organized as follows: Section 2 reviews related work on industrial Augmented Reality, the integration of Vision-Language Models in robotics, and Human-in-the-Loop teleoperation. Section 3 details the methodology of the proposed ISAR framework, elaborating on the hybrid perception architecture, the screen-to-world raycasting algorithm for spatial registration, and the implementation of the Action Manager. Section 4 presents the experimental setup and a comprehensive analysis of the results, validating the system's performance in terms of latency and semantic decision-making capabilities. Finally, Section 5 concludes the study and outlines directions for future research.

2. Literature Review

2.1. Augmented Reality for Industrial Teleoperation

Teleoperation is essential for executing tasks in hazardous environments where human presence is risky. However, traditional interfaces that rely on 2D video feedback often result in high cognitive load, as operators must mentally reconstruct 3D spatial relationships from flat images [3]. To address this, Augmented Reality (AR) has been widely adopted to enhance situational awareness by overlaying navigation cues and robot states directly onto the physical view [4].

Despite these advantages, most industrial AR systems currently rely on fiducial markers (e.g., ArUco tags) or pre-built maps for spatial registration [5]. While effective in controlled settings, these methods face significant limitations in unstructured environments—such as sewers or disaster sites—where lighting is poor, and physical markers cannot be pre-deployed or may be obscured [6]. Therefore, developing markerless registration techniques that do not depend on prior environmental modeling is a critical research gap for enabling robust teleoperation in dynamic fields [7].

2.2. Embodied AI and Vision Language Models

The field of robotic perception has undergone a paradigm shift from closed-set object detection to open-vocabulary understanding. Traditional models, such as the YOLO series, are effective for specific tasks but struggle to recognize novel objects or interpret complex states absent from their training data [8–15]. In contrast, recent advancements in Vision Language Models (VLMs), such as PaLM-E and GPT-4V, leverage large-scale pre-training to achieve zero-shot generalization [16–20].

These models endow robots with semantic reasoning capabilities, allowing them not only to detect objects but also to diagnose their conditions (e.g., distinguishing between a "safe valve" and a "leaking

 doi:10.20944/preprints202512.2091.v1

valve") [21–27]. However, a significant barrier to their widespread adoption in teleoperation is latency. The high computational cost of querying cloud-based VLMs often disrupts the real-time feedback loop required for safe robotic control [28–30]. Furthermore, few studies have explored how to translate these unstructured semantic insights into structured AR visualizations for Human-in-the-Loop decision-making, highlighting a gap this study aims to bridge [31–33].

2.3. Edge-Cloud Collaboration and Human-in-the-Loop Control

Deploying computationally intensive VLMs directly on mobile robots is often infeasible due to limited onboard power and processing capabilities [39]. To address this, Edge-Cloud collaborative architectures have emerged as a standard paradigm, in which real-time tasks (e.g., obstacle detection) are executed locally at the edge, while heavy reasoning tasks are offloaded to the cloud [34–37]. This hierarchical approach optimizes bandwidth usage and ensures basic reactive capabilities even under network instability [38].

However, latency in cloud communication remains a challenge for continuous control[40]. Moreover, in safety-critical industrial inspections, relying solely on autonomous AI decision-making introduces significant risks[41]. Human-in-the-Loop (HITL) frameworks integrate human judgment into the robotic control loop, serving as a critical safeguard [42]. By keeping the human operator as the final authority, HITL systems ensure reliability while leveraging AI for enhanced perception.

Despite the maturity of teleoperation, most existing HITL systems function at a low level of abstraction (e.g., direct joystick control based on raw video)[43]. There is a notable scarcity of semantic-aware HITL frameworks that enable the robot not only to transmit video but also to provide structured, VLM-driven action recommendations (e.g., "Suggest Stop due to Gas Leak") for operator validation [44,45]. This study aims to bridge this gap by integrating an Action Manager that facilitates high-level, semantic interactions within the control loop.

Summary: While recent advancements in AR, VLMs, and Edge-Cloud architectures offer individual solutions, a unified framework that combines markerless semantic registration with interactive HITL control in unstructured environments remains lacking. The following section details our proposed methodology, the ISAR framework, designed to address these integrated challenges.

3. Methodology

3.1. System Architecture: Asynchronous Edge Processing

The proposed Interactive Semantic-Augmented Reality (ISAR) framework is designed as a distributed system comprising three distinct layers, as illustrated in Figure 1. To address the latency bottlenecks inherent in sequential processing, we redesigned the Perception Layer on the edge device (Raspberry Pi 4) using a **Dual-Thread Asynchronous Architecture**:

1. **Thread A (Streaming Task):** This high-priority thread captures raw frames from the camera and encodes them into binary data (MJPEG/Base64) for real-time FPV (First-Person View) transmission via WebSocket. This ensures the operator receives smooth, low-latency video feedback regardless of AI processing load.
2. **Thread B (Inference Task):** This thread runs the YOLO object detection model asynchronously. Upon completing an inference cycle, it updates a shared data structure with the latest BBox, Class ID, and Confidence Score.

The Data Processor then packages the image data from Thread A and the latest available inference results from Thread B into a unified JSON packet. This decoupling strategy guarantees that the visual stream remains real-time, while semantic labels are updated at the inference rate.

 doi:10.20944/preprints202512.2091.v1

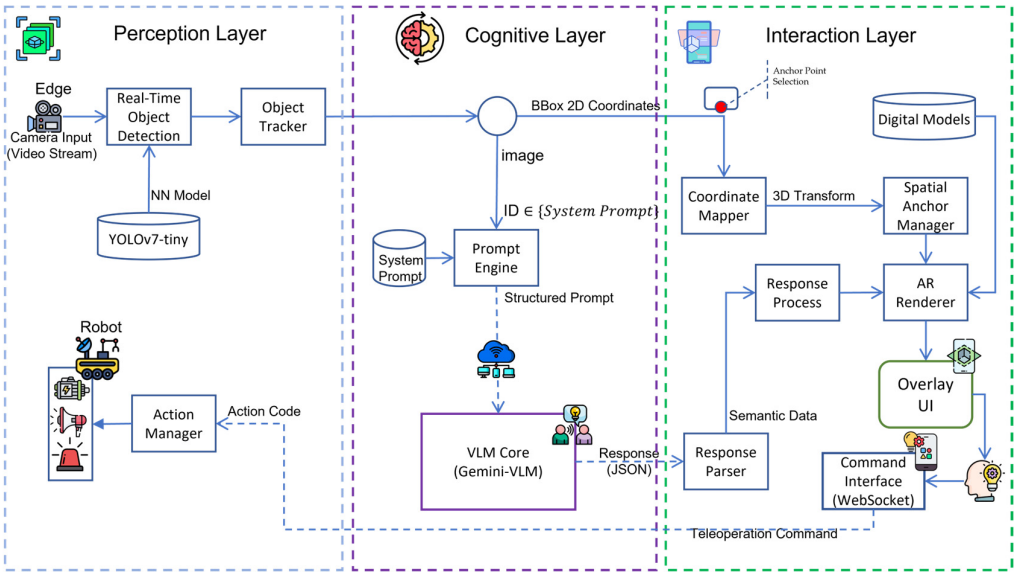


Figure 1. Comprehensive system architecture of the Interactive Semantic-Augmented Reality framework integrating Edge AI and Vision Language Models.

The architecture is structured into three layers: (1) The Perception Layer (Robot-side) executes real-time object detection using YOLOv7-tiny. (2) The Cognitive Layer (Cloud-side) utilizes Gemini-VLM to process complex visual semantics, returning standardized JSON recommendations via structured prompt engineering. (3) The Interaction Layer (User-side) overlays digital information onto the physical world using AR technology and provides an interface for users to validate VLM suggestions. Final control commands are transmitted back to the robot's Action Manager for execution, establishing a complete bi-directional control loop.

3.2. Cognitive Layer: Confidence-Based Semantic Verification

Instead of indiscriminately querying the VLM or relying on a static cache, we implement a **Confidence-Based Triggering Mechanism** to optimize computational resources and accuracy. The logic flow is as follows:

- **Data Parsing & Visualization:** Upon receiving the JSON packet, the system parses the object detection data. If valid detections exist, the Bounding Box (BBox) and Confidence Score are rendered onto the raw image.
- **Threshold Evaluation:** The system evaluates the Confidence Score (S).
- **IF $S \geq 70$:** The detection is considered reliable. The system bypasses the VLM and directly proceeds to spatial registration and AR rendering using the local YOLO labels.
- **IF $S < 70$:** The detection is flagged as ambiguous. The system triggers the VLM verification process.
- **Visual Prompting Strategy:** Unlike previous approaches that cropped the ROI, we retain the full-frame image to preserve spatial context. To focus the VLM's attention, we employ a **Visual Prompting** technique where the target BBox is visually drawn (e.g., a green rectangle) on the image sent to the VLM.
- **Prompt Engineering:** The Prompt Engine selects a specific query based on the YOLO Class ID (e.g., "Verify the object inside the green box. Is it a valve or a cap? Describe its condition.") and transmits it to the Gemini VLM via API.

 doi:10.20944/preprints202512.2091.v1

3.3. Markerless Spatial Registration Pipeline

To achieve precise semantic anchoring without fiducial markers, we implemented a coordinate transformation pipeline that bridges the robot's perception space and the user's AR space. The process consists of three formalized stages:

3.3.1. Anchor Selection and Coordinate Normalization

To ensure the AR information aligns with the physical footprint of the object, we select the bottom-center of the detected bounding box as the target anchor point P_{pixel} (u, v). However, a coordinate system discrepancy exists between the robotic vision module (typically following the standard image coordinate system with the origin at the top-left) and the AR rendering engine (Unity, utilizing a normalized viewport system with the origin at the bottom-left).

To bridge this gap, we define a **Normalization Function** Φ that maps the pixel coordinates (u, v) from the source image resolution (W, H) to the Normalized Device Coordinates (NDC) $(x_{ndc}, y_{ndc}) \in [0, 1]^2$. The transformation process is visually illustrated in Figure 2 and formalized below:

$$x_{ndc} = \frac{u}{W} \tag{1}$$

$$y_{ndc} = 1.0 - \frac{v}{H} \tag{2}$$

The term (1.0 - $\frac{v}{H}$) accounts for the vertical axis inversion (Y-flip) required to align the robot's downward-pointing Y-axis with the AR engine's upward-pointing Y-axis. The resulting vector $P_{ndc} = (x_{ndc}, y_{ndc})$ is then encapsulated in a JSON packet and transmitted to the AR application, ensuring resolution independence and correct spatial orientation.

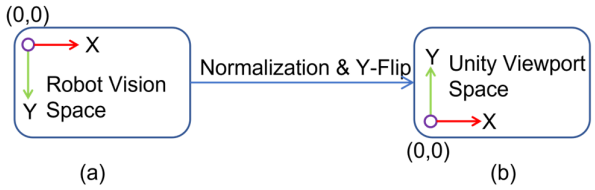


Figure 2. Schematic illustration of the coordinate system transformation from robotic vision to AR viewport. (a) The raw pixel coordinate system of the robot vision module (top-left origin); (b) The normalized viewport coordinate system of the Unity AR engine (bottom-left origin). The arrow indicates the mapping process from pixel space to Normalized Device Coordinates (NDC), incorporating the necessary vertical axis inversion (Y-flip).

3.3.2. Raycasting and Spatial Mapping

Upon receiving P_{ndc}, the AR system executes a Screen-to-World Raycasting operation. We define a ray generation function Ψ that projects a ray R from the AR camera's optical center O through the unprojected screen point:

$$R(t) = \Psi(P_{ndc}, K^{-1}) \tag{3}$$

where K represents the camera intrinsic matrix managed by the AR subsystem (e.g., AR Foundation). Finally, the spatial anchor A_{world} is determined by calculating the intersection between ray R and the detected environmental plane Π:

$$A_{world} = R(t) \cap \Pi \tag{4}$$

This approach effectively maps the 2D perception result to a 3D physical coordinate, allowing the semantic UI to maintain spatial coherence during user movement.

 doi:10.20944/preprints202512.2091.v1

3.4. Action Manager and HITL Control

The framework incorporates an Action Manager to close the control loop and enable Human-in-the-Loop (HITL) teleoperation, as shown in Figure 3. The process follows a specific workflow:

1. Recommendation: The AR interface parses the *action_code* from the VLM response (e.g., INSPECT) and displays a corresponding interactive button overlaid on the object.
2. Validation: The human operator evaluates the visual evidence provided by the VLM. If the operator agrees with the AI's diagnosis, they confirm the action via the UI.
3. Execution: The app transmits the confirmed command ID to the robot's Action Manager. The Action Manager maintains a Finite State Machine (FSM) and triggers the appropriate actuator behaviors (e.g., stopping motors, adjusting camera zoom, or logging the incident).

This mechanism ensures that high-level semantic decisions are verified by humans, mitigating the risks associated with potential AI hallucinations in safety-critical environments.

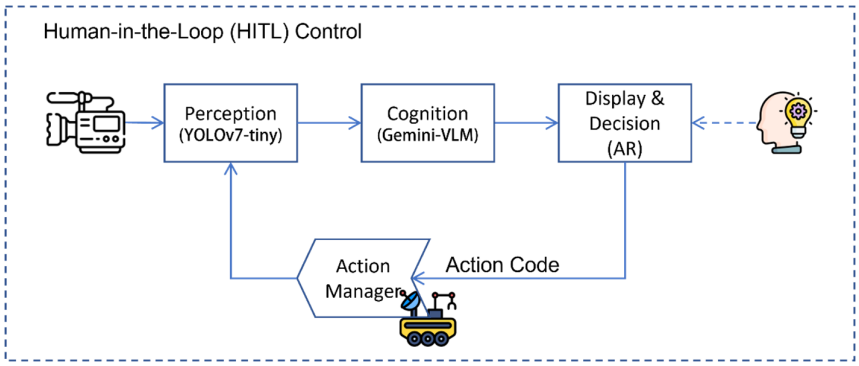


Figure 3. The proposed Human-in-the-Loop (HITL) control cycle for semantic-aware teleoperation.

The cycle consists of four phases: (1) Perception: The robot detects environmental states via Edge AI; (2) Cognition: The cloud-based VLM (Gemini) analyzes visual semantics and generates operational recommendations; (3) Display & Decision: The AR interface visualizes these insights, allowing the operator to make informed decisions and issue commands; and (4) Action: The robot executes the command via the Action Manager, modifying the environment and triggering the next cycle. This framework ensures operational safety and adaptability in complex scenarios.

4. Experiments & Results

To validate the feasibility of the proposed framework, we implemented the data flow architecture illustrated in Figure 4. On the robotic edge, ROS 2 (Robot Operating System) is employed as the middleware to orchestrate communication between the high-level computing unit (Raspberry Pi 4) and the low-level driver (RP2350B). Visual data and sensor states are serialized into JSON packets and transmitted in real-time to the mobile AR application via the WebSocket protocol. The AR App functions as a central bridge, forwarding visual queries to the Gemini VLM API and parsing the returned semantic recommendations. Consequently, the validated action commands are transmitted back to the robot's Action Manager node via WebSocket, triggering the RP2350B to drive the actuators, thus completing the closed-loop from perception to action.

 doi:10.20944/preprints202512.2091.v1

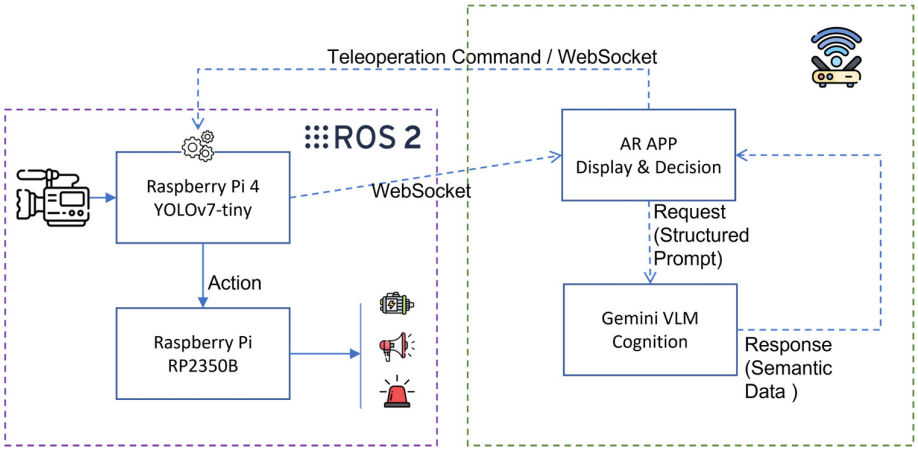


Figure 4. The implementation architecture and data flow diagram for experimental validation. The diagram illustrates the closed-loop communication pipeline connecting the ROS 2-based robotic nodes, the Unity AR application via WebSocket bridge, and the cloud-based Gemini VLM.

4.1. Experiment 1: Semantic Verification and Detail Refinement

Objective:

To validate the effectiveness of the VLM in resolving low-confidence detections ($S < 70$) and providing fine-grained details that YOLO misses.

Scenario:

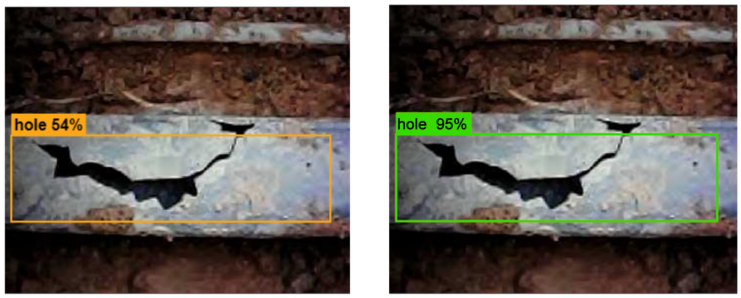
The robot inspected a rusted industrial valve in a dimly lit environment. YOLO detected the object as a "Valve" but with a low confidence score of 54% due to the rust and lighting conditions.

Process:

1. **Trigger:** Since the detected confidence score ***S*=54%** was below the predefined 70% threshold, the VLM pipeline was triggered. **Figure 5a** depicts this baseline detection provided by the Edge AI.
2. **Visual Prompt:** The full image with an **orange BBox** overlay was sent to Gemini.
3. **VLM Response:** The VLM returned the analysis: "The object inside the orange box is indeed a valve, but it shows signs of heavy corrosion on the handle. Status: Warning." and assigned a high verification confidence.

Result:

Upon receiving the VLM's confirmation, the AR interface updated the visualization. The label was effectively elevated from the initial unstable "Valve (54%)" to the high-confidence, semantically enriched label "Corroded Valve (95%)", as shown in Figure 5b. This experiment confirms that the VLM successfully acts as a verifier, correcting initial uncertainty and adding critical contextual status to the inspection.



(a) (b)

Figure 5. Comparison of Detection Results.

 doi:10.20944/preprints202512.2091.v1

4.2. Experiment 2: Semantic Navigation at Intersections

Objective: To demonstrate the **Human-in-the-Loop (HITL)** decision process where the system utilizes VLM capabilities to interpret traffic constraints and dynamically guide robot navigation.

Scenario: The robot approached a complex intersection. While the geometric path was clear, a traffic sign reading "STOP" was positioned at the front-left corner, implying a restriction on the left turn.

Process:

1. **Detection:** As shown in **Figure 6a**,**b**, the Edge AI successfully identified the scene as an "Intersection" with a high confidence score of **92%**. Concurrently, the visual system captured the "STOP" sign.
2. **Cognitive Analysis:** The system forwarded the scene context to the VLM. The VLM analyzed the spatial relationship between the intersection and the "STOP" sign, determining that the left path was restricted.
3. **Semantic Augmentation:** Instead of a simple 2D text warning, the system utilized the **Spatial Anchor Manager** to instantiate a virtual **3D Barrier** object directly onto the left path in the AR view, as illustrated in **Figure 6c**. This provides intuitive, immersive feedback to the operator.
4. **Decision Interface:** Based on the VLM's recommendation (Action Code: BLOCK_LEFT), the AR UI dynamically generated the control panel shown in **Figure 6d**. The system filtered out the restricted "Turn Left" option, presenting only valid commands: "**Straight**" and "**Turn Right**".
5. **Result:** The operator acknowledged the virtual barrier and selected "Turn Right". The robot executed the command safely. This experiment validates the framework's ability to translate high-level semantic understanding into actionable, safe control constraints in real-time.

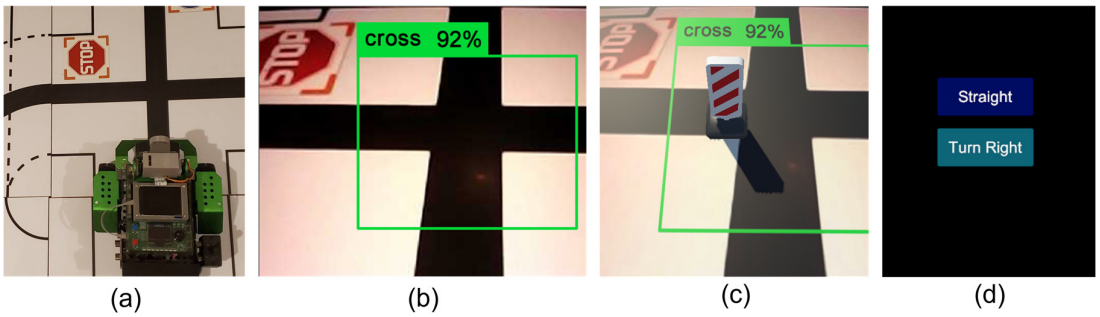


Figure 6. Sequential workflow of the Semantic Navigation experiment. (**a**) Third-person view of the robot positioned at an intersection; (**b**) The FPV AR interface displaying the initial AI detection of an "Intersection" (92%) and a physical "STOP" sign; (**c**) VLM-driven semantic augmentation, where a virtual **3D Barrier** is spatially anchored to the restricted path based on the analysis; (**d**) The generated HITL control interface presenting only valid directional commands ("Straight" and "Turn Right") for operator decision.

5. Conclusion and Future Work

5.1. Conclusion

This paper presented the **Interactive Semantic-Augmented Reality (ISAR)** framework, a novel approach designed to enhance robotic teleoperation in unstructured industrial environments. By integrating a **Dual-Thread Asynchronous Architecture** on the edge with a **Confidence-Based VLM Triggering** mechanism in the cloud, we successfully bridged the semantic gap while maintaining real-time responsiveness.

The experimental results validate the efficacy of our proposed methods:

1. **Real-time FPV Performance:** The dual-thread design effectively decouples image streaming from AI inference. This ensures that operators receive low-latency visual feedback regardless of the computational load from object detection tasks.

 doi:10.20944/preprints202512.2091.v1

2. **Optimized Resource Utilization:** The **Confidence-Based Triggering mechanism** balances the speed of Edge AI with the intelligence of Cloud AI. By invoking the VLM only when edge detection confidence falls below 70%, the system efficiently resolves ambiguity without incurring unnecessary latency or token costs.
3. **Enhanced Semantic Understanding:** The implementation of **Visual Prompting**, which retains the full-frame context rather than cropping, proved crucial. Experiments demonstrated that the VLM could effectively act as a verifier, elevating low-confidence detections (e.g., from 54% to 95%) and identifying fine-grained details such as corrosion.
4. **Semantic-Driven HITL Control:** The system demonstrated the ability to translate high-level semantic analysis into intuitive AR visualizations (e.g., **3D Barriers**) and constrained control interfaces. This establishes a robust **Human-in-the-Loop** control cycle, ensuring safe navigation in semantically complex environments.

5.2. Limitations and Future Work

Despite the promising results, this study has limitations. First, while the confidence-based mechanism reduces the frequency of cloud queries, the system still introduces a brief pause in decision-making when the VLM is triggered, which depends on network stability. Second, the current raycasting approach approximates spatial anchors on planar surfaces, which may lack precision on highly irregular 3D geometries.

For future work, we plan to explore two directions:

1. **Edge-VLM Deployment:** We aim to investigate quantized Vision Language Models (e.g., NanoLLM) that can run directly on the robot's NPU. This would eliminate network latency entirely, enabling continuous semantic analysis even in offline environments.
2. **Advanced 3D Reconstruction:** We intend to integrate **3D Gaussian Splatting** technology to achieve denser environmental understanding. This will allow for more precise semantic occlusion and interaction in complex, non-planar 3D spaces.

Author Contributions: Conceptualization, C. M. Lo and Y. J. Huang; methodology, C. M. Lo and Y. J. Huang; software, C. M. Lo and Y. J. Huang; validation, C. M. Lo and Y. J. Huang; formal analysis, F. L. Hsiao, C. M. Lo and H. W. Wang; investigation, Y. J. Huan; resources, F. L. Hsiao and H. W. Wang; data curation, Y. J. Huan; writing—original draft preparation, C. M. Lo and Y. J. Huang; writing—review and editing, H. W. Wang and F. L. Hsiao; visualization, C. M. Lo and Y. J. Huang; project administration, F. L. Hsiao; funding acquisition, H. W. Wang. All authors have read and agreed to the published version of the manuscript.

References

1. Van Krevelen, D.; Poelman, R. *Augmented Reality: Technologies, Applications, and Limitations*; Department of Computer Sciences, Vrije University Amsterdam: Amsterdam, The Netherlands, 2007.
2. Seetohul, J.; Shafiee, M.; Sirlantzis, K. Augmented Reality (AR) for Surgical Robotic and Autonomous Systems: State of the Art, Challenges, and Solutions. Sensors 2023, 23, 6202. https://doi.org/10.3390/s23136202
3. Zhao, F., Deng, W., & Pham, D. T. (2024). A Robotic Teleoperation System with Integrated Augmented Reality and Digital Twin Technologies for Disassembling End-of-Life Batteries. *Batteries*, *10*(11), 382.
4. Rosa-Garcia, A. D. L., Marrufo, A. I. S., Luviano-Cruz, D., Rodriguez-Ramirez, A., & Garcia-Luna, F. (2025). Bridging Remote Operations and Augmented Reality: An Analysis of Current Trends. *IEEE Access*, *13*, 36502-36526.
5. Tourani, A., Avsar, D. I., Bavle, H., Sanchez-Lopez, J. L., Lagerwall, J., & Voos, H. (2025). Unveiling the Potential of iMarkers: Invisible Fiducial Markers for Advanced Robotics. *arXiv preprint arXiv:2501.15505*.
6. Friske, M. D. (2024). Integration of Augmented Reality and Mobile Robot Indoor SLAM for Enhanced Spatial Awareness. *arXiv preprint arXiv:2409.01915*.
7. Su, Y. P., Chen, X. Q., Zhou, C., Pearson, L. H., Pretty, C. G., & Chase, J. G. (2023). Integrating virtual, mixed, and augmented reality into remote robotic applications: A brief review of extended reality-enhanced

 doi:10.20944/preprints202512.2091.v1

robotic systems for intuitive telemanipulation and telemanufacturing tasks in hazardous conditions. *Applied Sciences*, *13*(22), 12129.

8. Li, Hua Dong. (2025). Open-vocabulary object detection for high-resolution remote sensing images. Computer Vision and Image Understanding. 263. 104566. 10.1016/j.cviu.2025.104566.
9. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., & Zhang, L. (2023). A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1020-1031).
10. Li, Z., Xiang, Z., West, J., & Khoshelham, K. (2024). From Open Vocabulary to Open World: Teaching Vision Language Models to Detect Novel Objects. *arXiv preprint arXiv:2411.18207*.
11. Lin, J., Shen, Y., Wang, B., Lin, S., Li, K., & Cao, L. (2024, March). Weakly supervised open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 4, pp. 3404-3412).
12. Zhang, H., Xu, J., Tang, T., Sun, H., Yu, X., Huang, Z., & Yu, K. (2024, September). OpenSight: A simple open-vocabulary framework for LiDAR-based object detection. In *European Conference on Computer Vision* (pp. 1-19). Cham: Springer Nature Switzerland.
13. Hosoya, Y., Suganuma, M., & Okatani, T. (2024). Open-vocabulary vs. Closed-set: Best Practice for Few-shot Object Detection Considering Text Describability. *arXiv preprint arXiv:2410.15315*.
14. Zhu, C., & Chen, L. (2024). A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 8954-8975.
15. Zhang, H., Xu, J., Tang, T., Sun, H., Yu, X., Huang, Z., & Yu, K. (2024, September). OpenSight: A simple open-vocabulary framework for LiDAR-based object detection. In *European Conference on Computer Vision* (pp. 1-19). Cham: Springer Nature Switzerland.
16. Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Wahid, A., ... & Florence, P. (2023). Palm-e: An embodied multimodal language model.
17. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., ... & Han, K. (2023, December). Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning* (pp. 2165-2183). PMLR.
18. Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., ... & Finn, C. (2024). Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.
19. Li, Q., Liang, Y., Wang, Z., Luo, L., Chen, X., Liao, M., ... & Guo, B. (2024). Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*.
20. Sapkota, R., Cao, Y., Roumeliotis, K. I., & Karkee, M. (2025). Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*.
21. Wake, N., Kanehira, A., Sasabuchi, K., Takamatsu, J., & Ikeuchi, K. (2024). Gpt-4v (ision) for robotics: Multimodal task planning from human demonstration. *IEEE Robotics and Automation Letters*.
22. Hu, Y., Lin, F., Zhang, T., Yi, L., & Gao, Y. (2023). Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*.
23. Zhang, J., Wang, Z., Lai, J., & Wang, H. (2025). GPTArm: An Autonomous Task Planning Manipulator Grasping System Based on Vision–Language Models. *Machines*, *13*(3), 247.
24. Duan, J., Yuan, W., Pumacay, W., Wang, Y. R., Ehsani, K., Fox, D., & Krishna, R. (2024). Manipulate-anything: Automating real-world robots using vision-language models. *arXiv preprint arXiv:2406.18915*.
25. Hussain, S., Biswas, S., Dutta, A., Saad, M., Baimagambetov, A., Saeed, K., & Polatidis, N. (2025). A Review of Advances in Large Language and Vision Models for Robotic Manipulation: Techniques, Integrations, and Challenges. *SN Computer Science*, *6*(6), 588.
26. Team, G. R., Abeyruwan, S., Ainslie, J., Alayrac, J. B., Arenas, M. G., Armstrong, T., ... & Zhou, Y. (2025). Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*.
27. Team, G. R., Abdolmaleki, A., Abeyruwan, S., Ainslie, J., Alayrac, J. B., Arenas, M. G., ... & Wulfmeier, M. (2025). Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*.
28. Li, P., An, Z., Abrar, S., & Zhou, L. (2025). Large language models for multi-robot systems: A survey. *arXiv preprint arXiv:2502.03814*.

 doi:10.20944/preprints202512.2091.v1

29. Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*.
30. Saxena, S., Sharma, M., & Kroemer, O. (2023, August). Multi-resolution sensing for real-time control with vision-language models. In *2nd Workshop on Language and Robot Learning: Language as Grounding*.
31. Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364-381.
32. Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, *64*, 243-252.
33. Chai, C., & Li, G. (2020). Human-in-the-loop Techniques in Machine Learning. *IEEE Data Eng. Bull.*, *43*(3), 37-52.
34. Wang, Y., Yang, C., Lan, S., Zhu, L., & Zhang, Y. (2024). End-edge-cloud collaborative computing for deep learning: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, *26*(4), 2647-2683.
35. Rong, G., Xu, Y., Tong, X., & Fan, H. (2021). An edge-cloud collaborative computing platform for building AIoT applications efficiently. *Journal of Cloud computing*, *10*(1), 36.
36. Banitalebi-Dehkordi, A., Vedula, N., Pei, J., Xia, F., Wang, L., & Zhang, Y. (2021, August). Auto-split: A general framework of collaborative edge-cloud AI. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 2543-2553).
37. Kehoe, B., Patil, S., Abbeel, P., & Goldberg, K. (2015). A Survey of Research on Cloud Robotics and Automation. *IEEE Transactions on Automation Science and Engineering*, 12(2), 398–409.
38. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal*, 3(5), 637–646.
39. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., & Liu, H. (2024). The Rise of Large Language Models in Robotics: A Survey of Recent Advances. *IEEE Robotics and Automation Magazine* (Early Access / Preprint)
40. **Hu, G., Tay, W. P., & Wen, Y. (2012).** Cloud robotics: architecture, challenges and applications. ***IEEE Network***, 26(3), 21-28.
41. **Villani, V., Pini, F., Leali, F., & Secchi, C. (2018).** Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. ***Mechatronics***, 55, 248-266.
42. Sheridan, T. B. (2016). Human–Robot Interaction: Status and Challenges. Human Factors, 58(4), 525-532
43. **Losey, D. P., McDonald, C. G., Battaglia, E., & O'Malley, M. K. (2018).** A Review of Intent Detection, Arbitration, and Communication Aspects of Shared Control for Physical Human–Robot Interaction. ***Applied Mechanics Reviews***, 70(1).
44. Firoozi, R., et al. (2023). Foundation Models in Robotics: Applications, Challenges, and the Future. *Annual Review of Control, Robotics, and Autonomous Systems*, 7.
45. Garg, S., Sünderhauf, N., Dayoub, F., Morrison, D., Cosgun, A., Carneiro, G., ... & Milford, M. (2020). Semantics for Robotic Mapping, Perception and Interaction: A Survey. *Foundations and Trends® in Robotics*, 8(1-2), 1-224.