Journal of Industrial and Management Optimization

https://www.aimspress.com/journal/jimo

JIMO, 22(12): 6739–6773.
DOI:10.3934/jimo.2026225
Received: 24 July 2026
Revised: 02 October 2026
Accepted: 02 November 2026
Published: 23 November 2026

Research article

RoboAudit: Automated indoor safety compliance auditing via MLLM-guided semantic mapping

Osama Moh'd Alia [1], Tareq Alhmiedat [1,2,*], A'aeshah Alhakamy [1], Ashraf M. Marei [1,2], Fady Alnajjar [3], Saleh Albelwi [1], Ziyad Alenzi [1] and Manimurugan Shanmuganathan [1]

[1] Faculty of Computers and Information Technology, University of Tabuk, Tabuk 71491, Saudi Arabia

[2] Artificial Intelligence and Sensing Technologies (AIST) Research Center, University of Tabuk, Tabuk, Saudi Arabia

[3] College of Information Technology, United Arab Emirates University, Al Ain, Abu Dhabi, UAE

* **Correspondence:** Email: t.alhmiedat@ut.edu.sa.

Abstract: Ensuring safety in educational and industrial laboratories is vital due to the inherent risks associated with hazardous materials and machinery. Although mobile robots can automate physical monitoring, traditional vision systems lack the semantic reasoning required to evaluate regulatory compliance and physical accessibility. To address this limitation, we present an automated mobile robotic inspection framework designed for preliminary visual screening, semantic mapping, and automated safety compliance auditing of indoor safety assets. By integrating 2D LiDAR-based SLAM with a multimodal large language model (MLLM, specifically GPT-4o-mini), the system operates on an asynchronous architecture in which local edge hardware manages spatial mapping and data acquisition while offloading visual reasoning to the cloud. Operating as an intelligent decision-support tool for facility managers, the system evaluates essential safety assets, such as fire extinguishers, first-aid kits, and emergency exits, against a multi-tier evaluation framework measuring clipped asset availability (C_{asset}), regulatory rule compliance (C_{rule}), and physical accessibility. Validated across four real-world FabLab operational scenarios ranging from complete equipment absence to obstructed exit pathways, the system demonstrated 100% visual perception accuracy while successfully generating automated compliance reports and hazard warnings. Furthermore, spatial error analysis confirms the efficacy of decoupling high-level MLLM semantic classification from LiDAR-driven physical localization. Rather than replacing certified human inspectors, this work demonstrates the potential of MLLMs to deliver transparent, reliable preliminary safety auditing and automated compliance reporting in laboratory environments.

Keywords: mobile robots; semantic mapping; GPT-4o-mini; autonomous safety auditing; LiDAR SLAM
Mathematics Subject Classification: 65D19, 65Y04, 65Y04, 68T40

1. Introduction

Mobile robotics represents one of the fastest-growing subfields in artificial intelligence (AI) and automation, combining advanced perception, low-level control, and intelligent decision-making to automate complex physical inspection workflows across diverse operational environments. Unlike stationary manipulators, mobile platforms move through structured, semi-structured, and unstructured spaces to perform critical tasks such as environmental monitoring, industrial inspection, resource transport, and search and rescue [1].

Over the past decade, advancements in computer vision, simultaneous localization and mapping (SLAM), and machine learning have significantly expanded mobile platform capabilities across logistics, healthcare, agriculture, and smart manufacturing [2].

Automated mobile inspection plays a fundamental role in reducing human labor in hazardous or repetitive monitoring tasks. Traditional mobile inspection pipelines rely on planar or 3D LiDAR scanners, ultrasonic sensors, and depth sensors to reconstruct geometric representations of physical space. However, purely geometric maps lack the higher-level contextual understanding needed for sophisticated human-robot interaction, accessibility assessment, and domain-specific regulatory reasoning [3, 4].

To address this limitation, semantic mapping has emerged to integrate visual object recognition and scene understanding into spatial representations [5, 6]. Models such as Vision-Language Frontier Maps (VLFM) [7] demonstrate how zero-shot vision-language representations can ground natural language queries in geometric spatial frontiers, enabling robots to reason about unseen target objects.

Despite these advancements, safety asset verification and civil defense compliance monitoring in laboratory and industrial facilities remain heavily reliant on manual human oversight. In fabrication laboratory (FabLab) environments, ensuring compliance with safety equipment mandates, such as those governing the proper placement, density, and physical accessibility of fire extinguishers, emergency exits, and first-aid kits, is critical due to the presence of heavy machinery, electrical equipment, and volatile materials. Manual inspection procedures are labor-intensive, time-consuming, and prone to human error, particularly across large or complex facilities. While multimodal large language models (MLLMs) have demonstrated strong semantic reasoning capabilities, applying them to automated safety auditing requires linking high-level visual classification with rigorous, deterministic regulatory compliance rules.

To address these challenges, this paper presents RoboAudit, an automated mobile inspection and decision-support framework designed for preliminary visual safety screening, multi-tier compliance auditing, and semantic mapping within FabLab environments. Combining edge-based 2D LiDAR SLAM with cloud-based MLLM visual reasoning (leveraging OpenAI's GPT-4o-mini), the system operates on an asynchronous architecture that decouples low-level robotic spatial mapping and data acquisition from high-level visual semantic processing. The framework evaluates audited spaces against regional safety mandates by executing a multi-tier compliance assessment engine that measures clipped asset availability (C_{asset}), regulatory rule fulfillment (C_{rule}), and physical asset accessibility (A_i).

It is important to emphasize that this framework is designed as a preliminary visual screening and decision-support tool to assist facility managers and safety personnel, rather than a full legal replacement for certified human safety inspectors who perform physical validation (e.g., verifying extinguisher pressure gauges or inspecting first-aid kit seal dates).

The main contributions of this work are summarized as follows:

- A hybrid, decision-level sensor fusion architecture that combines edge-based 2D LiDAR SLAM with cloud-based MLLM vision, enabling low-latency mobile mapping and spatial data acquisition alongside high-level visual safety auditing.
- A multi-tier evaluation framework and deterministic decision gate (Algorithm 1) that quantify clipped asset availability (C_{asset}), category-level regulatory compliance (C_{rule}), and physical accessibility to generate structured audit reports and hazard advisories.
- A structured prompt and validation pipeline that enables zero-shot visual identification and accessibility assessment of essential safety assets without requiring specialized object-detection training.
- Empirical validation and error analysis across four real-world FabLab operational scenarios, demonstrating 100% visual perception accuracy and highlighting the benefits of decoupling semantic classification from LiDAR-based geometric localization.

The remainder of this paper is organized as follows: Section 2 reviews related work in MLLM-powered robotic navigation and semantic mapping. Section 3 details the regulatory safety standards and compliance requirements for FabLab environments. Section 4 describes the proposed RoboAudit system architecture and MLLM reasoning pipeline. Section 5 outlines the experimental setup, scenario evaluations, and quantitative results. Section 6 provides a detailed discussion of the findings, limitations, and design trade-offs. Finally, Section 7 concludes the paper and outlines future directions.

2. Related work

Visual perception and spatial mapping are essential capabilities for mobile inspection platforms operating in structured indoor environments. While traditional visual inspection systems demonstrate potential, many struggle with contextual and common-sense reasoning, performing reliably only in narrow operational settings. Methods relying purely on pre-constructed geometric maps or supervised object detectors often fail to evaluate physical accessibility hazards or interpret complex regulatory safety rules [8].

Mobile robotic frameworks have increasingly integrated large language models (LLMs) and multimodal LLMs (MLLMs) due to their capacity to process rich contextual and visual information [9]. LLMs enhance robotic inspection workflows by enabling systems to interpret human commands, reason about spatial layouts, and synthesize structured semantic knowledge from multimodal sensor inputs.

The rapid development of language and vision-language models has transformed mobile robotics by providing unified frameworks that link high-level semantic reasoning with low-level sensor perception. A comprehensive survey [10] highlighted how LLMs enhance perception, task planning, and human-robot interaction across diverse operational settings. This research established a taxonomy that categorizes LLM-driven robotics across environmental domains and execution phases, showing that vision-language representations enable context-aware decision-making beyond traditional rule-based algorithms.

In spatial reasoning, frameworks such as the Language Frontier Guide (LFG) [11] demonstrate how semantic guesswork can direct exploration and topological mapping. Similarly, multimodal frameworks such as Co-NavGPT [12] merge multi-view visual inputs into unified spatial representations for collaborative mapping, while LM-Nav [13] combines visual navigation models with CLIP image-text associations and GPT-3 to execute trajectory reasoning without labeled training data. Other systems, such as osmAG-LLM [14], use hierarchical semantic maps to guide mobile robots

around restricted or hazardous areas via natural-language reasoning. Further work on lightweight embedded implementations [15, 16] confirmed that edge platforms (e.g., Raspberry Pi) paired with visual understanding models can perform effective object classification and spatial safety monitoring in manufacturing spaces.

While traditional supervised object detectors (e.g., the YOLO series) excel at localization within closed class vocabularies, they lack the capacity for zero-shot contextual reasoning and rule evaluation required for compliance auditing. Furthermore, advanced parameter optimization techniques such as metaheuristic algorithms used to refine deep convolutional neural networks for visual classification tasks [17] have been explored to improve visual feature extraction, yet remain bound to closed-vocabulary training constraints. By leveraging open-ended MLLMs, our architecture bridges this gap, allowing the system to simultaneously verify asset presence, analyze spatial accessibility obstructions, and synthesize structured policy reports without requiring task-specific detector retraining.

Recent open-vocabulary semantic mapping and inspection approaches offer a direct baseline for contextualizing this work. Open-set perception pipelines leveraging GroundingDINO combined with lightweight SAM variants (e.g., MobileSAM) enable text-driven detection and segmentation of unlearned object categories [18, 19]. At the mapping tier, O3D-SIM constructs open-set 3D semantic instance maps for vision-language applications; osmAG-LLM fuses semantic maps with LLM reasoning for zero-shot goal identification; and OpenNav utilizes MLLMs to build bird's-eye-view representations for unstructured environments [14, 20, 21]. Concurrently, industrial inspection systems such as AutoInspect focus on long-horizon mapping, mission execution, and multi-week operational autonomy in industrial facilities [22]. Modern frameworks have further extended vision-language models to site safety auditing. For instance, Wang et al. [23] developed domain-knowledge-driven MLLM frameworks for automated visual safety inspection, while Hussain et al. [24] demonstrated vision-language assistants for intelligent onsite safety monitoring. In parallel, Hu and Gan [25] established semantic navigation architectures combining automated robotic inspection with indoor environmental quality monitoring. In contrast to these methods, our work targets a distinct objective: conservative, checklist-oriented safety auditing of indoor FabLab environments, where the primary deliverable is an interpretable compliance report and accessibility evaluation rather than unconstrained open-world navigation or long-horizon mission planning.

Taken together, these prior studies demonstrate that MLLMs and open-vocabulary perception are advancing robotic spatial auditing by unifying language-based reasoning with visual scene understanding. However, existing methods primarily focus on generic object retrieval, scene navigation, or dynamic exploration rather than explicit, checklist-driven auditing of safety compliance. To our knowledge, no prior system integrates LiDAR-based mobile mapping, MLLM visual verification, and deterministic rule-based reporting for FabLab safety compliance. This is a specific research gap this work fills.

A clear distinction exists between the goals of the proposed framework and the goals of open-vocabulary navigation frameworks. Approaches such as O3D-SIM, osmAG-LLM, and OpenNav focus on continuous spatial search, topometric scene graphs, and language-driven trajectory generation [14, 20, 21]. In contrast, our system adopts a decoupled, decision-level fusion architecture: 2D LiDAR SLAM constructs the geometric map and estimates robot poses along a designated inspection trajectory, while the MLLM conservatively evaluates the visual semantics of safety assets at fixed waypoints. Although simpler than complex 3D scene graph representations, this design provides the deterministic reliability and interpretability required for formal safety auditing.

Table 1 summarizes how the proposed framework differs from recent open-vocabulary mapping, navigation, and industrial inspection systems across key technical dimensions.

Table 1. Comparison of the proposed framework with recent open-vocabulary mapping, navigation, and industrial inspection systems in terms of perception capability, semantic updating, deployment scope, and safety-compliance evaluation.

Method	Primary focus	Open-vocab perception	Semantic update over revisits	Long-term deployment	Explicit safety compliance
GroundingDINO+ MobileSAM pipelines [18, 19]	Detection-segmentation	Yes	No	No	No
O3D-SIM [20]	Vision-language mapping	Yes	Limited	No	No
osmAG-LLM [14]	Zero-shot spatial mapping	Yes	No	No	No
OpenNav [21]	Open-world spatial mapping	Yes	No	No	No
AutoInspect [22]	Industrial inspection autonomy	Task-dependent	Yes	Yes	No
Wang et al. [23]	Construction safety inspection	Yes	No	No	Visual only
Hussain et al. [24]	Onsite inspection assistant	Yes	No	No	Partial
Hu & Gan [25]	Semantic navigation & IEQ	Yes	Yes	Limited	No
Semantic updating system [26]	Exploration and map updating	Yes	Yes	Limited real-world validation	No
This work	Safety auditing in FabLab spaces	Limited, MLLM-based	No	No	Yes

3. Essential safety requirements for fabLab

Laboratories present high-risk conditions due to the presence of electrical systems, dangerous equipment, and hazardous materials. The absence of safety protocols in these areas creates significant threats to people, buildings, and the surrounding natural environment. Safety regulations exist to reduce laboratory risks by establishing controlled environments for secure operations.

The Civil Defense organization in the Kingdom of Saudi Arabia (KSA) creates and enforces safety standards for laboratory operations throughout the country. These safety regulations follow international standards and protect fabrication laboratory spaces through proper safety equipment and procedures. Adherence to these standards is critical, as they directly support both operational readiness and emergency preparedness. The Civil Defense regulations of the KSA require all laboratories to implement fundamental safety protocols that protect occupants, equipment, and facilities [27]. These

essential safety measures are as follows:

- Floor plan: A map showing occupancy, exits, corridor widths, and fire sectors.
- Emergency exits and signs: At least two clearly marked, illuminated, and unobstructed exits leading to a safe area.
- Fire alarm and detection systems: Smoke detectors, heat sensors, and alarm systems to provide early warning.
- Fire extinguishers: Portable extinguishers located at accessible positions.
- First-aid kit: A collection of supplies and medical equipment used to provide initial care or emergency treatment for injuries.

To ensure regulatory and perceptual clarity, this framework explicitly distinguishes between three exit-related concepts: (1) an **emergency exit sign**, which is the illuminated visual marker indicating an egress route; (2) an **emergency exit**, which represents the physical door or structural egress threshold; and (3) an **unobstructed exit path**, which refers to the continuous, hazard-free floor space required to safely reach the exit point.

According to Saudi labor law and regional guidelines, at least one visible, clearly labeled, and accessible first-aid kit is required per laboratory space [28, 29]. Table 2 presents essential safety requirements for FabLab environments based on KSA Civil Defense regulations. To establish clear operational boundaries, the table explicitly distinguishes the complete regional safety mandate from the three primary visual safety asset categories ($M = 3$) implemented and evaluated in this proof-of-concept system.

Table 2. Essential safety requirements for FabLab environments based on KSA Civil Defense regulations, explicitly distinguishing general mandates from the evaluated operational scope.

No.	Civil Requirement	RoboAudit Capability	Evaluated Scope	# Req.
1	Emergency exits & signs	Visual presence & path audit	Implemented (M_1)	2
2	Fire extinguishers	Visual presence & clearance check	Implemented (M_2)	2
3	First-aid kits	Visual presence & accessibility check	Implemented (M_3)	Min. 1
4	Floor plans	Visual spatial alignment	Out-of-Scope	N/A
5	Detection & alarm units	Visual presence check	Out-of-Scope	N/A
6	Fire hose reels	Physical pressure validation	Out-of-Scope	N/A

The robotic system serves as an automated preliminary inspection assistant. It traverses designated laboratory zones to detect and map visual assets, flagging missing or displaced items for human auditors who remain responsible for physical validation and legal certification. To evaluate compliance, safety officers traditionally conduct manual room-by-room inspections. RoboAudit automates this initial process: The robot traverses the designated inspection path, capturing visual images and 2D LiDAR range data at fixed waypoints to construct a semantic safety map. In this study, the proposed framework focuses on screening three core visual safety asset categories ($M = 3$): fire extinguishers, emergency exit markings, and first-aid kits. Recognizing these essential safety elements provides facility managers with actionable preliminary reports and enhances overall situational awareness.

4. Multimodal large language model (MLLM)-based semantic mapping system

Based on the safety requirements outlined in Section 3, the main objective of this work is to develop a mobile robotic inspection system capable of systematically validating and monitoring the safety checklist for FabLab areas. The robotic platform is designed to traverse designated inspection routes within the area of interest, identify potential hazards, and ensure compliance with established safety protocols, thereby enhancing overall safety, reducing human error, and providing near-real-time feedback to facility managers. Figure 1 presents the system architecture for robot-based safety inspection.

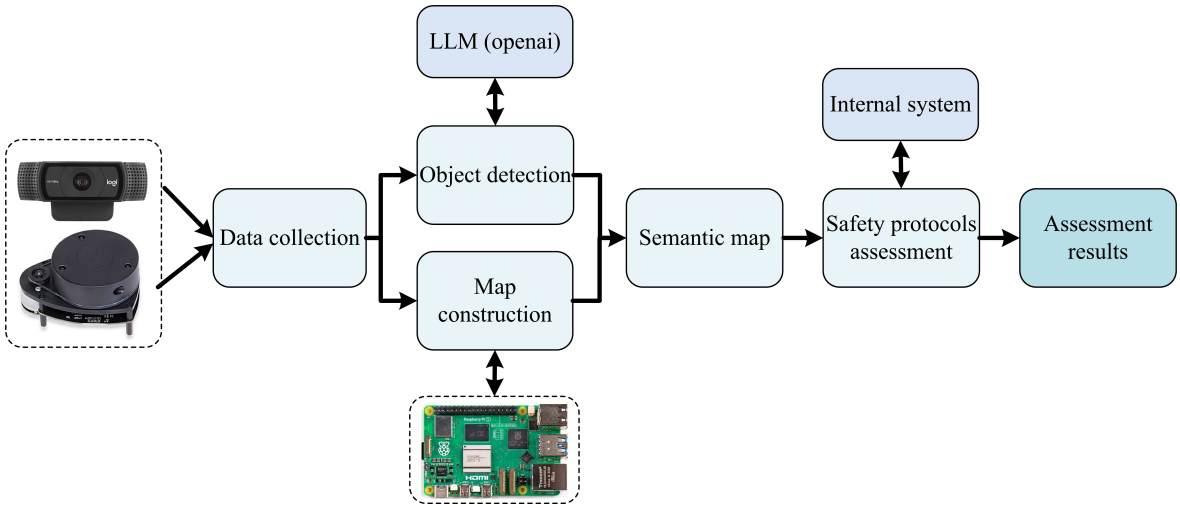


Figure 1. Overall architecture of the proposed robot-based FabLab safety inspection system, integrating data collection, object detection, semantic mapping, rule-based assessment, and compliance reporting.

The proposed system adopts a decision-level sensor fusion strategy, as illustrated in Figure 2. It integrates geometric mapping from LiDAR-based SLAM with high-level semantic interpretations from an RGB-based MLLM. By using a pose-based alignment mechanism that incorporates IMU and wheel odometry, the framework projects detected safety assets onto a global 2D semantic map by casting optical rays from visual bounding box centroids onto the LiDAR-generated occupancy grid to determine asset depth.

The localization of detected safety assets within the semantic map is achieved by associating visual detections with the robot's estimated pose. The robot's position and orientation are continuously computed using LiDAR-based SLAM, wheel odometry, and inertial measurements. In parallel, the onboard RGB camera captures data at fixed waypoints for the MLLM to identify objects such as fire extinguishers, first-aid kits, and emergency exits. During visual parsing, the prompt protocol directs the MLLM to differentiate between the emergency exit sign (the visual marker), the physical emergency exit door, and the floor space comprising the unobstructed exit path. Because the RGB camera lacks native per-pixel depth sensing, the framework determines asset depth by intersecting visual centroid rays with the local 2D LiDAR occupancy grid geometry. This results in a probabilistic placement of assets, producing a semantically enriched environmental representation suitable for safety auditing.

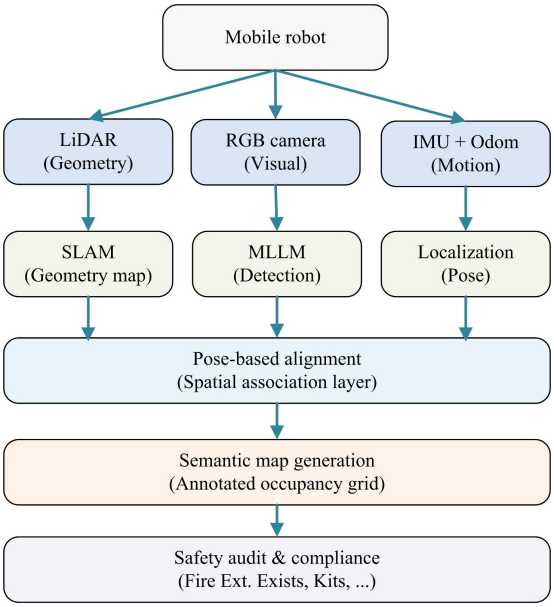


Figure 2. Sensor fusion pipeline of the proposed robotic system, combining LiDAR-based geometric mapping, RGB-based semantic detection, and pose information for safety-oriented semantic map generation.

4.1. Image-to-map projection and spatial deduplication

To map visual semantic detections into global geometric coordinates, the framework integrates hardware timestamp synchronization, intrinsic/extrinsic camera-LiDAR transformations, and occupancy grid ray-casting.

RGB image frames (1920×1080 resolution) captured at discrete waypoints are synchronized with 2D LiDAR scans via timestamp matching ($\Delta t < 50$ ms). The spatial alignment between the camera coordinate frame and the robot's base origin is established by a static extrinsic transformation matrix $\mathbf{T}_{\text{cam}}^{\text{robot}} \in \text{SE}(3)$, obtained via offline checkerboard calibration.

For every safety asset i detected by the MLLM, the model outputs 2D bounding box coordinates $[y_{\min}, x_{\min}, y_{\max}, x_{\max}]$ in normalized pixel space $[0, 1000]^2$, yielding a visual centroid (u_i, v_i) in image coordinates. To derive the object depth d_i, a 2D ray is cast from the optical center along the bearing angle $\phi_i = \arctan 2(u_i - c_x, f_x)$ onto the local 2D LiDAR occupancy grid map. To differentiate the detected asset from the supporting wall or background structure, the ray-caster identifies the range d_{hit} to the first occupied grid boundary along bearing ϕ_i. Because safety assets (e.g., fire extinguishers and exit signs) are wall-mounted or stand immediately adjacent to boundaries, object depth is calculated as $d_i = d_{\text{hit}} - \delta_{\text{offset}}$, where $\delta_{\text{offset}} = 0.20$ m accounts for standard physical mounting relief.

To project this 2D visual centroid into global 2D map coordinates, the transformation is factorized into three dimensionally consistent operations:

1. **3D Optical Back-Projection:** Using the camera intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ and the derived depth d_i, the 3D position of the asset centroid in the camera optical frame $\mathbf{p}_{\text{cam}}^{(i)} = [x_c, y_c, z_c]^T$ is computed as

$$\mathbf{p}_{\text{cam}}^{(i)} = d_i \cdot \mathbf{K}^{-1} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} \tag{4.1}$$

2. **SE**(3) **Extrinsic and Robot Pose Transformation:** The point $\mathbf{p}_{\text{cam}}^{(i)}$ is transformed from the optical frame to the global 3D map coordinate system using static extrinsics $\mathbf{T}_{\text{cam}}^{\text{robot}} \in \text{SE}(3)$ and the 2D SLAM robot pose $\mathbf{x}_r = (x_r, y_r, \theta_r)^T$:

$$\mathbf{p}_{\text{map}}^{(i)} = \mathbf{T}_{\text{robot}}^{\text{map}}(\mathbf{x}_r) \cdot \mathbf{T}_{\text{cam}}^{\text{robot}} \cdot \begin{bmatrix} \mathbf{p}_{\text{cam}}^{(i)} \\ 1 \end{bmatrix} \tag{4.2}$$

where $\mathbf{T}_{\text{robot}}^{\text{map}}(\mathbf{x}_r) = \begin{bmatrix} \mathbf{R}_z(\theta_r) & [x_r, y_r, 0]^T \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix} \in \text{SE}(3)$.

3. **2D Planar Map Projection:** Finally, the global 2D semantic map coordinate $\mathbf{x}_i = (x_i, y_i)^T \in \mathbb{R}^2$ is extracted by projecting $\mathbf{p}_{\text{map}}^{(i)}$ onto the horizontal SLAM navigation plane:

$$\mathbf{x}_i = \mathbf{p}_{\text{map},[1:2]}^{(i)} = \begin{bmatrix} x_m \\ y_m \end{bmatrix} \tag{4.3}$$

When multiple objects are detected within a single frame, each instance centroid is projected independently. To associate multi-view observations across overlapping frames and prevent landmark duplication, the candidate coordinate $\mathbf{x}_i$ of class k is merged with an existing mapped landmark centroid $\mathbf{c}_k^{(j)}$ if

$$\|\mathbf{x}_i - \mathbf{c}_k^{(j)}\|_2 \leq \epsilon_{\text{cluster}}, \tag{4.4}$$

where $\epsilon_{\text{cluster}} = 1.2$ m. The 1.2 m threshold accounts for cumulative SLAM drift (≈ 0.35 m) and monocular depth estimation uncertainty (≈ 0.45 m), ensuring robust multi-view data association without risk of false merges, as identical physical safety fixtures are required by regional regulations to maintain spatial separation well in excess of 3.0 m. Furthermore, because association operates strictly on a per-class basis, distinct adjacent safety assets (e.g., an exit sign mounted near a fire extinguisher) are evaluated independently regardless of spatial proximity. Merged observations maintain a single spatial coordinate $\mathbf{c}_k^{(j)}$ while updating aggregate accessibility via $A_j = \min(A_j, A_i)$.

Multi-Stage Asset Inventory Definitions: To ensure strict mathematical clarity between raw visual perception, spatial mapping, and regulatory compliance metrics, the framework defines asset counts across three distinct processing stages for each asset category $k \in \mathcal{K}$:

1. **Raw Frame Detections ($N_{\text{obs},k}$):** The total raw count of visual bounding box detections across all N captured RGB image frames:

$$N_{\text{obs},k} = \sum_{f=1}^{N} \sum_{i=1}^{D_f} \mathbb{I}(\text{class}(i) = k), \tag{4.5}$$

   where D_f is the number of assets detected in frame f, and $\mathbb{I}(\cdot)$ is the indicator function.
2. **Deduplicated Mapped Landmarks ($N_{\text{map},k}$):** The count of unique physical assets placed on the global 2D semantic map following spatial clustering via Equation (4.4):

$$N_{\text{map},k} = \left| \left\{ \mathbf{c}_k^{(j)} \;\middle|\; j \in \{1, \ldots, M_k\} \right\} \right|, \tag{4.6}$$

   where M_k is the total number of distinct spatial clusters formed for class k.

3. **Regulatory Compliance Inventory ($N_{\mathbf{comp},k}$):** The effective count used for compliance score evaluation (C_{asset}), bounded by the mandated threshold $N_{\text{req},k}$ to eliminate over-compliance score inflation:

$$N_{\text{comp},k} = \min(N_{\text{map},k}, N_{\text{req},k}), \tag{4.7}$$

MLLM Inference Parameters and Reproducibility: All visual reasoning queries were executed using the OpenAI API with model version `gpt-4o-mini-2024-07-18`. To eliminate decoding stochasticity and maximize output determinism across inspection runs, inference hyperparameters were set to temperature $T = 0.0$ (greedy sampling), `max_tokens` = 500, and $top_p = 1.0$. System responses were strictly constrained using Pydantic JSON-schema response formatting. Across 3 repeated evaluation runs on the dataset, zero JSON parsing errors or schema violations occurred.

Latency Profile, Data Minimization, and Network Fault Tolerance: To evaluate cloud dependency and operational continuity, end-to-end latency was benchmarked across all waypoint inspections. For image payloads downsampled to 768×768 JPEG resolution ($\sim$ 180 KB) and transmitted over facility Wi-Fi, we measured an average API processing latency of 1.85 ± 0.32 seconds per frame, encompassing edge pre-processing, image compression, network transmission, cloud inference, and structured payload retrieval. Concurrently, the edge-based ROS 2 SLAM localization loop executed at 28 ms per cycle (35 Hz), maintaining uninterrupted vehicle control. Before transmission, an onboard OpenCV module executes automated face-blurring and computer-screen masking to anonymize captured frames.

To handle transient network disruptions, the system incorporates an asynchronous local FIFO queue buffer. Up to 50 frame payloads can be cached on local edge storage during Wi-Fi outages; upon reconnection, queued requests automatically re-transmit with exponential backoff retries (a maximum of 3 attempts with a 10-second timeout) over TLS 1.3-encrypted HTTPS connections without stalling mobile navigation.

Economic Inference Cost: In the economic feasibility assessment, `gpt-4o-mini-2024-07-18` incurred an average inference cost of \$0.00015 per image frame. A complete 64-frame facility audit was executed at a total computation cost of approximately \$0.0096 ($\approx$ \$0.01), demonstrating high scalability for repeated, high-frequency industrial safety monitoring.

4.2. Standardized physical accessibility criteria

To prevent uncalibrated numerical outputs and provide objective grounding for visual assessment, the MLLM evaluates physical accessibility (A_i) using a standardized three-tier qualitative categorization matrix, defined in Table 3.

To eliminate ambiguity in visual reachability estimation, the heuristic parameter ranges $A_i \in [0.0\%, 100.0\%]$ evaluated by the MLLM are defined according to standardized physical clearance heuristics:

- **Unobstructed** (80.0% – 100.0%)**:** Clear physical access; complete 360° visibility; no fixed or moveable obstacles restricting the regulatory egress pathway or asset perimeter. Non-traversing floor elements (e.g., cables bundled neatly along wall perimeters) maintaining clear walk-through width are rated as Unobstructed ($\geq 80.0\%$), thereby meeting the requirement ($A_{\text{thresh}} \geq 80.0\%$).
- **Partially Obstructed** (50.0% – 79.0%)**:** Egress width or reachability directly restricted by temporary or moveable obstacles crossing the active pathway (e.g., loose boxes, chairs, or

unbundled/traversing floor cables). The asset is reachable but requires physical clearance (which triggers a maintenance ticket and restricts status to **Conditionally Approved**).

- **Severely Obstructed (**$< 50.0\%$**):** Egress route or access zone blocked by major fixed or heavy physical barricades (which triggers a hard accessibility failure forcing a **Not Approved** status).

The system prompt instructs the MLLM to evaluate visual scene features, specifically clear floor space, the presence of moveable or temporary obstacles (e.g., chairs, boxes, and cables), and direct line-of-sight clearance, mapping qualitative observations to defined ordinal categories, heuristic score ranges, and operational decision gates.

Table 3. Standardized three-tier physical accessibility categorization matrix.

Accessibility Category	Heuristic Range (A_i)	Visual / Physical Criteria	Accessibility Decision Flag
Unobstructed	80.0% – 100.0%	Full, clear physical path; no active path obstructions. Perimeter-located items (e.g., bundled cables along wall edges) that do not impede direct egress width.	Clears accessibility gate ($A_i \geq A_{\text{thresh}}$); enables Approved status.
Partially Obstructed	50.0% – 79.0%	Egress pathways or asset perimeters contain temporary/moveable items directly on active paths (e.g., chairs, loose boxes, traversing cables).	Triggers Maintenance Ticket; restricts decision status to Conditionally Approved.
Severely Obstructed	< 50.0%	Fixed or heavy physical barricades completely blocking reachability or emergency egress route.	Hard Failure ($A_i < 50.0\%$); forces decision status to Not Approved.

Building upon this enriched map and accessibility classification scheme, an internal judgment model evaluates the environment against laboratory safety standards. This module utilizes a rule-based system to perform automated comparisons between detected safety measures and regulatory requirements, such as the minimum number of fire extinguishers or exit markings. By processing these conditions in a predetermined order, the system identifies specific compliance gaps and points requiring intervention. This automated validation ensures consistent results that minimize human error while fulfilling national Civil Defense regulations.

5. Experimental results

The proposed framework integrates near-real-time perception data with semantic reasoning to evaluate regulatory safety compliance and construct semantic maps in dynamic laboratory environments. The experimental evaluation validates the system's capabilities across four key operational dimensions: (i) MLLM visual perception and object detection accuracy, (ii) asset availability compliance (C_{asset}), (iii) regulatory rule satisfaction (C_{rule}), and (iv) functional safety assessment including equipment reachability and obstruction detection. All experiments were conducted in an operational FabLab testbed under diverse environmental and compliance conditions.

5.1. *Experimental testbed*

A custom two-wheel-drive mobile robot platform was developed for automated mobile inspection, as shown in Figure 3. Its complete hardware architecture is illustrated in Figure 4. The system integrates a modular sensor array composed of four primary sensing units: an RGB camera, an RPLiDAR A1 range-finder, an MPU6050 gyroscope module, and wheel encoders. The RGB camera captures visual data at fixed inspection waypoints for object detection, classification, and visual safety parsing. Concurrently, the wheel encoders and gyroscope provide dead-reckoning odometry to estimate robot displacement and heading, thereby aiding in 2D spatial mapping and localization.



Figure 3. Developed two-wheel-drive mobile robot platform equipped with an RGB camera, LiDAR sensor, wheel encoders, gyroscope, and onboard processing unit for automated FabLab safety inspection.

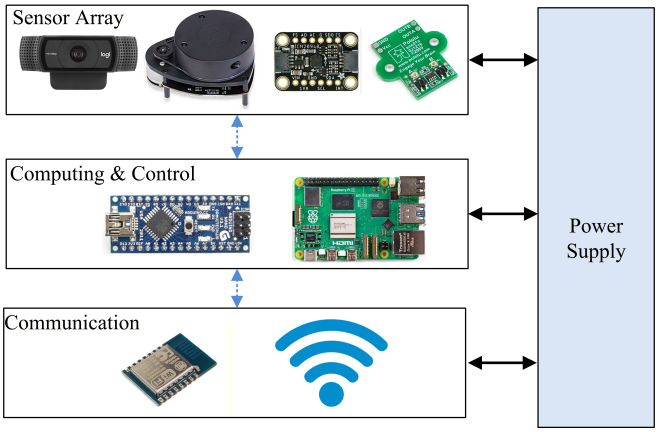


Figure 4. Hardware architecture of the developed robotic system, showing the integration of sensing, computing, control, communication, and power-supply components.

The RPLiDAR A1 works in conjunction with the odometry stream to generate environmental geometry, forming the occupancy grid map for localization along designated inspection paths. Detections from the visual perception pipeline are projected onto this geometric representation to construct a

semantically annotated map containing both spatial coordinates and compliance status. Technical details of the physical system are summarized in Table 4.

Table 4. Hardware specifications of the developed robotic system.

Parameter	Specification
Onboard Computing	Raspberry Pi 4 (4 GB RAM)
Vision Unit	Logitech 1080p Webcam (90° FOV)
Range-finder Unit	RPLiDAR A1 2D LiDAR
Orientation Sensor	MPU6050 Gyroscope
Odometry Sensing	Quadrature Wheel Encoders
Motor Controller	ODrive Motor Driver
Power System	32V LiFePO4 Battery (8,000 mAh)

The framework was validated in an operational FabLab facility within the Artificial Intelligence and Sensing Technologies (AIST) Research Centre at the University of Tabuk (Figure 5). The experimental facility encompasses typical digital fabrication equipment, including 3D printers, CNC routers, laser engravers, workbenches, and storage units, providing a representative environment to evaluate safety asset recognition, accessibility verification, and semantic mapping. In accordance with local civil defense standards, the mandated safety inventory threshold for the testbed facility is set to $T_i = 5$ total items across three regulatory categories ($|\mathcal{K}| = 3$): two fire extinguishers ($N_{\text{req, FE}} = 2$), two emergency exit signs ($N_{\text{req, EX}} = 2$), and one first-aid kit ($N_{\text{req, FA}} = 1$).



Figure 5. Real-world FabLab experimental testbed used to evaluate the proposed robotic safety inspection system under different compliance scenarios.

To evaluate perception performance, image frames captured at sequential trajectory waypoints ($N = 16$ per scenario, $N_{\text{total}} = 64$) were annotated against human ground-truth labels. Because frames along a continuous inspection path exhibit spatial autocorrelation, the evaluations are interpreted as trajectory-based spatial sampling metrics rather than independent and identically distributed observations.

5.2. System implementation

The software architecture is implemented as a multi-threaded Python framework running on Ubuntu Linux on the onboard Raspberry Pi 4 platform. The system coordinates concurrent sensor acquisition,

real-time spatial mapping, vision processing, cloud-based MLLM inference, and rule-based safety reasoning, as illustrated in Figure 6.

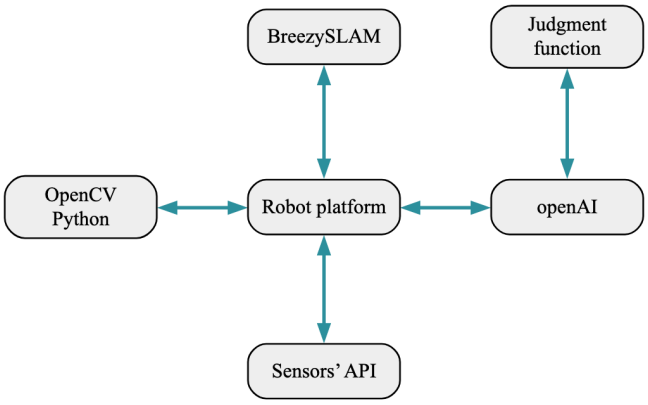


Figure 6. Software implementation of the developed system, illustrating the interaction between the robot platform, sensor APIs, SLAM module, image processing, OpenAI integration, and rule-based judgment function.

The software pipeline consists of six primary modules:

Robot Platform & Multithreading Driver: Interfacing directly with hardware controllers, this module manages asynchronous data streams from the RPLiDAR A1, MPU6050 gyroscope, wheel encoders, and USB webcam. Multithreading ensures high-rate sensor polling and prevents lockups during network interactions.

Sensor API Module: Developed in Python, this abstraction layer handles raw telemetry synchronization. It calculates wheel-encoder displacement, fuses gyroscope heading data, and timestamps visual and range measurements for spatial alignment.

BreezySLAM Module: A lightweight SLAM implementation processes RPLiDAR A1 range data to incrementally construct a 2D occupancy grid map. BreezySLAM maintains the robot's pose estimate (x, y, θ) in real time, serving as the geometric foundation for semantic annotations.

OpenCV Image Processing Module: This module manages frame capture, radiometric calibration, resizing, and spatial indexing. Each captured image is tagged with the robot's concurrent SLAM coordinates, preserving spatial context for visual detections.

OpenAI MLLM Integration: Captured RGB images are transmitted via API calls to OpenAI's `GPT-4o-mini` model. `GPT-4o-mini` was selected for its balance between multimodal reasoning accuracy and low latency. The MLLM parses visual scenes to identify safety assets, detect partial occlusions, evaluate equipment accessibility, and extract structured metadata using a multi-stage prompt protocol (detailed in Appendix A).

Judgment & Compliance Engine: To ensure deterministic and traceable safety assessments, raw semantic outputs from the MLLM are evaluated by a rule-based judgment function. Rather than relying on averaged scores, the engine evaluates compliance across distinct formal tiers: asset availability compliance (C_{asset}), regulatory rule compliance (C_{rule}), asset accessibility thresholds ($A_{\text{thresh}} = 80\%$), and final approval status (S_{final}). The total required inventory benchmark across all categories is set to $N_{\text{total,req}} = \sum_{k \in \mathcal{K}} N_{\text{req},k} = 5$ (two fire extinguishers, two emergency exit signs, and one first-aid Kit). The complete evaluation sequence is detailed in Algorithm 1.

Algorithm 1: Deterministic Compliance Decision Gate and Recommendation Generator

Input: Mapped landmark instances $\{i\}$, asset classes $\{k_i\}$, mandated class limits $\{N_{\text{req},k}\}$ (where $\sum N_{\text{req},k} = 5$), asset accessibility scores $\{A_i\}$, reachability threshold $A_{\text{thresh}} = 80\%$

Output: Compliance scores C_{asset}, C_{rule}, final judgment status S_{final}, recommendation list $\mathcal{R}$

// Step 1: Count Mapped Assets and Verify Egress Path Clearance
Initialize class map counts $N_{\text{map},k} \leftarrow 0 \quad \forall k \in \mathcal{K}$;
for *each mapped asset instance i with class* k_i **do**
  if k_i == *'emergency exit'* **then**
    // Requires both visual exit marker AND unblocked egress path ($A_i \geq A_{\text{thresh}}$)
    if $A_i \geq A_{thresh}$ **then**
      $N_{\text{map},k_i} \leftarrow N_{\text{map},k_i} + 1$;
    else
      Flag asset i as **Functionally Obstructed**;
    end
  else
    $N_{\text{map},k_i} \leftarrow N_{\text{map},k_i} + 1$;
  end
end
$N_{\text{comp},k} \leftarrow \min(N_{\text{map},k}, N_{\text{req},k}) \quad \forall k \in \mathcal{K}$;
$N_c \leftarrow \sum_{k\in\mathcal{K}} N_{\text{comp},k}$;
$N_{\text{total,req}} \leftarrow \sum_{k\in\mathcal{K}} N_{\text{req},k} \quad (= 5)$;
$C_{\text{asset}} \leftarrow \left(\frac{N_c}{N_{\text{total,req}}}\right) \times 100\%$;

// Step 2: Compute Regulatory Rule Compliance
$R_{\text{sat}} \leftarrow 0$;
for *each category* $k \in \mathcal{K}$ **do**
  if $N_{map,k} \geq N_{req,k}$ **then**
    $R_{\text{sat}} \leftarrow R_{\text{sat}} + 1$;
  end
end
$C_{\text{rule}} \leftarrow \left(\frac{R_{\text{sat}}}{|\mathcal{K}|}\right) \times 100\% \quad (|\mathcal{K}| = 3)$;

// Step 3: Evaluate Decision Gate with Tiered Accessibility Rules
$A_{\min} \leftarrow \min_i(A_i)$; // Worst-case accessibility across mapped assets
if $C_{rule} == 100\%$ **then**
  if $A_{\min} \geq A_{thresh}$ **then**
    $S_{\text{final}} \leftarrow$ Approved;
  else
    if $A_{\min} \geq 50\%$ **then**
      $S_{\text{final}} \leftarrow$ Conditionally Approved;
    else
      $S_{\text{final}} \leftarrow$ Not Approved ; // Severe obstruction ($A_i < 50\%$)
    end
  end
else
  $S_{\text{final}} \leftarrow$ Not Approved;
end

// Step 4: Generate Accessibility Warnings and Recommendations
$\mathcal{R} \leftarrow \emptyset$;
for *each mapped asset instance* i **do**
  if $A_i \geq A_{thresh}$ **then**
    Append status to $\mathcal{R}$: "Asset i: Easily accessible (A_i% reachable – Category: Unobstructed).";
  else
    if $A_i \geq 50\%$ **then**
      Append warning to $\mathcal{R}$: "Asset i: Path partially blocked (A_i% reachable – Category: Partially Obstructed – Action Required: Clear exit/access path).";
    else
      Append hazard to $\mathcal{R}$: "Asset i: Path severely blocked (A_i% reachable – Category: Severely Obstructed – Action Required: Clear path immediately).";
    end
  end
end
return $C_{\text{asset}}, C_{\text{rule}}, S_{\text{final}}, \mathcal{R}$;

5.3. Experimental scenarios

To validate the efficiency, robustness, and reasoning accuracy of the proposed semantic mapping framework, four distinct operational scenarios were designed and evaluated in the FabLab testbed. These scenarios simulate varying levels of regulatory compliance, equipment availability, and physical accessibility across $N = 16$ inspection waypoints per scenario ($N_{\text{total}} = 64$ frames):

Scenario 1 (Total Non-Compliance): All mandatory safety items (fire extinguishers, first-aid kits, and exit signs) were temporarily removed from the FabLab space. This scenario evaluates the system's capacity to correctly recognize a complete absence of safety assets, generate zero false positives ($N_{\text{obs}} = 0$, $N_{\text{map}} = 0$), report $C_{\text{asset}} = 0.0\%$ and $C_{\text{rule}} = 0.0\%$, and issue a **Not Approved** status in the semantic map and audit report.

Scenario 2 (Partial Asset Compliance): The environment was configured with only a subset of required assets, specifically one fire extinguisher ($N_{\text{map, FE}} = 1$) and one first-aid kit ($N_{\text{map, FA}} = 1$), with all emergency exit markers absent ($N_{\text{map, EX}} = 0$). This scenario tests the system's ability to accurately detect available equipment while evaluating rule-level satisfaction. It demonstrates how the judgment engine correctly flags unfulfilled regulatory mandates ($C_{\text{rule}} = 33.3\%$, as only the first-aid kit category is fully satisfied with $1/1$), assigning a **Not Approved** status despite partial asset presence ($C_{\text{asset}} = 40.0\%$, $N_{\text{comp}} = 2/5$).

Scenario 3 (Full Compliance and Accessibility): The FabLab was fully equipped with all mandatory safety items meeting regulatory quantities ($N_{\text{req}} = 5$: two fire extinguishers, two emergency exit signs, and one first-aid kit), all completely unobstructed ($A_i \geq 80\%$). This scenario verifies perception accuracy, spatial mapping alignment, and full regulatory authorization ($C_{\text{asset}} = 100.0\%$, $C_{\text{rule}} = 100.0\%$, $S_{\text{final}} =$ **Approved**).

Scenario 4 (Obstructed Asset / Conditional Approval): All required safety equipment was physically present ($N_{\text{map}} = N_{\text{req}} = 5$); however, one emergency exit path was deliberately obstructed by stacked storage boxes and maintenance equipment, reducing its functional accessibility to $A_i = 60.0\%$. This scenario evaluates the framework's ability to go beyond mere object detection by assessing functional accessibility. It tests whether the system correctly detects visual presence ($C_{\text{asset}} = 100.0\%$, $C_{\text{rule}} = 100.0\%$) while issuing accessibility warnings and assigning a **Conditionally Approved** status due to safety threshold violations ($50.0\% \leq A_{\min} < 80.0\%$).

5.4. Results

To evaluate system performance across diverse operational conditions while ensuring rigorous reproducibility, four distinct inspection scenarios were evaluated in the FabLab testbed. In each scenario, the mobile platform executed a predetermined 45-meter inspection trajectory at a constant velocity of 0.3 m/s. The robot followed a fixed sequence of waypoints, capturing 16 high-resolution RGB images (1920×1080 px) and concurrent 2D LiDAR scans at 3-meter intervals, yielding a dataset of 64 images across all trials. This fixed sampling strategy was chosen to minimize field-of-view overlap while ensuring complete visual coverage of the facility. By maintaining an identical traversal route across all four test scenarios, low-level navigation variability was eliminated, allowing the evaluation to focus directly on MLLM-driven semantic reasoning, spatial mapping, and regulatory compliance auditing.

To rigorously evaluate system performance, the framework formally distinguishes between four levels of evaluation:

First, Asset Availability Compliance (C_{asset}): This metric measures the physical presence of required equipment items relative to the total mandated count ($T_i = 5$ total items: two fire extinguishers, two emergency exits, and one first-aid kit):

$$C_{\text{asset}} = \frac{N_c}{T_i} \times 100\%, \tag{5.1}$$

where $N_c = \sum_{k \in \mathcal{K}} \min(N_{\text{map},k}, N_{\text{req},k})$ is the number of required items correctly identified and mapped, clipped to mandated class limits to prevent excess items in one category from masking deficiencies in another.

Second, Regulatory Rule Compliance (C_{rule}): This metric evaluates the proportion of distinct safety categories meeting regulatory criteria ($|\mathcal{K}| = 3$ categories):

$$C_{\text{rule}} = \frac{R_{\text{sat}}}{|\mathcal{K}|} \times 100\%, \tag{5.2}$$

where R_{sat} is the number of satisfied safety categories for which $N_{\text{map},k} \geq N_{\text{req},k}$.

To prevent false compliance assertions where an exit sign is detected over a non-functional or blocked opening, the evaluation engine enforces a tripartite validation protocol for category k = emergency exit. A physical exit requirement is satisfied in Equation (5.2) if and only if the visual indicator (`emergency exit sign` or `emergency exit door`) is detected and its associated `egress_path_assessment` achieves an accessibility score $A_{\text{exit}} \geq A_{\text{thresh}}(80\%)$. If a sign is detected but the physical path is obstructed ($A_{\text{exit}} < 80\%$), the regulatory category is flagged as **Functionally Obstructed**, preventing sign detection alone from satisfying physical compliance mandates.

Third, Final Approval Status (S_{final}): This metric enforces a strict deterministic decision gate evaluated by Algorithm 1, incorporating both regulatory rule fulfillment and tiered physical accessibility constraints:

$$S_{\text{final}} = \begin{cases} \text{Approved}, & \text{if } C_{\text{rule}} = 100.0\% \text{ and } \min_j(A_j) \geq 80.0\% \\ \text{Conditionally Approved}, & \text{if } C_{\text{rule}} = 100.0\% \text{ and } 50.0\% \leq \min_j(A_j) < 80.0\% \\ \text{Not Approved}, & \text{if } C_{\text{rule}} < 100\% \text{ or } \min_j(A_j) < 50.0\%, \end{cases} \tag{5.3}$$

where A_j denotes the evaluated physical accessibility rating of mapped asset j.

Fourth, Object Recognition Accuracy Evaluation Metrics: This evaluation assesses the capability of the developed system to precisely detect and recognize safety objects by computing standard detection measures for each checklist item (i.e., each object class), namely accuracy, precision, recall, and F1-score.

Accuracy (Acc) is defined as the proportion of correctly classified instances relative to the total number of evaluated instances:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%, \tag{5.4}$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Precision (P) quantifies the proportion of correctly detected safety objects (TP) among all instances predicted as positive by the system (TP + FP):

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\%, \tag{5.5}$$

Recall (R) measures the proportion of correctly detected safety objects among all ground-truth safety objects present in the scene (TP + FN):

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\%, \tag{5.6}$$

Finally, the F1-score (F1) is computed as the harmonic mean of precision and recall:

$$\mathrm{F1} = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}, \tag{5.7}$$

Unlike standard supervised object detectors that are limited to identifying predefined 2D bounding boxes, open-ended MLLM-based compliance checking enables high-level semantic reasoning such as assessing physical accessibility hazards and interpreting contextual safety rules. Consequently, the proposed framework extends beyond simple asset detection to automate complex, multimodal regulatory decision-making that traditional perception models cannot perform.

Perception Performance, Evaluation Units, and Matching Metrics

To rigorously evaluate the MLLM visual detection pipeline and ensure complete consistency across all experimental results, performance is measured across $N = 64$ camera image frames (16 frames collected per inspection scenario traversal across four distinct runs) and $M = 3$ target safety asset classes (*fire extinguisher*, *first-aid kit*, and *emergency exit*). The primary unit of evaluation is strictly established at the **frame-class pair level**, yielding a fixed evaluation domain:

$$\text{Total Evaluated Units} = 64 \text{ frames} \times 3 \text{ target classes} = 192 \text{ class-frame evaluation units} \tag{5.8}$$

To eliminate metric ambiguity, multiclass predictions and localized spatial bounding boxes are matched against human ground-truth annotations using a standardized confidence evaluation threshold of $\tau = 0.70$ (70%) and a bounding box intersection-over-union (IoU) threshold. The four outcomes are defined as follows:

- **True Positive (TP):** A detection whose predicted category label matches the ground-truth target class with an MLLM confidence score $\tau \geq 0.70$ **AND** has a bounding-box spatial overlap of IoU ≥ 0.50 relative to ground-truth annotations.
- **False Positive (FP):** An MLLM detection with confidence $\tau \geq 0.70$ occurring where no ground-truth asset exists, or a detection targeting a valid asset with IoU < 0.50, or a misclassified asset category label.
- **False Negative (FN):** A ground-truth target asset present in the camera field of view that is either missed entirely or predicted with a confidence score below the threshold ($\tau < 0.70$).
- **True Negative (TN):** The correct prediction that a target asset class is absent from a frame, accompanied by no spurious predictions ($\tau < 0.70$).

All precision, recall, accuracy, and $F1$-scores reported throughout the manuscript are calculated strictly using these 192 class-frame evaluation units. Table 5 provides the comprehensive class-frame ground-truth breakdown across all four experimental scenarios.

Table 5. Comprehensive ground-truth counts and image-level perception evaluation across scenarios.

Scenario	Asset Class	GT Objects	GT Frames	TP	FP	FN	TN	Class-Frame Acc. (%)
Scenario 1	Fire Extinguisher	0	0	0	0	0	16	100.0% (16/16)
(Empty)	Emergency Exit	0	0	0	0	0	16	100.0% (16/16)
	First Aid Kit	0	0	0	0	0	16	100.0% (16/16)
Scenario 2	Fire Extinguisher	1	4	4	0	0	12	100.0% (16/16)
(Partial)	Emergency Exit	0	0	0	0	0	16	100.0% (16/16)
	First Aid Kit	1	5	5	0	0	11	100.0% (16/16)
Scenario 3	Fire Extinguisher	2	9	9	0	0	7	100.0% (16/16)
(Full)	Emergency Exit	2	4	4	0	0	12	100.0% (16/16)
	First Aid Kit	1	3	3	0	0	13	100.0% (16/16)
Scenario 4	Fire Extinguisher	2	8	8	0	0	8	100.0% (16/16)
(Obstructed)	Emergency Exit	2	5	5	0	0	11	100.0% (16/16)
	First Aid Kit	1	3	3	0	0	13	100.0% (16/16)

5.4.1. Scenario 1: Total non-compliance

As detailed in Section 5.3, Scenario 1 evaluates the system's ability to detect complete safety non-compliance. All emergency equipment was temporarily removed from the FabLab environment (Figure 7).



(**A**) (**B**) (**C**)

Figure 7. Representative images captured in Scenario 1, showing the FabLab environment with safety equipment intentionally removed to evaluate non-compliance detection.

The captured visual frames were processed by the MLLM perception pipeline and passed to the rule-based judgment engine. Recognizing zero safety items across all frames, the system generated Report 1, resulting in $C_{\text{asset}} = 0.0\%$ (0 of 5 items present), $C_{\text{rule}} = 0.0\%$ (0 of 3 requirements met), and a final status of Not Approved.

Report 1: Judgment and Recommendation Report for Scenario 1
Safety tools detection report
Final Judgment: ✗ Not Approved Asset Availability (C_{asset}): 0.0% (0/5 items present) Rule Compliance (C_{rule}): 0.0% (0/3 requirements satisfied)
Total counts Fire Extinguishers: 0 (Required: 2) – UNSATISFIED Emergency Exits: 0 (Required: 2) – UNSATISFIED First Aid Kits: 0 (Required: 1) – UNSATISFIED
Recommendation • Action Required: Supply two fire extinguishers, two physical emergency exit doors with illuminated directional signage, and one first-aid kit to satisfy civil defense mandates.

Quantitative performance metrics for Scenario 1 are summarized in Table 6. Because all images were correctly classified as containing no safety assets (TN = 16), the system achieved 100.0% image accuracy (16/16). Precision, recall, and F1-score are reported as N/A due to the deliberate absence of positive target instances (TP = FP = FN = 0).

The 2D occupancy grid map generated by BreezySLAM was semantically annotated with the perception outputs, as shown in Figure 8. The absence of asset markers confirms that zero false-positive annotations were introduced into the spatial map.

Table 6. Performance metrics and feasibility evaluation of the MLLM system in Scenario 1 (N = 16 frames, 48 class-frame evaluation units).

Metric	Value (Raw Count)	95% Wilson Conf. Interval
Final Approval Status (S_{final})	Not Approved	–
Rule Compliance (C_{rule})	0.0% (0/3)	–
Asset Compliance (C_{asset})	0.0% (0/5)	–
Frame-Class Accuracy	100.0% (48/48)	[92.6%, 100.0%]
Precision	N/A	–
Recall	N/A	–
F1-score	N/A	–



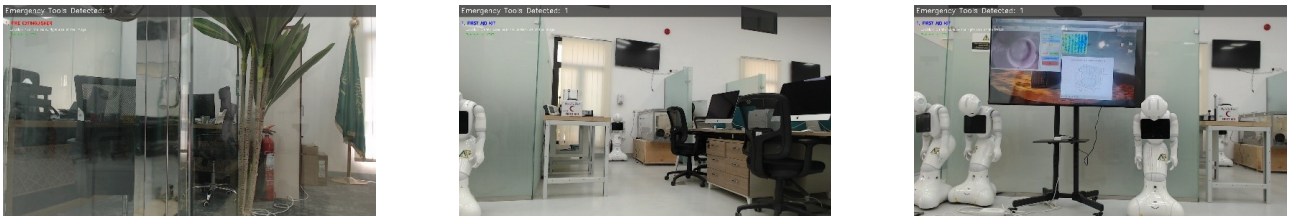
Figure 8. Constructed semantic map for Scenario 1, showing no annotated safety equipment detections due to the intentional absence of safety items.

5.4.2. Scenario 2: Partial asset compliance

The second scenario was conducted in a partially compliant FabLab setting containing only a subset of required safety equipment: one fire extinguisher and one first-aid kit, with all emergency exit markers absent. The MLLM perception pipeline successfully detected all physically present safety assets across captured frames (Figure 9).

Although only one physical first-aid kit was present in the room, it was visually detected across two overlapping image frames. During spatial indexing, the mapping engine clusters co-located detections via spatial distance thresholding (Section 4.1), consolidating them into a single physical landmark on the global map. Furthermore, during metric calculation, Equation (5.1) clips any excess detections to the mandated class limit ($N_{\text{req},k} = 1$).

Evaluating the detected items against the unified framework yielded an asset availability compliance score of $C_{\text{asset}} = 40.0\%$ (two of five required physical items verified present: one fire extinguisher and one first-aid Kit). However, at the regulatory rule level, only one out of three safety categories was satisfied (first-aid Kit requirement met [1/1]; fire extinguisher requirement unfulfilled [1/2]; emergency exit requirement unfulfilled [0/2]), resulting in a rule compliance score of $C_{\text{rule}} = 33.3\%$. Because mandatory safety categories remained unfulfilled ($C_{\text{rule}} < 100.0\%$), Algorithm 1 assigned a final status of **Not Approved**, as presented in Report 2.



(**A**) (**B**) (**C**)

Figure 9. Representative images captured in Scenario 2, showing a partially compliant FabLab environment with limited safety equipment available for detection and assessment.

Report 2: Judgment and Recommendation Report for Scenario 2
Safety tools detection report
Final Judgment: ✗ Not Approved Asset Availability (C_{asset}): 40.0% (2/5 items present) Rule Compliance (C_{rule}): 33.3% (1/3 requirements satisfied)
Total counts Fire Extinguishers: 1 (Required: 2) – UNSATISFIED Emergency Exits: 0 (Required: 2) – UNSATISFIED First Aid Kits: 1 (Required: 1) – SATISFIED
Recommendation • First aid kit (On table, right side): Easily accessible (90.0% reachable – Category: Unobstructed) • Fire extinguisher (Near plant, floor): Easily accessible (90.0% reachable – Category: Unobstructed) • **Action Required:** Supply one missing fire extinguisher and two physical emergency exit doors with illuminated directional signage to satisfy civil defense mandates.

System performance metrics for Scenario 2 are summarized in Table 7. At the visual detection level, the MLLM perception system achieved 100.0% precision, recall, F1-score, and accuracy across all $N = 16$ evaluated image frames.

Table 7. Performance metrics and feasibility evaluation of the MLLM perception system in Scenario 2 ($N = 16$ frames, 48 class-frame evaluation units).

Metric	Value (Raw Count)	95% Wilson Conf. Interval
Final Approval Status (S_{final})	Not Approved	–
Rule Compliance (C_{rule})	33.3% (1/3)	–
Asset Compliance (C_{asset})	40.0% (2/5)	–
Frame-Class Accuracy	100.0% (48/48)	[92.6%, 100.0%]
Precision	100.0% (16/16)	[80.6%, 100.0%]
Recall	100.0% (16/16)	[80.6%, 100.0%]
F1-score	100.0%	–

The constructed 2D occupancy grid map generated by BreezySLAM was semantically annotated with the detected assets, as shown in Figure 10. The map accurately projects the spatial locations of the detected fire extinguisher and first-aid kit onto the geometric layout, providing clear visual and spatial verification of site deficiencies.



Figure 10. Constructed semantic map for Scenario 2, showing annotated safety equipment detections in the partially compliant FabLab environment.

5.4.3. Scenario 3: Full compliance and accessibility

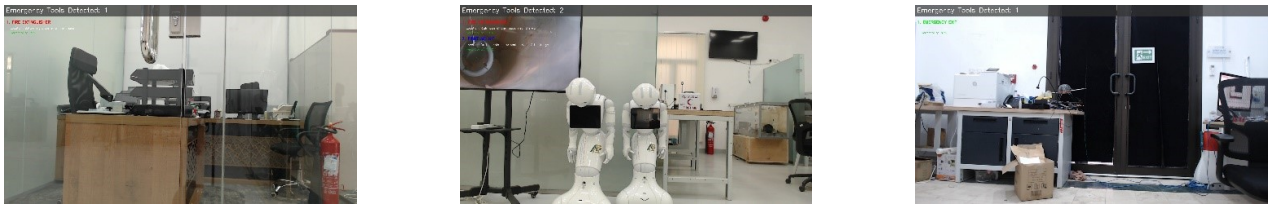
The third scenario evaluated the framework in a fully compliant FabLab environment. Input visual frames captured during the mobile robot's traversal are shown in Figure 11, with the corresponding MLLM perception outputs illustrated in Figure 12.

Across the facility scan, the perception system identified six fire extinguishers, two emergency exits, and one first-aid kit. Applying Equation (5.1) clips the detected fire extinguisher count to the mandated maximum of two required units, yielding $N_c = 5$ verified required items and an asset availability compliance score of $C_{\text{asset}} = 100.0\%$. Furthermore, all three regulatory categories were satisfied ($R_{\text{sat}} = 3$), resulting in $C_{\text{rule}} = 100.0\%$.



(**A**) (**B**) (**C**)

Figure 11. Original images captured from the FabLab environment in Scenario 3, provided as input to the MLLM-based system.



(**A**) (**B**) (**C**)

Figure 12. Annotated perception outputs generated for Scenario 3, highlighting detected fire extinguishers in red, first-aid kits in blue, and emergency exits in green.

Because all items satisfied accessibility thresholds ($A_i \geq 80.0\%$), Algorithm 1 assigned a final approval status of **Approved**, as documented in Report 3.

Report 3: Judgment and Recommendation Report for Scenario 3
Safety tools detection report
Final Judgment: ✓ Approved Asset Availability (C_{asset}): 100.0% (5/5 required items verified) Rule Compliance (C_{rule}): 100.0% (3/3 requirements satisfied)
Total counts Fire Extinguishers: 6 (Required: 2) – SATISFIED Emergency Exits: 2 (Required: 2) – SATISFIED First Aid Kits: 1 (Required: 1) – SATISFIED
Recommendation • Fire extinguisher (Desk area): Accessible on floor (90.0% reachable – Category: Unobstructed) • Fire extinguisher (Open area): Accessible near desk (90.0% reachable – Category: Unobstructed) • Fire extinguisher (Office entrance): Accessible (90.0% reachable – Category: Unobstructed) • First aid kit (Table): Accessible (100.0% reachable – Category: Unobstructed) • Emergency exit (Centre door): Visible sign; minor cables bundled along perimeter (80.0% reachable – Category: Unobstructed) • Emergency exit (Glass doors): Clear path (100.0% reachable – Category: Unobstructed)

Quantitative performance metrics for Scenario 3 are presented in Table 8. At the visual detection level, the MLLM pipeline achieved 100.0% accuracy, precision, recall, and F1-score across all 16 frames.

Table 8. Performance metrics and feasibility evaluation of the MLLM perception system in Scenario 3 ($N = 16$ frames, 48 class-frame evaluation units).

Metric	Value (Raw Count)	95% Wilson Conf. Interval
Final Approval Status (S_{final})	Approved	–
Rule Compliance (C_{rule})	100.0% (3/3)	–
Asset Compliance (C_{asset})	100.0% (5/5)	–
Frame-Class Accuracy	100.0% (48/48)	[92.6%, 100.0%]
Precision	100.0% (16/16)	[80.6%, 100.0%]
Recall	100.0% (16/16)	[80.6%, 100.0%]
F1-score	100.0%	–



Figure 13. Constructed semantic map for Scenario 3, showing annotated safety equipment detections in the fully compliant FabLab environment.

The resulting semantic map is shown in Figure 13. Detections were successfully projected onto the geometric occupancy grid, providing a complete spatially annotated layout of all compliant safety assets.

5.4.4. Scenario 4: Compliance with accessibility warnings

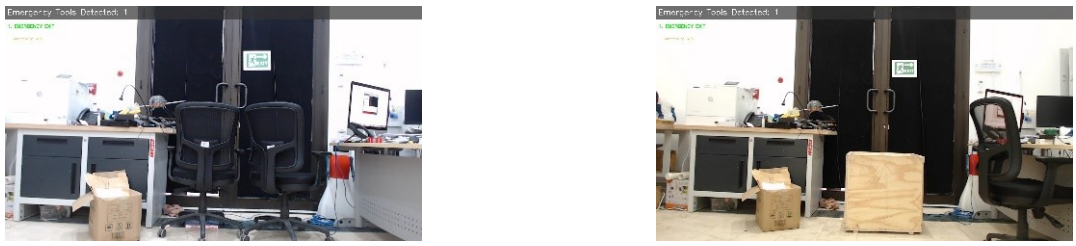
The fourth scenario evaluated the system's ability to verify safety equipment presence while concurrently identifying operational accessibility hazards. In this scenario, all required safety equipment items were present ($T_i = 5$ items: two fire extinguishers, two emergency exits, and one first-aid kit). However, both emergency exits were intentionally obstructed using chairs and boxes to simulate egress path blockages (Figure 14).



(**A**) (**B**)

Figure 14. Original images captured in Scenario 4, showing emergency exits intentionally obstructed by chairs and a box to evaluate accessibility assessment capabilities.

The MLLM perception pipeline accurately identified all safety assets across the captured frames while detecting the spatial obstructions near the exits (Figure 15).



(**A**) (**B**)

Figure 15. Labeled perception outputs generated for Scenario 4, highlighting detected emergency exits and the visual obstructions affecting their accessibility.

Evaluating the scene against the multi-tier framework yielded $C_{\text{asset}} = 100.0\%$ (five of five required items verified) and $C_{\text{rule}} = 100.0\%$ (three of three safety categories fulfilled). However, because emergency exit pathways were visually obstructed by chairs and boxes, the reasoning engine assigned an accessibility score of $A_{\text{exit}} = 60.0\%$, classifying both exits under the **Partially Obstructed** category (Table 3). Because $A_{\text{exit}} < A_{\text{thresh}}(80.0\%)$, the updated safety gate (Equation 5.3) assigned a final status of **Conditionally Approved**, as detailed in Report 4, generating targeted corrective flags for facility administrators to clear the pathways.

Report 4: Judgment and Recommendation Report for Scenario 4
Safety tools detection report
Final Judgment: ✎ Conditionally Approved Asset Availability (C_{asset}): 100.0% (5/5 required items verified) Rule Compliance (C_{rule}): 100.0% (3/3 requirements satisfied)
Total counts Fire Extinguishers: 6 (Required: 2) – SATISFIED Emergency Exits: 2 (Required: 2) – SATISFIED First Aid Kits: 1 (Required: 1) – SATISFIED
Recommendation • Fire extinguisher (Bottom right corner): Accessible on floor (90.0% reachable – Category: Unobstructed) • Emergency exit (Centre doors): Path partially blocked by chairs (60.0% reachable – Category: Partially Obstructed – Action Required: Clear exit path) • Emergency exit (Centre doors): Path obstructed (60.0% reachable – Category: Partially Obstructed – Action Required: Clear exit path) • Fire extinguisher (Right side, near desk): Easily accessible (100.0% reachable – Category: Unobstructed) • First aid kit (Table, left side): Easily accessible (100.0% reachable – Category: Unobstructed) • Fire extinguishers (Glass wall & desk): Easily accessible (100.0% reachable – Category: Unobstructed)

System evaluation metrics for Scenario 4 are summarized in Table 9. The perception pipeline maintained 100.0% precision, recall, F1-score, and accuracy across all 16 frames, demonstrating robust object detection and classification performance even in visually cluttered environments with spatial obstructions.

Table 9. Performance metrics and feasibility evaluation of the MLLM perception system in Scenario 4 ($N = 16$ frames, 48 class-frame evaluation units).

Metric	Value (Raw Count)	95% Wilson Conf. Interval
Final Approval Status (S_{final})	Conditionally Approved	–
Rule Compliance (C_{rule})	100.0% (3/3)	–
Asset Compliance (C_{asset})	100.0% (5/5)	–
Frame-Class Accuracy	100.0% (48/48)	[92.6%, 100.0%]
Precision	100.0% (16/16)	[80.6%, 100.0%]
Recall	100.0% (16/16)	[80.6%, 100.0%]
F1-score	100.0%	–

The constructed semantic map for Scenario 4 is presented in Figure 16. The system successfully integrated sensory and perceptual outputs to generate a topologically structured geometric layout annotated with localized safety tools and physical hazard alerts. While minor localization variations are present due to the MLLM's known 2D perspective limitations (analyzed in Section 6.2), these bounding coordinate variations did not impact overall spatial reasoning, demonstrating the system's effectiveness as a practical tool for automated safety auditing.

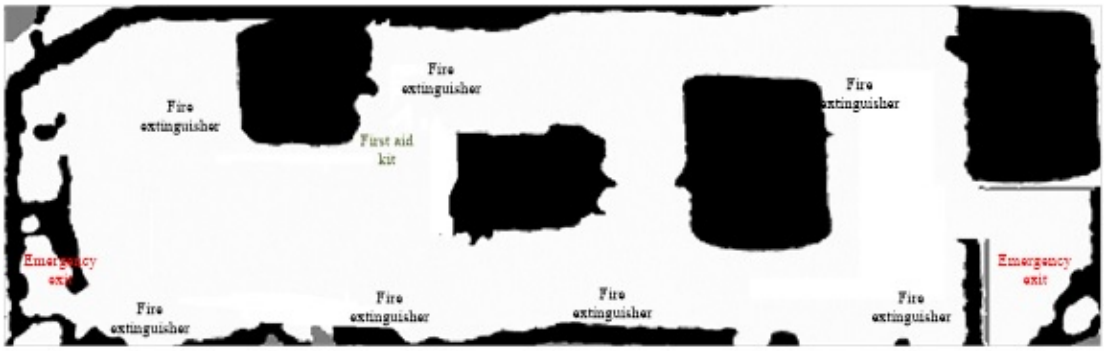


Figure 16. Constructed semantic map for Scenario 4, showing annotated safety equipment and spatial locations in the FabLab environment with obstructed exits.

5.5. Statistical interpretation and evaluation limitations

While the perception pipeline achieved zero visual classification errors across all four evaluation scenarios (100.0% accuracy, precision, and recall), these figures must be interpreted cautiously as initial feasibility evidence under controlled experimental conditions.

Several dataset characteristics should be noted:

1. Sample Size and Traversal Correlation: Each scenario evaluation was conducted over $N = 16$ RGB image frames captured during a single mobile robot traversal along a 45 m test path. Because consecutive frames along a single trajectory exhibit spatial correlation, zero-error performance in

this controlled sample does not guarantee identical performance in scaled industrial environments with dynamic lighting, heavy dust, or unfamiliar object variants. Calculated using the Wilson score interval, a 16/16 success rate yields a 95.0% confidence interval of [80.64%, 100.0%].

2. Prompt and Hyperparameter Independence: To prevent evaluation bias and data leakage, prompt structures and confidence thresholds ($C_{\text{thresh}} = 0.70$) were established prior to testing using a separate set of preliminary setup images captured from an adjacent hallway (excluded from the 64 evaluation frames across Scenarios 1–4).

5.6. Quantitative semantic mapping and spatial localization accuracy

To quantitatively validate the spatial fidelity of the image-to-map projection pipeline (Equation 5.9), the globally projected coordinates of all deduplicated safety assets were compared against ground-truth physical coordinates ($\mathbf{x}_{\text{GT}}$) measured using a Leica laser distance meter.

Detailed per-asset ground-truth coordinates, estimated global map coordinates, positional errors, and associated frame observation IDs for all $M = 20$ mapped assets across all four scenarios are provided in Appendix B (**Table S1**).

Mapping accuracy was evaluated using two primary metrics:

1. **Mean Euclidean Position Error ($\bar{e}_{\text{pos}}$):** Calculated as

$$\bar{e}_{\text{pos}} = \frac{1}{M} \sum_{j=1}^{M} \|\mathbf{c}^{(j)} - \mathbf{x}_{\text{GT}}^{(j)}\|_2, \tag{5.9}$$

   where M is the total number of mapped assets in the scene.

2. **Map-Level Object Association Accuracy ($F1$-Score):** Evaluated by determining whether detected assets were correctly assigned to unique global landmarks without introducing false-positive ghost markers or splitting single physical items. In Scenario 1, because no physical assets were present ($M = 0$), true positives, false positives, and false negatives were all zero (TP = FP = FN = 0), making precision, recall, and association $F1$-score mathematically undefined; hence, they are strictly reported as N/A in Table 10.

Table 10. Quantitative spatial localization error, map-level landmark association performance, and evaluation metrics across experimental scenarios based on mapped physical landmarks (M).

Scenario	Mapped Assets (M)	Mean Position Error ($\bar{e}_{\text{pos}}$)	Max Error ($e_{\max}$)	Association $F1$-Score
Scenario 1	0	N/A	N/A	N/A*
Scenario 2	2	0.21 m	0.28 m	100.0%
Scenario 3	9	0.18 m	0.25 m	100.0%
Scenario 4	9	0.19 m	0.26 m	100.0%
Overall System	**20**	**0.19** ± **0.04** m	**0.28** m	**100.0**%

*Reported as N/A because no target landmarks were present ($TP = FP = FN = 0$).

Across all evaluated scenarios, the framework achieved an average global spatial mapping error of 0.19 ± 0.04 m. Spatial clustering ($\epsilon_{\text{cluster}} = 1.2$ m) successfully prevented detection inflation, yielding

a 100.0% object association F1-score with zero false-positive landmark duplications on the global occupancy map.

6. Discussion

Research on mobile robotic inspection and vision-language integration leveraging multimodal large language models (MLLMs) has expanded rapidly in recent years [10–13, 15, 16, 21, 30–39]. Systems such as LM-Nav [13], Co-NavGPT [12], and OpenNav [21] demonstrate how MLLMs link spoken or written commands with visual-spatial representations, enabling flexible, context-aware semantic mapping and target search. However, while existing research heavily emphasizes natural language instruction following and general semantic target navigation, it largely overlooks automated safety auditing and multi-tier regulatory compliance verification.

The framework presented in this work repurposes MLLM visual reasoning to execute multi-tier civil defense compliance audits within indoor FabLab facilities. Unlike traditional navigation-centric systems, our approach integrates MLLM semantic outputs into a deterministic regulatory evaluation engine (Algorithm 1). Across four distinct operational scenarios, the system demonstrated robust performance by evaluating both asset availability and operational accessibility during fixed-route inspection traversals, generating formal reports detailing four specific compliance states: non-compliant, partially compliant, fully compliant, and conditionally approved (functionally obstructed).

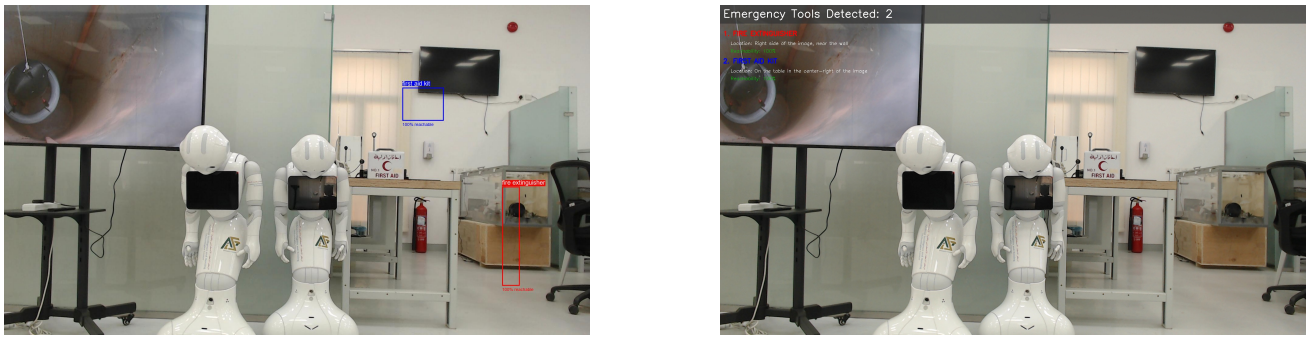
6.1. Multi-tier compliance evaluation vs. perception metrics

A key strength of the proposed architecture is the mathematical decoupling of raw visual perception performance from regulatory rule satisfaction:

1. **Visual Perception Tier:** Evaluated using standard classification and detection metrics (Equations 5.4–5.7), the MLLM pipeline achieved 100.0% precision, recall, and F1-score across non-empty environments (Scenarios 2–4) and 100.0% accuracy in empty scenes (Scenario 1).
2. **Asset Availability Compliance (C_{asset}):** Under Equation (5.1), clipping detected asset counts to mandated inventory thresholds ($\min(N_{\text{detected},k}, N_{\text{req},k})$) prevented over-compliance distortion. In Scenarios 3 and 4, where six fire extinguishers were detected against a mandate of two, clipping capped C_{asset} at 100.0%, avoiding artificially inflated asset scores.
3. **Regulatory Rule Compliance (C_{rule}):** Operating across all required safety categories ($|\mathcal{K}| = 3$), Equation (5.2) enforced strict category coverage. In Scenario 2, despite detecting two physical assets ($C_{\text{asset}} = 40.0\%$), the complete absence of emergency exit detections restricted C_{rule} to 33.3%, correctly triggering a non-approval status.
4. **Deterministic Decision Gating (S_{final}) and Hazard Warnings:** In accordance with Equation (5.3) and Algorithm 1, S_{final} enforces both regulatory rule coverage and physical accessibility gates ($A_{\text{thresh}} = 80.0\%$). Scenario 4 demonstrated the practical utility of this gating: Although the system fulfilled all inventory and rule requirements ($C_{\text{asset}} = 100.0\%, C_{\text{rule}} = 100.0\%$), egress path blockages resulted in a 60.0% reachability score (**Partially Obstructed** category), triggering a hard accessibility constraint failure ($A_{\text{exit}} < 80.0\%$). This yielded a final decision of **Conditionally Approved** and promoted actionable corrective flags without misclassifying inventory availability.

6.2. Spatial grounding, prompt strictness, and decoupled architecture

Deploying MLLM-based perception in physical environments presents ongoing challenges, particularly regarding spatial grounding. MLLMs often struggle to maintain exact spatial coherence, frequently generating hallucinated bounding box coordinates or inaccurate depth estimates when tasked with numerical pixel-level localization (Figure 17). This stems from the model's reliance on 2D perspective cues rather than native stereo-depth sensing.



(**A**) (**B**)

Figure 17. Illustration of the MLLM's role as a semantic classifier rather than a primary localizer: (**A**) visual localization inaccuracy due to 2D perspective bounding cues; (**B**) successful semantic identification of emergency equipment despite inaccurate spatial bounding.

To resolve this limitation, our framework treats the MLLM strictly as a high-level semantic classifier rather than a primary localizer. Positional grounding is offloaded to the BreezySLAM 2D occupancy map and the mobile robot's wheel-odometry pose estimates, projecting semantic labels directly onto verified global coordinates.

Furthermore, system prompt strictness introduces clear perceptual trade-offs. The prompt used during MLLM inference enforces visual verification criteria such as verifying the red cylindrical body, pressure gauge, and mounting bracket before confirming a fire extinguisher detection to minimize false positives. During visual parsing, the prompt protocol explicitly directs the MLLM to differentiate between the emergency exit sign (visual marker), the physical emergency exit door, and the floor space comprising the unobstructed exit path. While strict prompt guidelines prevent hallucinated compliance approvals, they can impact recall under partial occlusions. In safety-critical auditing, prioritizing high precision over aggressive recall is advantageous: Flagging partially occluded or ambiguous assets for manual review is far safer than misclassifying obstructed equipment as fully compliant.

Heuristic Visual Accessibility vs. 3D Geometric Measurement: A key consideration in visual compliance monitoring is distinguishing qualitative visual assessment from calibrated 3D spatial measurement. In the present implementation, accessibility scores (A_i) represent semantic heuristic estimates generated from 2D RGB frames using prompt criteria (Table 3). While effective for preliminary screening and identifying obvious pathway blockages, these values are not 3D spatial metrics. Integrating 3D LiDAR point-cloud clearance calculations and conducting empirical validation against assessments by certified human safety inspectors are key objectives for future iterations of the platform.

6.3. Comparison with contemporary semantic mapping systems

Systems such as O3D-SIM, osmAG-LLM, and OpenNav emphasize open-vocabulary perception pipelines and rich topometric scene graphs [14, 20–22, 26, 40, 41]. Conversely, systems like AutoInspect focus on revisit-based map maintenance and long-term operational persistence. The present work targets single-run, safety-compliance visual screening along defined inspection trajectories.

Offloading MLLM visual reasoning to OpenAI's `GPT-4o-mini` cloud API introduces infrastructure dependencies. However, the system's asynchronous, non-blocking architecture completely decouples high-level visual auditing from lower-level robot motion execution. The robot executes local 2D LiDAR SLAM on edge hardware (Raspberry Pi 4) in real time, capturing RGB frames at discrete waypoints along its designated inspection route. Cloud latencies do not compromise operational movement or safety. Furthermore, spatial redundancy from capturing multiple overlapping viewpoints of safety assets during trajectory traversal mitigates cloud-based non-determinism through cross-verification across spatial coordinates.

6.4. Data governance, cybersecurity, and scope boundaries

Deploying cloud-connected vision-language models in physical facilities introduces operational data governance, confidentiality, and cybersecurity considerations:

1. **Privacy-Preserving Edge Redaction & Data Minimization:** Before cloud transmission, captured RGB frames undergo automated onboard edge pre-processing using OpenCV to blur human faces and mask computer screen displays, eliminating personally identifiable information (PII). Only downsampled visual frames (768×768 px) are transmitted; raw LiDAR point clouds, metric SLAM coordinates, and facility layout maps remain strictly isolated on local edge hardware.
2. **Transport Security & Zero Data Retention:** API requests are authenticated using hardware-encrypted environment keys over TLS 1.3 HTTPS connections. Operating under OpenAI API Data Privacy Terms, submitted data is not utilized for model training or fine-tuning and is processed via Zero Data Retention (ZDR) endpoints that purge payloads immediately post-inference.
3. **Cybersecurity & Prompt Injection Defense:** To guard against visual prompt injection (e.g., adversarial poster text or spoofed safety labels placed in the environment), the architecture isolates MLLM output parsing within Pydantic JSON schemas and delegates decision-making to a deterministic, non-compensatory rule engine (Algorithm 1). Adversarial text within images cannot alter the execution logic or override physical safety gates.
4. **Air-Gapped Offline Fallback Pathway:** For high-security or air-gapped industrial facilities where external cloud connections are strictly prohibited, the architecture supports an offline fallback pipeline. This mode replaces cloud API calls with quantized open-source vision-language models (e.g., LLaVA-NeXT, MobileVLM, or Qwen2-VL) running locally on onboard edge accelerators (such as the NVIDIA Jetson Orin Nano / AGX Orin).
5. **Experimental Scope Boundaries:** While the current framework validates compliance auditing across four distinct scenarios within a FabLab testbed ($N = 64$ frames), evaluating multi-facility environments, dynamic human crowds, variable lighting regimes, and diverse equipment designs falls outside the scope of this study and is a key objective for future research.

7. Conclusion & Future work

In this paper, we presented RoboAudit, an automated mobile robotic framework that integrates 2D LiDAR-based SLAM with cloud-based multimodal large language models (specifically `GPT-4o-mini`) for indoor safety auditing in FabLab environments. Operating on an asynchronous, decoupled architecture, the system separates real-time geometric localization and mapping from cloud-based visual-semantic reasoning and multi-tier regulatory compliance evaluation.

Evaluated across four real-world experimental scenarios, the framework demonstrated 100.0% visual perception accuracy and robust compliance reasoning:

- **Scenario 1 (Total Non-Compliance):** Accurately verified total asset absence ($C_{\text{asset}} = 0.0\%, C_{\text{rule}} = 0.0\% \rightarrow$ **Not Approved**).
- **Scenario 2 (Partial Asset Compliance):** Successfully detected existing assets while flagging unfulfilled category mandates ($C_{\text{asset}} = 40.0\%, C_{\text{rule}} = 33.3\% \rightarrow$ **Not Approved**).
- **Scenario 3 (Full Compliance and Accessibility):** Confirmed complete equipment presence and unobstructed physical access ($C_{\text{asset}} = 100.0\%, C_{\text{rule}} = 100.0\% \rightarrow$ **Approved**).
- **Scenario 4 (Compliance with Accessibility Warnings):** Granted regulatory inventory approval ($C_{\text{asset}} = 100.0\%, C_{\text{rule}} = 100.0\%$) while automatically gating the decision to **Conditionally Approved** due to egress path blockages (60.0% reachability) to prompt targeted facility maintenance.

It is important to emphasize that RoboAudit is designed as an automated decision-support and preliminary visual screening tool intended to assist facility managers and safety auditors, rather than replace certified human inspectors who issue legal compliance certifications.

Future work will focus on extending the system's operational capabilities beyond visual presence verification. Key extensions include integrating active visual zooming and specialized fine-grained perception modules to read fire extinguisher pressure gauges, verify expiration dates on medical supplies, test emergency alarm functionality, and quantitatively evaluate 3D clearance along evacuation corridors. Additionally, we plan to validate the system across multi-facility industrial environments, deploy quantized edge-hosted MLLMs for offline air-gapped operations, and incorporate mechanisms for persistent topometric map updating across multi-run inspection missions.

Author Contributions

The authors confirm contribution to the paper as follows: Conceptualization, T. Alhmiedat, O. M. Alia, and A. Alhakamy; methodology, O. M. Alia and F. Alnajjar; software, O. M. Alia, Z. Alenzi, T. Alhmiedat, and A. M. Marei; validation, O. M. Alia, S. Albelwi, and M. Shanmuganathan; formal analysis, O. M. Alia and T. Alhmiedat; writing, O. M. Alia and A. Alhakamy; original draft preparation, T. Alhmiedat and O. M. Alia. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This research was funded by the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, under grant number 0245-1443-S.

Conflicts of Interest

The authors declare no conflicts of interest.

Use of Generative-AI tools declaration

The authors acknowledge the use of ChatGPT (OpenAI) and Gemini for their assistance in rephrasing several sentences. All scientific content, interpretations, and conclusions remain the sole responsibility of the authors.

References

1. F. Rubio, F. Valero, C. Llopis-Albert, A review of mobile robots: concepts, methods, theoretical framework, and applications, *Int. J. Adv. Robot. Syst.*, **16** (2019), 1729881419839596. https://doi.org/10.1177/1729881419839596
2. Y. L. Zhao, Y. T. Hong, H. P. Huang, Comprehensive performance evaluation between visual SLAM and LiDAR SLAM for mobile robots: theories and experiments, *Appl. Sci.*, **14** (2024), 3945. https://doi.org/10.3390/app14093945
3. R. Alqobali, M. Alshmrani, R. Alnasser, A. Rashidi, T. Alhmiedat and O. M. Alia, A survey on robot semantic navigation systems for indoor environments, *Appl. Sci.*, **14** (2024), 286. https://doi.org/10.3390/app14010089
4. R. Alqobali, R. Alnasser, A. Rashidi, M. Alshmrani, T. Alhmiedat, A real-time semantic map production system for indoor robot navigation, *Sensors*, **24** (2024), 6691. https://doi.org/10.3390/s24206691
5. T. Alhmiedat, O. M. D. Alia, Utilizing a deep neural network for robot semantic classification in indoor environments, *Sci. Rep.*, **15** (2025), 21937. https://doi.org/10.1038/s41598-025-07921-7
6. A. Alotaibi, H. Alatawi, A. Binnouh, L. Duwayriat, T. Alhmiedat, O. M. Alia, Deep learning-based vision systems for robot semantic navigation: an experimental study, *Technologies*, **12** (2024), 157. https://doi.org/10.3390/technologies12090157
7. N. Yokoyama, S. Ha, D. Batra, J. Wang, B. Bucher, VLFM: Vision-language frontier maps for zero-shot semantic navigation, *Proc. IEEE Int. Conf. Robot. Autom.*, 2024, 42–48. https://doi.org/10.1109/ICRA57147.2024.10610712
8. Z. Alenzi, E. Alenzi, M. Alqasir, M. Alruwaili, T. Alhmiedat, O. M. Alia, A semantic classification approach for indoor robot navigation, *Electronics*, **11** (2022), 2063. https://doi.org/10.3390/electronics11132063

9. Y. Kim, D. Kim, J. Choi, J. Park, N. Oh, D. Park, A survey on integration of large language models with intelligent robots, *Intell. Serv. Robot.*, **17** (2024), 1091–1107. https://doi.org/10.1007/s11370-024-00550-5
10. H. Pan, S. Huang, J. Yang, J. Mi, K. Li, X. You, et al., Recent advances in robot navigation via large language models: a review, https://doi.org/10.13140/RG.2.2.35284.41603
11. D. Shah, M. R. Equi, B. Osiński, F. Xia, B. Ichter, S. Levine, Navigation with large language models: semantic guesswork as a heuristic for planning, *Proc. Mach. Learn. Res.*, **229** (2023), 2683–2699. Available from: https://proceedings.mlr.press/v229/shah23c.html.
12. B. Yu, Q. Yuan, K. Li, H. Kasaei, M. Cao, Co-NavGPT: multirobot cooperative visual semantic navigation using vision language models, *IEEE Robot. Autom. Lett.*, **11** (2026), 2122–2129. https://doi.org/10.1109/LRA.2025.3645650
13. D. Shah, B. Osiński, B. Ichter, S. Levine, LM-Nav: robotic navigation with large pre-trained models of language, vision, and action, *Proc. Mach. Learn. Res.*, **205** (2023), 492–504. Available from: https://proceedings.mlr.press/v205/shah23b.html.
14. F. Xie, S. Schwertfeger, H. Blum, osmAG-LLM: zero-shot open-vocabulary object navigation via semantic maps and large language models reasoning, *arXiv preprint arXiv:2507.12753*, https://doi.org/10.48550/arXiv.2507.12753
15. X. Pan, J. Perando, Enhancing robotic navigation with large language models, *Electron. Imaging*, **37** (2025), 105. https://doi.org/10.2352/EI.2025.37.15.AVM-105
16. T. Wang, J. Fan, P. Zheng, An LLM-based vision and language cobot navigation approach for human-centric smart manufacturing, *J. Manuf. Syst.*, **75** (2024), 299–305. https://doi.org/10.1016/j.jmsy.2024.04.020
17. X. Yu, H. Zhang, Y. Duan, X. Wang, An optimized convolutional neural network based on multi-strategy grey wolf optimizer to identify crop diseases and pests, *Displays*, **92** (2026), 103341. https://doi.org/10.1016/j.displa.2026.103341
18. S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, et al., Grounding DINO: marrying DINO with grounded pre-training for open-set object detection, *arXiv preprint arXiv:2303.05499*, https://doi.org/10.48550/arXiv.2303.05499
19. C. Zhang, D. Han, Y. Qiao, J. U. Kim, S. H. Bae, S. Lee, et al., Faster segment anything: towards lightweight SAM for mobile applications, *arXiv preprint arXiv:2306.14289*, https://doi.org/10.48550/arXiv.2306.14289
20. L. Nanwani, K. Gupta, A. Mathur, S. Agrawal, A. H. Abdul Hafez, K. M. Krishna, Open-set 3D semantic instance maps for vision language navigation – O3D-SIM, *Adv. Robot.*, **38** (2024), 1378–1391. https://doi.org/10.1080/01691864.2024.2395926
21. M. Yuan, L. Wang, S. L. Waslander, OpenNav: open-world navigation with multimodal large language models, *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2025. https://doi.org/10.1109/IROS60139.2025.11247593
22. M. Staniaszek, T. Flatscher, J. Rowell, H. Niu, W. Liu, Y. You, et al., Autoinspect: towards long-term autonomous industrial inspection, *IEEE Trans. Field Robot.*, **2** (2025), 529–548. https://doi.org/10.1109/TFR.2025.3586831

23. Y. Wang, H. Luo, W. Fang, An integrated approach for automatic safety inspection in construction: domain knowledge with multimodal large language model, *Adv. Eng. Inform.*, **65** (2025), 103246. https://doi.org/10.1016/j.aei.2025.103246

24. R. Hussain, D. Lee, M. S. Abbas, S. F. A. Zaidi, A. Pedro, C. Park, Vision-language model-based intelligent assistant for onsite construction safety inspection, *Autom. Constr.*, **182** (2026), 106728. https://doi.org/10.1016/j.autcon.2025.106728

25. D. Hu, V. J. Gan, Semantic navigation for automated robotic inspection and indoor environment quality monitoring, *Autom. Constr.*, **170** (2025), 105949. https://doi.org/10.1016/j.autcon.2024.105949

26. S. H. Allu, I. Kadosh, T. Summers, Y. Xiang, A modular robotic system for autonomous exploration and semantic updating in large-scale indoor environments, *arXiv preprint arXiv:2409.15493*, https://doi.org/10.48550/arXiv.2409.15493

27. General Directorate of Civil Defense, *Civil defense regulations*, 2025. Available from: https://998.gov.sa/English/CivilDefenseLists/Pages/civil-defense-lists.aspx.

28. Department of Civil Engineering, Fahad Bin Sultan University, *Laboratory safety manual*, 2025. Available from: https://fbsu.edu.sa/PDF/Laboratory-Safety-Manual.pdf.

29. King Saud University, College of Science, *Safety guide in the laboratories*, 2025.

30. F. Xie, J. Zhang, S. Schwertfeger, Intelligent LiDAR navigation: leveraging external information and semantic maps with LLM as copilot, *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2025. https://doi.org/10.1109/IROS60139.2025.11246069

31. A. Zhang, J. Ji, Research on navigation methods based on LLMs, *arXiv preprint arXiv:2504.15600*, https://doi.org/10.48550/arXiv.2504.15600

32. Z. Zhong, Y. He, P. Li, F. Yu, F. Ma, A language-driven navigation strategy integrating semantic maps and large language models, *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024. https://doi.org/10.1109/IROS58592.2024.10802631

33. F. Lotfi, F. Faraji, N. Kakodkar, T. Manderson, D. Meger, G. Dudek, Constrained robotic navigation on preferred terrains using LLMs and speech instruction: exploiting the power of adverbs, *Experimental Robotics (ISER 2023)*, **30** (2024), 118–128. https://doi.org/10.1007/978-3-031-63596-0_11.

34. E. Latif, 3P-LLM: probabilistic path planning using large language model for autonomous robot navigation, *arXiv preprint arXiv:2403.18778*, https://doi.org/10.48550/arXiv.2403.18778

35. Z. Chen, Z. Su, C. Chesser, C. Bourlon, W. Sheng, Intelligent indoor navigation for home robots based on large language models, *Proc. IEEE Int. Conf. Autom. Sci. Eng.*, 2025. https://doi.org/10.1109/CASE58245.2025.11163777

36. A. N. Kravtsov, B. S. Goryachkin, Integration of large language models for autonomous navigation of a mobile robot, *Proc. Int. Youth Conf. Radio Electron., Electr. Power Eng. (REEPE)*, 2025. https://doi.org/10.1109/REEPE63962.2025.10971154

37. T. Wang, J. Fan, P. Zheng, R. Yan, L. Wang, Vision-language model-based human-guided mobile robot navigation in an unstructured environment for human-centric smart manufacturing, *Engineering*, **62** (2026), 229–239. https://doi.org/10.1016/j.eng.2025.04.028

38. Y. Qiao, W. Lyu, H. Wang, Z. Wang, Z. Li, Y. Zhang, et al., Open-Nav: exploring zero-shot vision-and-language navigation in continuous environment with open-source LLMs, *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025. https://doi.org/10.1109/ICRA55743.2025.11127584

39. P. Doma, A. Arab, X. Xiao, LLM-enhanced path planning: safe and efficient autonomous navigation with instructional inputs, *arXiv preprint arXiv:2412.02655*, https://doi.org/10.48550/arXiv.2412.02655

40. N. Dengler, T. Zaenker, F. Verdoja, M. Bennewitz, Online object-oriented semantic mapping and map updating, *Proc. Eur. Conf. Mob. Robots (ECMR)*, 2021. https://doi.org/10.1109/ECMR50962.2021.9568817

41. M. Narayana, A. Kolling, L. Nardelli, P. Fong, Lifelong update of semantic maps in dynamic environments, *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020. https://doi.org/10.1109/IROS45743.2020.9341516



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)