ALARMbot: Autonomous Laboratory Safety Inspection and Operable Hazard Intervention Robot Enabled by Foundation Models

Shoujie Li, Yushan Liu, Xintao Chao, Yan Huang, Xiao-Ping Zhang, *Fellow IEEE*, Wenbo Ding

***Abstract*—Ensuring laboratory safety is a critical challenge due to the presence of hazardous materials and complex operational environments. In this paper, we develop an autonomous laboratory inspection robot (ALARMbot) leveraging foundation models for intelligent safety management. The system integrates a mobile platform, multi-modal sensors, and a 6-DoF manipulator. By fusing LiDAR-based mapping with vision-and-language models, the robot achieves semantic navigation and fine-grained functional zone recognition. A hierarchical framework combines YOLOv8-OBB visual perception (achieving 93.1% mAP on custom datasets) with vision-language risk reasoning for real-time hazard detection. The robot autonomously intervenes in operable risks with an average response time of 8.5 seconds and navigates complex laboratory layouts with a 96.3% success rate. Extensive real-world experiments demonstrate reliable navigation, accurate risk detection, and effective hazard mitigation. This work offers a practical solution for intelligent laboratory safety and highlights the advantages of foundation models in autonomous inspection robotics.**

***Index Terms*—Laboratory inspection, Navigation, Foundation models**

I. Introduction

LABORATORY safety is a critical concern in research and industrial environments, where hazardous chemicals, complex instruments, and dynamic activities pose significant risks to both personnel and facilities[1]. Traditional safety inspections, typically performed by human operators, are labor-intensive, error-prone, and often fail to provide timely responses in rapidly changing or high-risk scenarios[2].

Recent advances in robotics and artificial intelligence have enabled the development of autonomous inspection systems. In particular, vision-based detection technologies have become increasingly prominent due to their non-contact, real-time monitoring capabilities [3], [4]. Deep learning-based object detection and semantic analysis methods have demonstrated strong adaptability and accuracy in identifying laboratory hazards, including misplaced equipment, chemical spills, and unsafe actions [5], [6]. Robotic platforms designed for inspection tasks now incorporate diverse mobility systems and multi-modal sensors, enabling them to navigate complex laboratory layouts and interact with various objects [7], [8]. The integration of simultaneous localization and mapping (SLAM) algorithms and semantic mapping strategies further enhances their ability to understand and operate within intricate environments.

Despite these advancements, several challenges remain. Existing systems often lack robust multi-task collaboration, sufficient reliability in cluttered or extreme environments, and effective human-robot interaction mechanisms. Moreover, precise risk detection and autonomous mitigation—such as upright positioning of tipped bottles or cleaning chemical spills—require advanced perception and manipulation capabilities that are not fully addressed by current approaches.

To address this, we propose an integrated robotic system for laboratory safety inspection, combining a mobile platform, multi-modal perception sensors, and a flexible manipulation subsystem. The main contributions of this work are as follows:

- We develop a semantic-aware navigation framework that fuses LiDAR-based mapping with vision-and-language models, enabling natural language instruction following and fine-grained semantic mapping in laboratory environments.
- We design a hierarchical inspection and risk mitigation framework that leverages YOLO-based visual perception and vision-language reasoning to detect and autonomously address laboratory hazards in real time.
- We introduce a modular manipulation strategy, allowing the robot to autonomously execute risk mitigation actions—such as upright positioning of fallen bottles, cleaning spills, and sorting objects—based on structured risk assessments.
- We implement and validate a web-based human–robot interaction interface, supporting intuitive task assignment, real-time monitoring, and traceable reporting, and demonstrate the system's effectiveness through extensive real-world experiments.

The remainder of this paper is organized as follows. Section II reviews the related work in vision-based safety detection and robotic inspection systems. Section III introduces the system design. Section IV describes the navigation and semantic mapping strategy. Section V presents the inspection and manipulation framework. Section VI reports experimental results and Section VII concludes the paper.

II. Related Work

A. Traditional Methods for Laboratory Safety Inspection

Traditional laboratory safety inspection methods primarily rely on manual checks or classical computer vision algorithms.

This work was supported by National Key R&D Program of China grant (2024YFB3816000), Shenzhen Key Laboratory of Ubiquitous Data Enabling (No. ZDSYS20220527171406015), Shenzhen Science and Technology Program (JCYJ20220530143013030), Guangdong Innovative and Entrepreneurial Research Team Program (2021ZT09L197), Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (No. SZPR2023005) and Meituan. *Shoujie Li, Yushan Liu and Xintao Chao contributed equally to this work.* (Corresponding author: Wenbo Ding, ding.wenbo@sz.tsinghua.edu.cn)

Shoujie Li, Yushan Liu, Xintao Chao, Yan Huang, Xiao-Ping Zhang and Wenbo Ding are with Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China.

The video and supplementary materials can be found at alarm-bot.github.io.

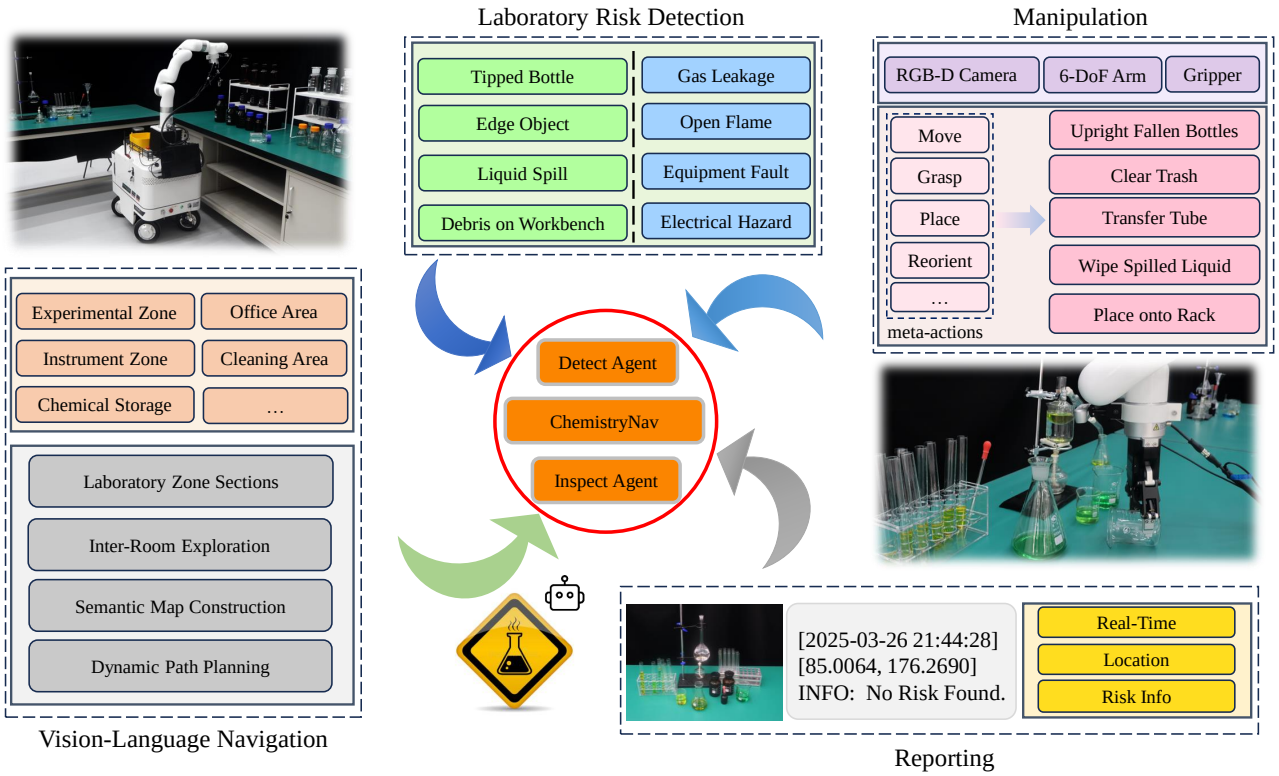

Fig. 1. **System architecture of the autonomous laboratory inspection framework.** The system consists of three core modules: **ChemistryNav** for navigation and semantic mapping, **Detect Agent** for risk detection and reasoning, and **Inspect Agent** for manipulation and user interaction. It enables autonomous exploration and partitioning of unknown laboratory environments, performs real-time risk identification and corresponding mitigation actions along the inspection route, and uploads structured results to the Web UI for visualization and logging.

Manual inspections are labor-intensive, time-consuming, and prone to human error, especially in dynamic or hazardous environments. Early automated systems adopted rule-based image processing or classical machine learning techniques, such as edge detection, color segmentation, and handcrafted feature extraction, to identify laboratory equipment or hazards [9]. While these approaches offered some automation, their performance was limited by poor generalization to diverse laboratory conditions, sensitivity to lighting changes, and inability to handle complex semantic tasks such as compliance verification or context-aware risk assessment.

B. Vision-Based Safety Detection and VLMs

Recent advances in deep learning, especially vision-based detection technologies, have significantly improved the capabilities of automated safety inspection systems in complex environments such as chemical laboratories. Deep learning-based object detection and semantic analysis methods have demonstrated strong adaptability and accuracy in identifying laboratory hazards, including misplaced equipment, chemical spills, and unsafe actions [3], [5], [4], [6]. Some studies further integrate object detection with OCR for comprehensive scene understanding [4], while others use vision-language models (VLMs) for multi-stage semantic reasoning and compliance verification [6].

In laboratory and industrial safety, these methods are increasingly applied to tasks such as detecting misplaced equipment, identifying chemical spills, and ensuring the correct use of personal protective equipment (PPE) [3], [5], [10]. By leveraging large-scale datasets and multi-modal learning, these models achieve robust performance under various lighting, clutter, and occlusion conditions, which are common in real-world laboratory settings.

C. Robotic Inspection Systems in Laboratory Environments

Autonomous robots for laboratory inspection typically combine mobile platforms, multi-modal perception, and manipulation capabilities. State-of-the-art systems utilize wheeled or hybrid mobile bases equipped with LiDAR and RGB-D cameras for navigation and mapping [7], [8], [11], [12]. The integration of simultaneous localization and mapping (SLAM) with semantic scene understanding enables precise localization and functional zone annotation, essential for targeted inspection and risk mitigation [8], [13].

Recent research also focuses on the role of modular manipulation subsystems, such as 6-DoF robotic arms with adaptive grippers, which extend the robot's ability to intervene in hazardous situations, including upright positioning of fallen bottles and cleaning chemical spills [8]. Multi-robot collaboration and specialized platforms for extreme environments have also been explored [14], [15], [13].

However, existing robotic inspection systems still face several critical limitations:

- **Lack of Closed-Loop Risk Mitigation:** Most current systems focus solely on perception and detection, without the ability to autonomously execute intervention or mitigation actions. When hazards are detected, the system typically only issues alerts or records data, requiring human intervention for actual risk handling.
- **Separation of Perception and Manipulation:** There is a clear disconnect between the perception modules (for hazard detection) and the manipulation modules (for physical intervention). As a result, robots are unable to seamlessly translate high-level semantic understanding or risk assessment into effective, context-aware actions.
- **Limited Autonomous Operation:** Many systems operate in a stepwise or semi-automatic manner, lacking the ability to perform end-to-end, fully autonomous inspection and risk mitigation in complex and dynamic laboratory environments.
- **Insufficient Semantic Reasoning for Manipulation:** Even with advanced perception, existing systems rarely leverage vision-language models to inform manipulation strategies. This limits their ability to handle nuanced, context-dependent tasks such as compliance verification or adaptive risk response.

The integration of large foundation models and multi-modal learning offers promising solutions for semantic perception, flexible reasoning, and adaptive task planning in autonomous laboratory inspection. However, the realization of a truly closed-loop system, which is capable of both detecting and autonomously mitigating risks, remains an open challenge.

III. Hardware Design

As shown in Fig. 2, we design a robotic system tailored specifically for inspection tasks in chemical laboratories. The hardware consists of three modules: a mobile chassis, the perception sensors, and a manipulation subsystem.

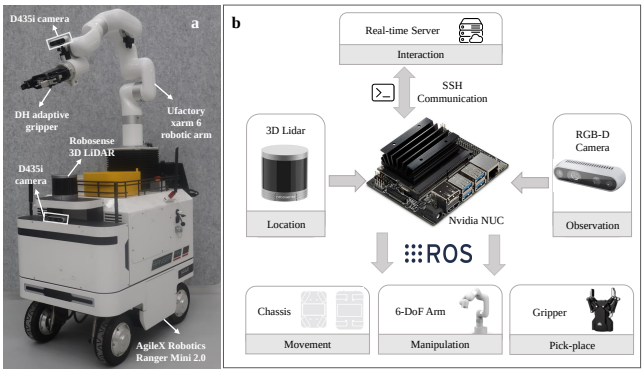

Fig. 2. **Overview of the robotic hardware platform and system architecture.** (a) Mobile manipulator platform used for laboratory inspection, equipped with a 6-DoF robotic arm, RGB-D cameras, LiDAR, and an adaptive gripper. (b) System architecture diagram illustrating the integration of perception, planning, and control components via ROS on a central computing unit, with remote monitoring enabled by SSH-based communication.

The mobile chassis utilizes the Ranger Mini 2.0 from AgileX Robotics, featuring a compact design capable of differential drive, omnidirectional, and translational movements, making it well-suited for navigating in narrow and complex laboratory environments. It is equipped with Ubuntu 18.04 and ROS Melodic, facilitating seamless integration and algorithm deployment. Environmental perception is provided by an Intel RealSense D435i RGB-D camera and a RoboSense Helios 16-channel 3D LiDAR. The RGB-D camera captures synchronized color and depth data, supplying rich semantic and spatial information for navigation, while the LiDAR accurately maps the 3D environment structure, significantly enhancing navigation performance and safety. Both sensors utilize ROS-compatible drivers, ensuring real-time data fusion and precise sensor calibration, essential for accurate robotic manipulation during movement.The manipulation subsystem consists of a UFactory 6-DoF robotic arm paired with a DH Robotics AG-160-95 adaptive gripper. The robotic arm provides flexible 6-DoF operation and ROS-based motion planning, while the adaptive gripper reliably grasps objects of varying shapes and sizes. An additional Intel RealSense D435i RGB-D camera integrated onto the arm supports target recognition, localization, and precise grasping, enabling robust task execution in complex laboratory scenarios.

IV. Navigation

Effective navigation is essential for autonomous laboratory inspection, requiring both accurate spatial localization and a deep understanding of complex environments. To address this, we developed a navigation system that integrates high-precision LiDAR-based mapping with vision-language models, enabling the robot to follow natural language instructions while constructing detailed semantic maps. This section describes our navigation strategy, including navigation architecture, full-floor exploration, laboratory identification, and functional lab zone annotation.

A. Semantic-Mapping Navigation Architecture

To address the specific challenges in laboratory environments, we reference the latest advancements in Vision-and-Language Navigation (VLN) and propose a semantic-driven, high-precision navigation framework that enables the robot to perform environment exploration and high-precision mapping based on instructions.

As shown in Fig. 3, the input to our VLN framework includes RGB image frames from visual sensors and instructions provided by humans. Given a specific navigation instruction, the Query Generator (frozen during training) formulates targeted queries. Simultaneously, the Vision Encoder processes the input RGB image frames to extract rich visual feature representations. The generated queries and visual features are then projected through dedicated Modality Fusion Projectors, producing instruction-aligned query tokens and image tokens. These tokens are subsequently integrated within our unified Observation Encoder. At each timestep, historical and current observations (denoted by **[HIS]** and **[OBS]**) are encoded alongside navigation-oriented tokens (**[NAV]**)[16]. These combined tokens are then processed by an end-to-end VLM, named ChemistryNav, which generates explicit navigation actions based on natural language instructions.

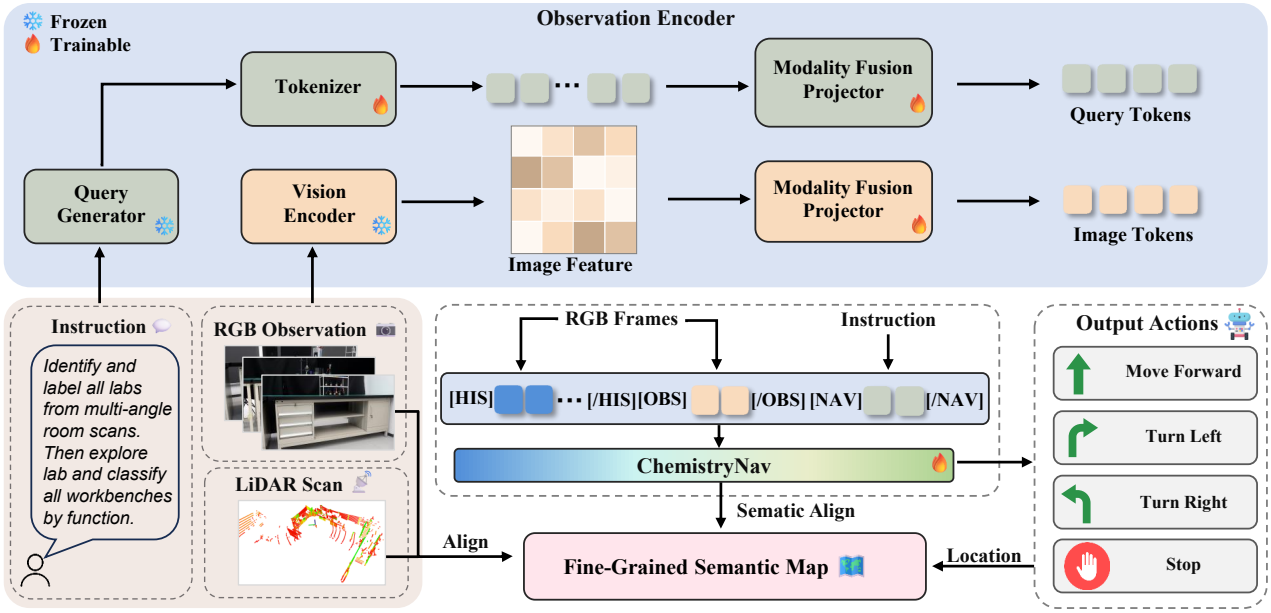

Fig. 3. **The overview of ChemistryNav.** Based on navigation instructions and multimodal sensor inputs—including RGB images and LiDAR scans—the system generates query tokens and image tokens through dedicated Modality Fusion Projectors, integrating them within a unified Observation Encoder. At each timestep, ChemistryNav processes both historical and current observations alongside navigation-oriented tokens to produce explicit navigation actions. Real-time robot positions, acquired via precisely calibrated LiDAR sensors, are dynamically aligned with an evolving Fine-Grained Semantic Map, providing comprehensive semantic and spatial information essential for effective dynamic route planning and targeted risk management tasks.

During robot movement driven by output action results, real-time position data from precisely calibrated LiDAR sensors is continuously recorded and dynamically aligned with the evolving semantic map. This map encodes key environmental regions with semantic labels, supporting spatial reasoning and memory essential for navigation decision-making.

Upon completion of the exploration tasks, the resulting map not only provides high spatial accuracy but also contains rich semantic information. This detailed semantic map thus provides the necessary spatial coordinates and contextual knowledge to support subsequent dynamic route planning and targeted risk management operations.

B. Full-Floor Exploration and Laboratory Identification

Based on VLN, we designed a strategy for exploring the laboratory floor to identify specific rooms. This approach balances extensive spatial exploration with semantic understanding, ensuring comprehensive mapping and accurate room identification, as illustrated in Fig. 4.

The process begins with instructing the agent to explore the entire floor. The prompt guides the agent to identify all rooms with open doors and mark their locations: "*You are on a floor with multiple rooms, each with an open door. Find all laboratory rooms and mark their locations. Your goal is to identify and locate all laboratory rooms based on these images.*" This ensures thorough exploration and relevant image capture for room identification.

During exploration, the robot uses a front-mounted camera and a VLM, named Inspect Agent, to evaluate if each room has key laboratory features. The VLM assesses predefined features such as laboratory equipment, experimental setups,

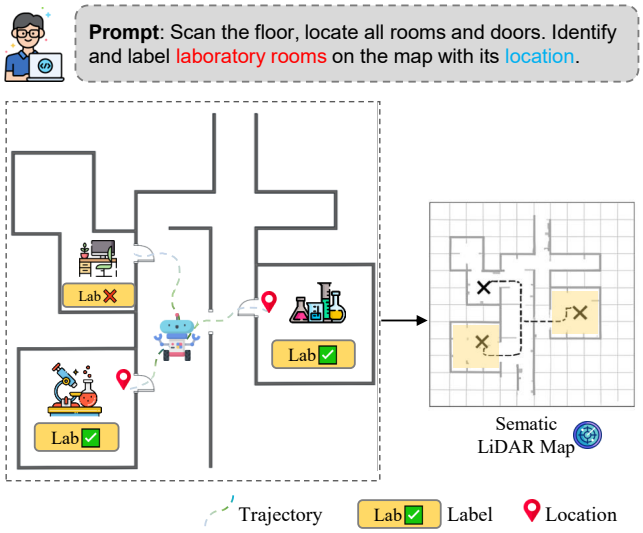

Fig. 4. **Full-floor exploration schematic.** User prompts trigger the robot to visit each room, record whether it is a laboratory, tag its location, and store the data for designing a comprehensive cross-room inspection route.

and chemical storage. Rooms scoring higher than 0.8 are identified as laboratories. The scoring system includes:

- **Laboratory Equipment** (40%): Presence of laboratory equipment like reagent bottles, analytical instruments, etc.
- **Laboratory Facilities** (30%): Workbenches or experimental tables, organized for experimental purposes.
- **Chemical Materials** (20%): Visible chemical containers or storage cabinets.
- **Environmental Cleanliness** (10%): Clean and organized room suggesting a controlled research environment.

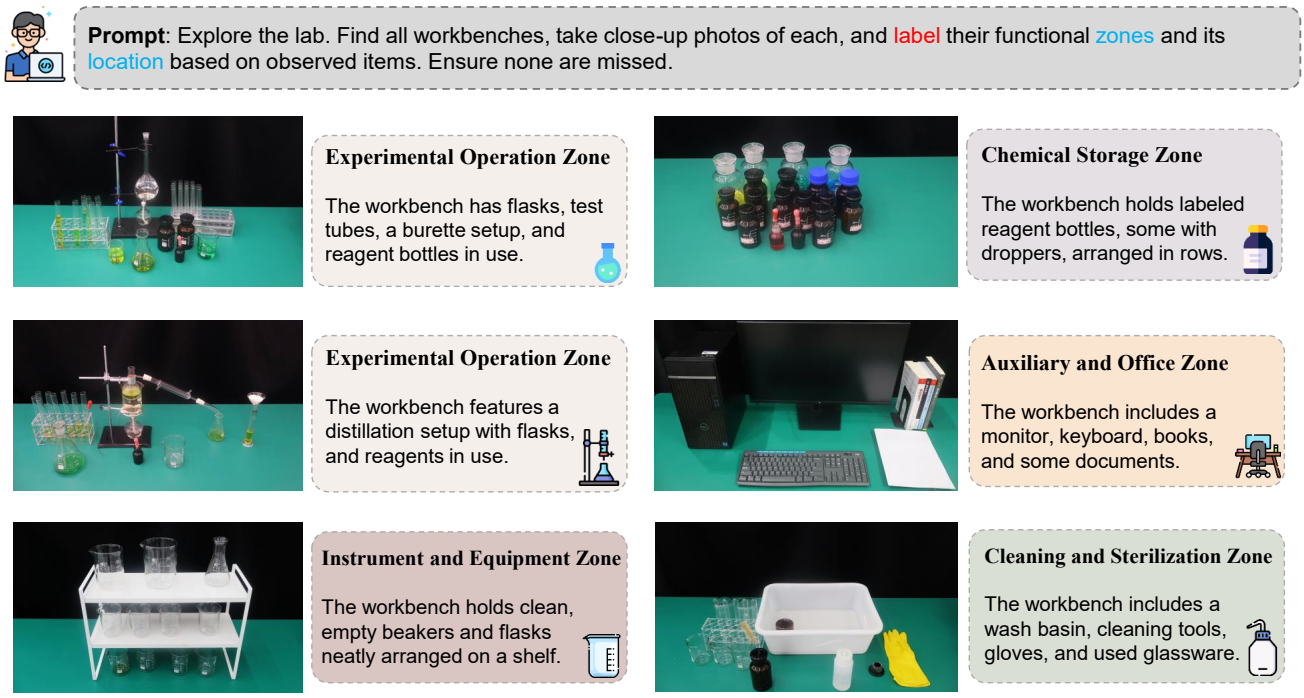

Fig. 5. **Workbench recognition and functional zone classification.** Once the robot reaches an appropriate position near a workbench, the camera mounted on the robotic arm captures desktop images, which are processed by the Inspect Agent to identify objects on the workbench and assign a unique functional-zone label based on predefined criteria.

This helps prioritize rooms suitable for further inspection.

As the VLM processes visual data, it aligns with a LiDAR-generated map, marking potential laboratory entrances for future navigation. After exploration, the agent returns to these identified coordinates for subsequent tasks. This strategy ensures reliable zone identification and task execution in the identified laboratory rooms.

C. Laboratory Mapping and Functional Zone Annotation

After identifying the laboratory rooms across the floor, we proceed to perform fine-grained exploration within each laboratory. Given that laboratory interiors are typically divided into distinct functional zones, we have developed a laboratory exploration and zoning classification strategy to ensure that the navigation agent can efficiently and accurately identify and label all workbenches and their corresponding functional areas.

At this stage, the agent operates according to the following prompt: "*You have entered a laboratory room. Your task is to conduct a thorough exploration to identify and label all workbenches. Each workbench corresponds to a specific functional zone. You must approach each workbench, capture detailed images of its surface, and determine the corresponding functional zone based on the observed items. Ensure that no workbench is missed, and that each is accurately labeled.*"

To simulate realistic chemical laboratory environments, we have constructed diverse functional scenarios across different workbenches. These scenarios are designed to reflect typical laboratory zones, defined as follows:

- **Experimental Operation Zone**: Dedicated to routine chemical experiments, featuring simulated experimental processes, laboratory equipment, and active reaction environments.
- **Chemical Storage Zone**: Used for the organized storage of chemical reagents, with chemicals categorized by properties and containers labeled with names, concentrations, and hazard symbols for safe handling.
- **Instrument and Equipment Zone**: Contains clean instruments and reagent bottles, neatly arranged for quick access by laboratory personnel.
- **Cleaning and Sterilization Zone**: For washing glassware, disinfecting tools, and preliminary wastewater treatment, equipped with sanitation facilities for a safe, hygienic environment.
- **Auxiliary and Office Zone**: Used for document storage, data recording, and providing workspace for lab staff.

These configurations ensure a comprehensive representation of functional areas, enabling the navigation agent to perform accurate semantic recognition and labeling during exploration.

To carry out the exploration tasks, the agent uses two cameras: one for room-level navigation and spatial understanding, and the other mounted on the robotic arm to capture detailed images of workbenches. The agent uses the front camera for initial scans and detecting workbench locations, with real-time depth data processing to ensure safe and precise approach. Once at the correct vantage point, the arm-mounted camera captures workbench images, which are processed by a VLM to analyze visible items and assign functional zones. The VLM outputs labels for each workbench, shown in Fig. 5. Each workbench's position and functional zone are annotated on a high-resolution 2D map, created with onboard LiDAR sensors, serving as a reference for subsequent tasks.

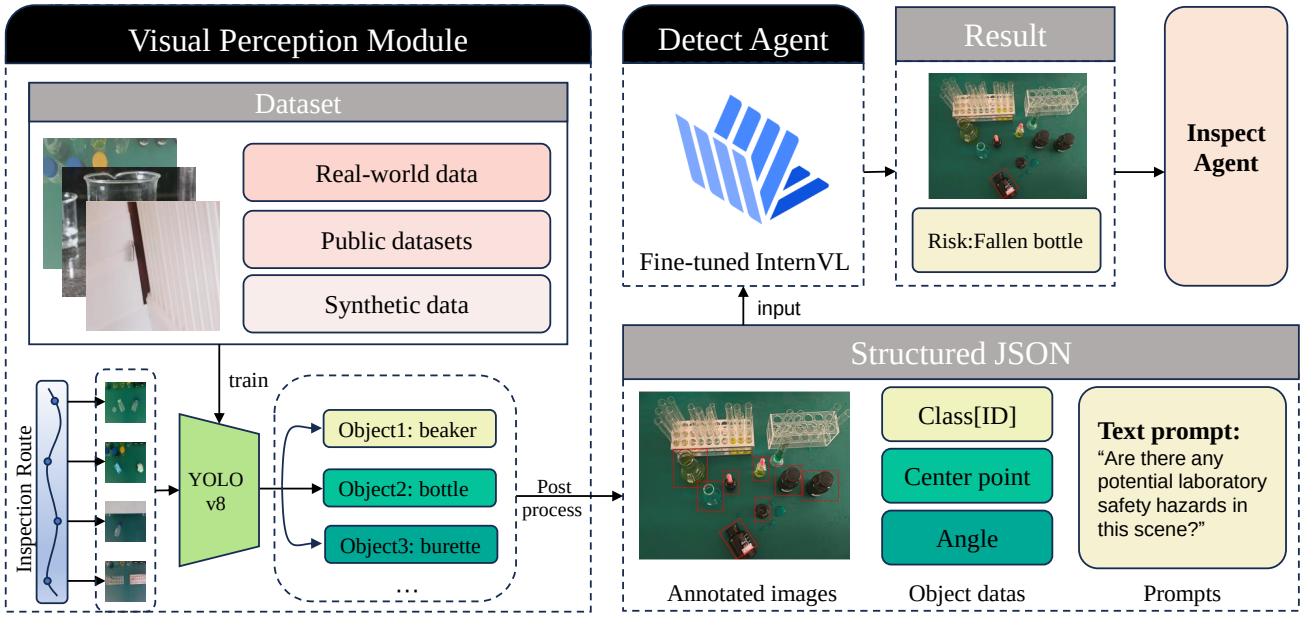

Fig. 6. **Overview of the laboratory risk detection framework.** The Visual Perception Module utilizes a YOLOv8 model trained on real-world, public, and synthetic datasets to detect objects from images collected along the inspection route. The detection results are post-processed into a structured JSON format, paired with prompt queries, and fed into the Detect Agent for risk reasoning. The identified risks are subsequently passed to the Inspect Agent to drive downstream decision-making and action planning.

After completing the exploration and classification, the agent compiles a complete semantic map showing the spatial distribution and functional categorization of work areas. The agent then proceeds to the next laboratory room, continuing until all rooms are explored and annotated, generating a detailed semantic representation of the laboratory environment for future inspection and monitoring tasks.

V. Risk Handling

After completing environment mapping and establishing the inspection route, the system enters the laboratory inspection phase. A two-stage detection framework, comprising a YOLO-based Visual Perception Module and a VLM-based Risk Reasoning Module, performs real-time object detection and hazard inference. The structured detection outputs are forwarded to the Inspect Agent, which serves as the high-level decision-making core. Based on the inferred risks, the agent plans appropriate actions or archives the information and uploads both risk assessments and execution results to the Interaction UI for human monitoring and record-keeping.

A. VLM-based Detection System

In the autonomous laboratory inspection workflow, the robot's initial route is planned to cover key functional zones, such as experimental workbenches and chemical storage areas. During real-time execution, the robot dynamically updates its navigation path for obstacle avoidance and captures images at predefined checkpoints. Each captured image is immediately transmitted to the risk detection module for real-time analysis, as illustrated in Fig. 6.

The detection system consists of two core modules: a YOLO-based Visual Perception Module and a VLM-based Detect Agent. Together, they perform semantic analysis of real-time visual data, identify potential safety hazards, and output structured risk information to support downstream robotic actions.

The Visual Perception Module is implemented using YOLOv8-OBB, which supports oriented bounding box detection for precise object localization in cluttered and rotated laboratory environments. To enhance generalization and robustness, we constructed a composite training dataset aggregated from three heterogeneous sources:

- **Real-World Data**: Approximately 1,000 images collected from laboratory environments, featuring common apparatuses (e.g., beakers, Erlenmeyer flasks, reagent bottles) under various layouts and lighting conditions.
- **Public Datasets**: Around 3,000 open-source images covering typical laboratory setups [17], [18].
- **Synthetic Transparent Object Data**: Approximately 1,000 images were generated by rendering CAD models of transparent laboratory containers, including reagent bottles, beakers, and Erlenmeyer flasks. We employed physically-based rendering techniques with randomized backgrounds and lighting conditions to create diverse and high-fidelity images, enhancing detection performance for low-contrast transparent objects.

The trained YOLO model detects objects with associated class labels, spatial coordinates, and rotation angles. Detection outputs are postprocessed into a structured JSON format, serving as input for subsequent reasoning and interaction.

To enable higher-level risk inference, we introduce the Detect Agent, built upon the open source InternVL2.5-4B model and fine-tuned for laboratory-specific semantic reasoning tasks. During fine-tuning, approximately 1,000 image-

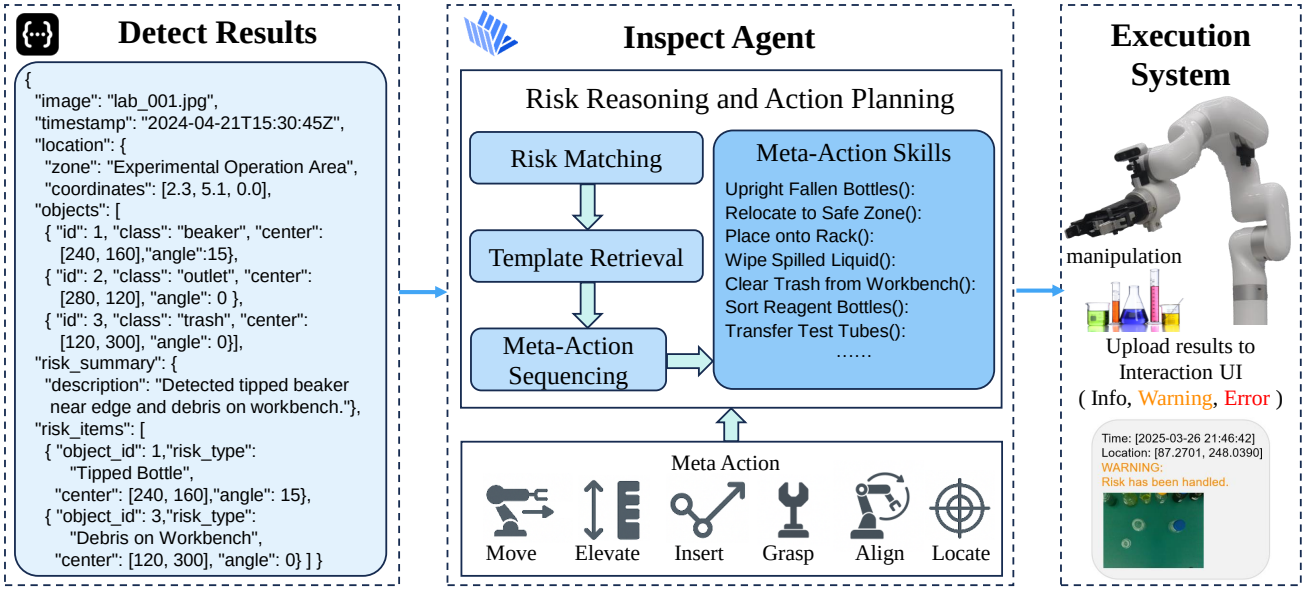

Fig. 7. **Overview of the autonomous risk mitigation and execution framework.** The Inspect Agent receives structured detection results containing identified objects and inferred risks. It performs risk reasoning and action planning through three stages: Risk Matching, Template Retrieval, and Meta-Action Sequencing. Based on predefined Meta-Action Skills, the system composes executable motion sequences, which are then carried out by the robotic manipulator. Execution results, including success or warnings, are uploaded to the Interaction UI for monitoring and logging.

text pairs were constructed based on outputs from the Visual Perception Module. Each sample includes:

- **Image**: Captured and processed by the YOLO module.
- **Text prompt**: Structured object descriptions (e.g., 'Object 1: Erlenmeyer flask, angle 15°, center [240, 180]'), followed by a guiding question (e.g., 'Are there any potential hazards on the laboratory workbench?').
- **Answer**: Human-annotated risk descriptions (e.g., 'The Erlenmeyer flask at position 1 may tip over.').

After fine-tuning, the Detect Agent accepts structured JSON inputs, performs multimodal reasoning to infer safety risks, and outputs structured feedback including risk types, associated object IDs, positions, and orientations. These results are subsequently passed to the Inspect Agent for downstream action execution and user-facing reporting.

B. Autonomous Risk Mitigation and Action Planning

After completing risk detection, the system enters the manipulation phase, where detected risks are assessed for automated mitigation or information archiving.

We introduce the Inspect Agent, a multimodal VLM serving as the high-level decision-making core. It receives structured image-text pairs from the detection phase and generates risk response strategies through multi-turn prompting. Internally, Inspect Agent employs a risk dictionary-based reasoning mechanism: when a detected risk matches a keyword in the predefined operable risk dictionary, the agent retrieves the corresponding action template and generates executable meta-action sequences; otherwise, the risk is categorized as non-operable and escalated for reporting. The overall risk reasoning and action planning process is illustrated in Fig. 7.

The manipulation phase consists of three operational modes:

Mode 1: No-Risk Reporting. If no risks are detected, the system packages the image, navigation position, and timestamp into an **Info** packet, uploading it to the Interaction UI for archiving and traceability.

Mode 2: Non-Executable Risk. For non-operable risks (e.g., uncovered hazardous chemicals, open flames), the system encapsulates the relevant information into an **Error** packet and issues an alert via the Interaction UI. When operable and non-operable risks coexist, operable risks are prioritized for mitigation.

Mode 3: Executable Risk. For operable risks (e.g., tipped bottles, scattered debris), Inspect Agent dynamically generates a meta-action sequence based on the "Laboratory Risk Mitigation Guide" prompt and detected bounding box information. Action templates are retrieved based on the risk type, with parameters like 3D positions and orientations adjusted in a real-time manner.

To enable flexible execution, we design a modular Meta-Action API system encompassing:

- **End-Effector Motion**: Move the robotic end-effector toward a target 3D position.
- **Target Localization**: Extract 3D coordinates of targets.
- **Gripper Actuation**: Control the gripper for object grasping or releasing.
- **Waypoint Insertion**: Insert intermediate points into the motion trajectory to reduce collision risk.
- **Height-Constrained Motion**: Maintain safe height to avoid table clutter or obstacles.
- **Orientation Alignment**: Adjust the end-effector orientation for optimal grasp stability.

Meta-actions are flexibly composed to form complete action sequences adapted to the current scenario. All actions rely on depth-calibrated spatial localization, and intermediate

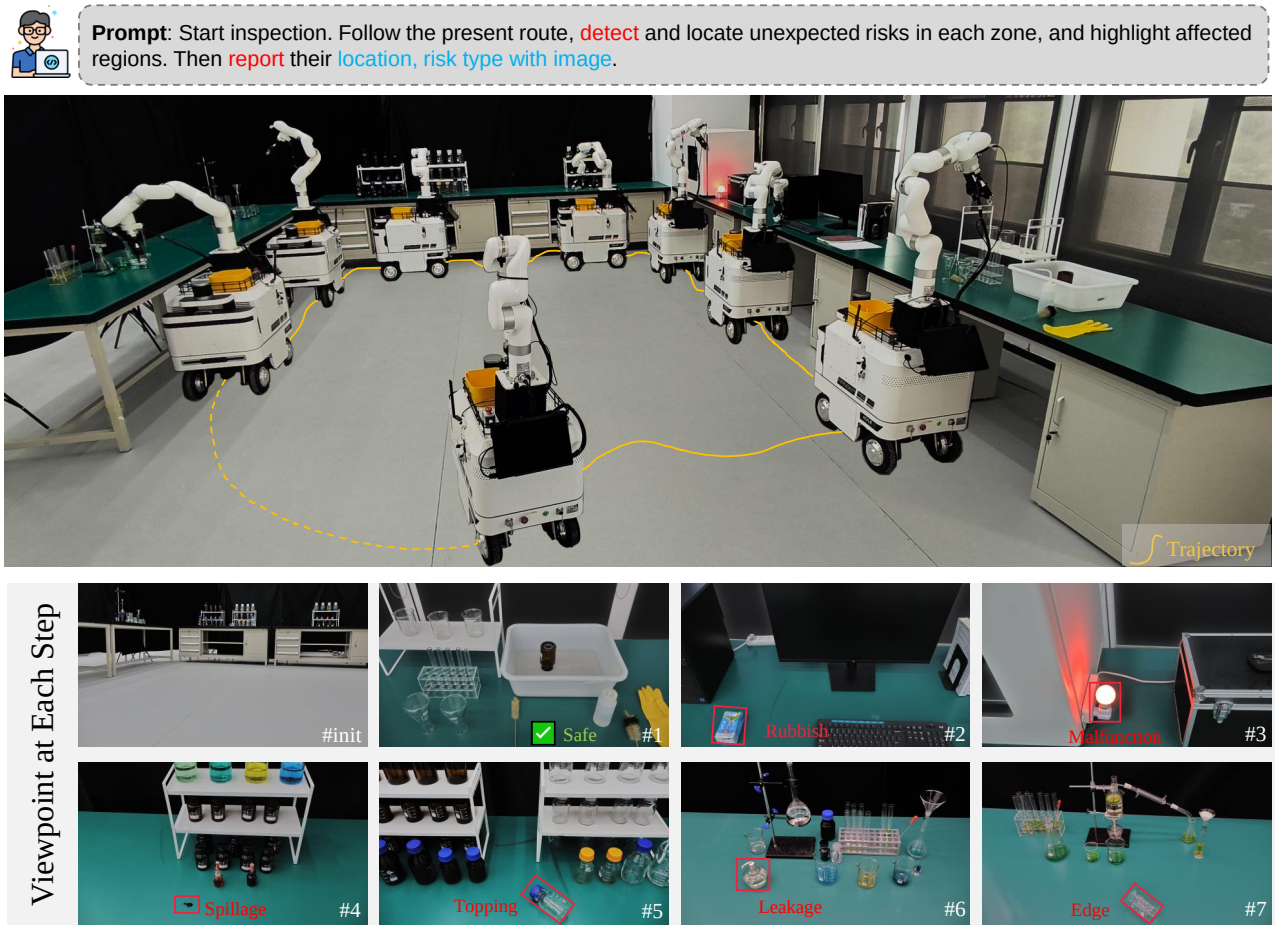

Fig. 8. **Robot inspection experiment.** Upon receiving the inspection command, the robot follows the predefined inspection route based on zone planning, and performs risk identification and labeling by analyzing the tabletop conditions captured by the camera, referencing the designated zones.

waypoints are incorporated for operational robustness.

The meta-action execution system currently does not implement retries. Upon task initiation, a pre-execution image and metadata are uploaded as an **Error** packet to the Interaction UI to mark the risk. After execution, whether successful or not, a post-execution image and status are uploaded as a **Warning** packet. The system then proceeds to the next inspection task without interruption.

When multiple risks are detected, the system prioritizes mitigating all operable risks before reporting non-operable risks, ensuring timely intervention for controllable hazards.

VI. Experiments

A comprehensive evaluation of our proposed robotic system was conducted in unfamiliar laboratory environments, assessing navigation performance, risk-detection accuracy, and risk-elimination manipulation. We also developed an intuitive human–robot interaction interface that accepts natural-language commands, enabling the robot to execute tasks with stability and precision. Experimental results demonstrate that the system delivers high reliability and robustness in complex and dynamic scenes.

A. Navigation Performance

Our ChemistryNav navigation framework runs in real time on a single RTX-4090 server and connects to the inspect agent via SSH. Predefined APIs process user instructions and preloaded prompts, parse live video from the vehicle-mounted camera, and generate corresponding motion commands.

To evaluate the practicality and robustness of ChemistryNav in real-world lab environments, we conducted deployment experiments focused on the robot's autonomous navigation. The experiment includes two main tasks: Full-Floor Exploration and Autonomous Laboratory Inspection. The robot autonomously explores an unknown environment, performs lab localization, identifies functional zones, and designs inspection paths for preparatory tasks before operation.

In the inspection experiment, the robot achieved an average response time of 8.5 s from command issuance to execution, with an average inspection success rate of 96.3%, demonstrating significant improvements using semantic-guided navigation. This strategy enabled efficient identification and semantic labeling of experimental areas, surpassing traditional methods in path control and deviation suppression. In complex environments, semantic-based dynamic adjustments effectively prevented detours and misjudgments. Real-time updates of the fine-grained semantic map maintained accurate alignment

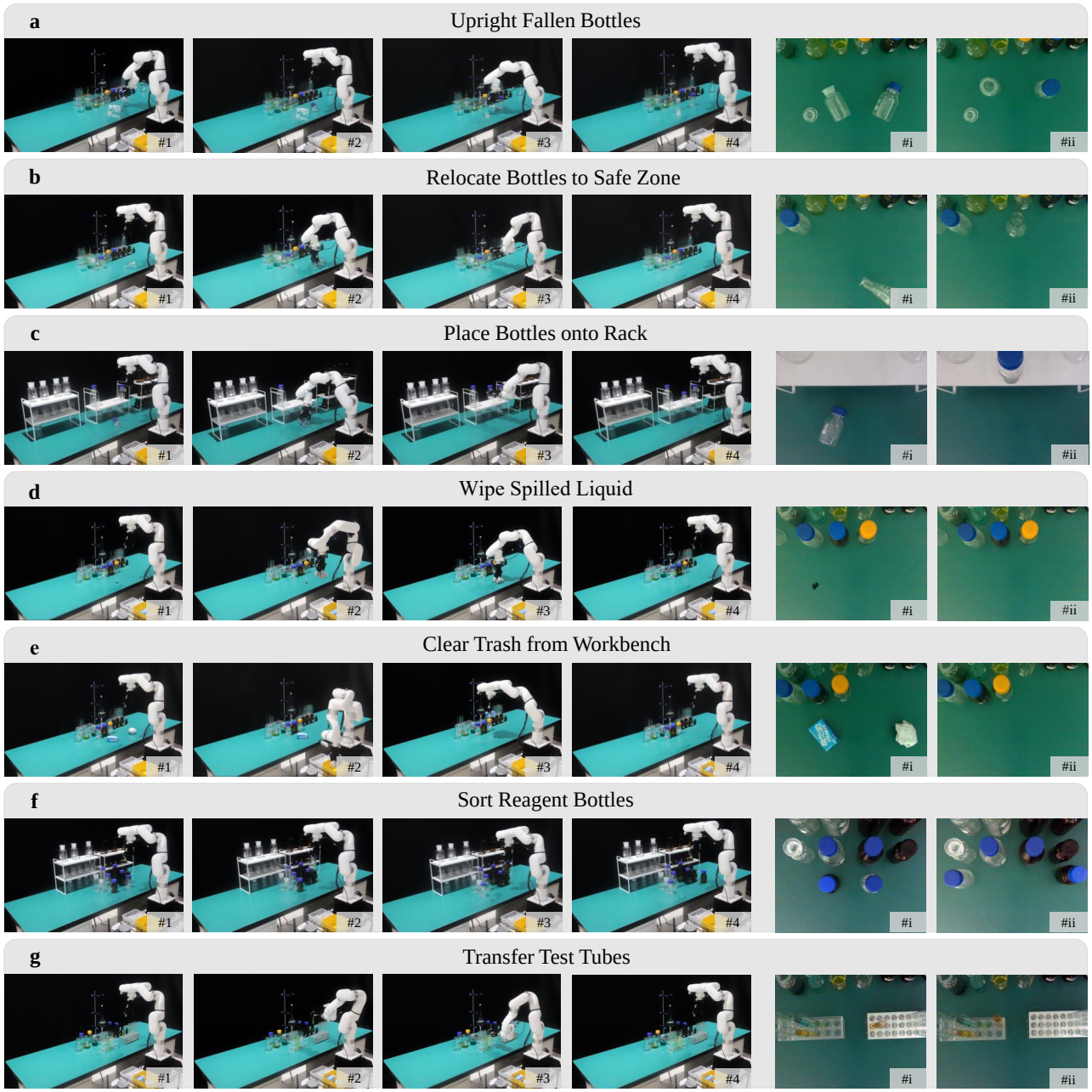

Fig. 9. **Robotic execution of risk-handling tasks in the laboratory environment.** (a) Upright two fallen transparent wide-mouth bottles to their vertical position. (b) Move a transparent conical flask away from the table edge toward a safer inner region. (c) Upright a fallen bottle and place it onto an empty slot of the visible bottle rack. (d) Use a sponge picked from the onboard storage basket to wipe spilled liquid on the tabletop. (e) Grasp trash items and place them into the onboard waste bin. (f) Sort transparent and opaque reagent bottles into correct categories. (g) Transfer solution-filled test tubes from one rack to another.

between robot positioning and environmental data, enhancing task stability and providing reliable support for subsequent risk detection and mitigation.

As shown in Fig. 8, during inspection, the robot follows a predefined route and adjusts the path based on environmental changes, performing risk assessments and operations in each area to successfully complete the inspection task.

B. Detection and Risk Reasoning Performance

The system employs a YOLOv8-OBB model to detect laboratory objects, including their categories, positions, and orientations. The model is trained on a combined dataset consisting of 3,000 public laboratory images, 1,000 real-world captured images, and 1000 synthetically generated transparent object images. The evaluation results on a held-out test set of 800 images are summarized in Table I.

To further investigate the impact of training data composition, an ablation study is conducted, comparing three configurations: Public only, Public + Real, and Public + Real + Synthetic. The overall detection success rates and transparent wide-mouth bottle detection rates are reported in Table I. As shown in the table, adding real-world captured

TABLE I
ABLATION STUDY ON TRAINING DATA COMPOSITION AND DETECTION SUCCESS RATES.

Training Data	Overall SR (%)	Transparent Object SR (%)
Public only	85.2	68.1
Public + Real	89.7	74.3
Public + Real + Synthetic	**93.1**	**88.6**

data improves the overall success rate from 85.2% to 89.7%. Further incorporating synthetic transparent object data boosts the overall detection success to 93.1%. Notably, the success rate for detecting transparent wide-mouth bottles increases dramatically from 68.1% to 88.6%, highlighting the significant benefit of synthetic data in enhancing the detection of low-contrast transparent objects.

Based on the detection outputs, the Detect Agent infers associated laboratory risks for each detected object. Manual verification indicates that the risk classification accuracy reaches 93.5%, demonstrating the agent's reliability and applicability in autonomous inspection tasks.

C. Action Execution Performance

To evaluate the system's capability for autonomous laboratory risk mitigation, we designed a series of representative task experiments, as illustrated in Fig. 9. The tasks include: (a) upright fallen bottles; (b) relocate edge-exposed objects; (c) place tipped bottles onto racks; (d) wipe spilled liquids; (e) clear debris from workbenches; (f) sort unclassified reagent bottles; and (g) transfer misplaced tubes. Each task was independently tested over 20 trials, recording both action execution status and final task outcomes. Two evaluation metrics were used:

- **Action Success Rate**: The proportion of trials where all meta-actions (e.g., localization, grasping, moving, placing) were successfully executed.
- **Task Completion Rate**: The proportion of trials where the risk was completely mitigated after execution.

TABLE II
ACTION EXECUTION AND TASK COMPLETION SUCCESS RATES ACROSS DIFFERENT RISK-HANDLING TASKS.

Risk Handling Task	Action Success Rate (%)	Task Completion Rate (%)
Upright Fallen Bottles	100.0	95.0
Relocate Bottles to Safe Zone	100.0	90.0
Place Bottles onto Rack	85.0	80.0
Wipe Spilled Liquid	90.0	70.0
Clear Trash from Workbench	90.0	85.0
Sort Reagent Bottles	85.0	80.0
Transfer Test Tubes	85.0	70.0
Overall Average	**90.7**	**81.4**

The detailed results are summarized in Table II. Overall, the system achieves an Action Success Rate exceeding 90% and a Task Completion Rate exceeding 80%. In structured tasks such as bottle uprighting and reagent bottle sorting, success rates remained consistently high. In contrast, more challenging tasks like liquid wiping and tube transferring exhibited slightly lower rates, primarily due to higher precision requirements where minor deviations impacted final task outcomes. These results demonstrate the effectiveness of combining vision-language reasoning with meta-action sequence control for autonomous laboratory risk mitigation, while also identifying opportunities for further improvements in precision-critical manipulations.

D. Interaction UI

To enhance human–robot interaction, we developed a web-based user interface that communicates in real time with VLM API. The UI lets operators issue commands to the robot and receive live status updates for each workspace zone.

The Inspect agent, responsible for human interaction, is implemented via the GPT-4V API running on an onboard Nvidia Jetson Nano, providing convenient access to vehicle peripherals. Upon receiving human task commands, the agent interprets and dispatches them to the appropriate backend modules—navigation and manipulation—through predefined APIs. The responses are displayed on the user interface.

For mobility, the Inspect agent communicates with the ChemistryNav navigation agent, retrieving motion commands that are relayed as ROS messages to the robotic base's motion controller. For risk assessment, the agent interacts with the detection agent, analyzing captured images and risk data (e.g., hazard type and location), determining whether the risks can be autonomously resolved or need escalation. Real-time log reports are generated, including timestamps, spatial coordinates, message types (**Info /Warning /Error**), Inspect Agent responses (risk assessments or operational outcomes), and relevant images.

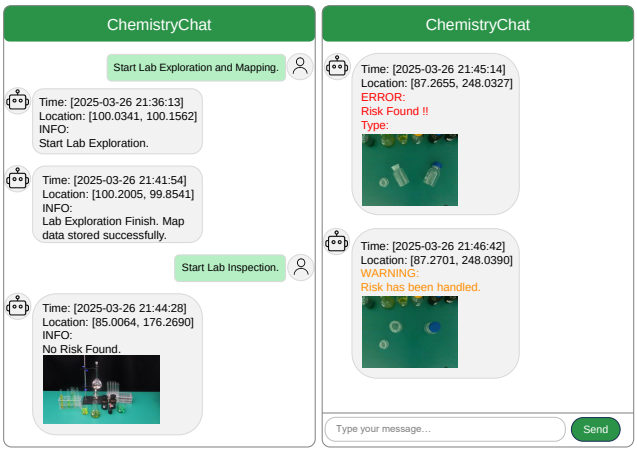

Fig. 10. **Screenshot of the Web-based UI.** Commands sent to the *Inspect agent* launch exploration and inspection tasks, while the agent returns real-time status in a predefined format reporting, e.g., whether exploration has finished, whether any risks were detected during inspection, and so on.

As shown in Fig. 10, issuing an Exploration command triggers automatic vision-language navigation (VLN), generating a high-precision 2D semantic map with functional zones and inspection paths. When an Inspection command is given,

the robot dynamically plans its route and outputs messages indicating identified issues in each zone.

VII. Conclusion

This paper presents an autonomous laboratory inspection robot (ALARMbot) that leverages foundation models for intelligent safety management in complex laboratory environments. The proposed system integrates a mobile platform, multimodal perception (including LiDAR and RGB-D cameras), and a 6-DoF manipulator, enabling comprehensive inspection and intervention tasks. By fusing LiDAR-based mapping with vision-and-language models, the robot achieves semantic-aware navigation and fine-grained functional zone recognition. The hierarchical framework combines YOLOv8-OBB visual perception (achieving 93.1% mAP on custom datasets) with vision-language risk reasoning for real-time hazard detection and analysis. Experimental results demonstrate that the robot autonomously intervenes in operable risks with an average response time of 8.5 seconds and navigates complex laboratory layouts with a 96.3% success rate, ensuring reliable and effective hazard mitigation. The core innovations of this work include the integration of foundation models for semantic perception and risk reasoning, as well as a modular manipulation system for autonomous intervention. In the future, this approach can be extended to other high-risk environments such as industrial plants and chemical warehouses. Further research will focus on enhancing multi-robot collaboration, adaptive learning in dynamic scenarios, and deeper human-robot interaction to achieve more intelligent and scalable safety management solutions.

References

[1] L. Ali, F. Alnajjar, M. M. A. Parambil, M. I. Younes, Z. I. Abdelhalim, and H. Aljassmi, "Development of YOLOv5-based real-time smart monitoring system for increasing lab safety awareness in educational institutions," *Sensors*, vol. 22, no. 22, p. 8820, 2022.

[2] S. Tang, D. Roberts, and M. Golparvar-Fard, "Human-object interaction recognition for automatic construction site safety inspection," *Automation in Construction*, vol. 120, p. 103356, 2020.

[3] A. Alayed, R. Alidrisi, E. Feras, S. Aboukozzana, and A. Alomayri, "Real-time inspection of fire safety equipment using computer vision and deep learning," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13290–13298, 2024.

[4] J. Feng, G. Hamilton-Fletcher, N. Ballem, M. Batavia, Y. Wang, J. Zhong, M. Porfiri, and J.-R. Rizzo, "Robust computer-vision based construction site detection for assistive-technology applications," *arXiv preprint arXiv:2503.04139*, 2025.

[5] L. C. O. Tiong, H. J. Yoo, N. Y. Kim, K.-Y. Lee, S. S. Han, and D. Kim, "Machine vision for vial positioning detection toward the safe automation of material synthesis," *arXiv preprint arXiv:2206.07272*, 2022.

[6] Z. Chen, H. Chen, M. Imani, R. Chen, and F. Imani, "Vision language model for interpretable and fine-grained detection of safety compliance in diverse workplaces," *Expert Systems with Applications*, vol. 265, p. 125769, 2025.

[7] P. Rea, E. Ottaviano, F. J. Castillo-García, and A. Gonzalez-Rodríguez, "Inspection robotic system: design and simulation for indoor and outdoor surveys," in *Innovations in Mechatronics Engineering*, pp. 313–321, 2022.

[8] Z. Chen, C. Song, B. Wang, X. Tao, X. Zhang, F. Lin, and J. C. Cheng, "Automated reality capture for indoor inspection using BIM and a multi-sensor quadruped robot," *Automation in Construction*, vol. 170, p. 105930, 2025.

[9] J. C. Wilson and P. A. Berardo, "Automatic inspection of hazardous materials by mobile robot," in *IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, vol. 4, pp. 3280–3285, 1995.

[10] S. N. Reddy, V. Kurrey, M. Nagar, and G. R. Gupta, "Action recognition based industrial safety violation detection," *arXiv preprint arXiv:2412.05531*, 2024.

[11] S. Lu, Y. Zhang, and J. Su, "Mobile robot for power substation inspection: A survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 830–847, 2017.

[12] D. Zhang and Z. Guo, "Mobile sentry robot for laboratory safety inspection based on machine vision and infrared thermal imaging detection," *Security and Communication Networks*, vol. 2021, no. 1, p. 6612438, 2021.

[13] M. Di Castro, *A novel robotic framework for safe inspection and telemanipulation in hazardous and unstructured environments*. PhD thesis, Industriales, 2019.

[14] B. Zhao, C.-d. Wu, X. Zhao, R.-h. Sun, and Y. Jiang, "Research on hybrid navigation algorithm and multi-objective cooperative planning method for electric inspection robot," *Energy Reports*, vol. 9, pp. 805–813, 2023.

[15] K. Hasselmann, M. Malizia, R. Caballero, F. Polisano, S. Govindaraj, J. Stigler, O. Ilchenko, M. Bajic, and G. De Cubber, "A multi-robot system for the detection of explosive devices," *arXiv preprint arXiv:2404.14167*, 2024.

[16] L. Zhang, X. Hao, Q. Xu, Q. Zhang, X. Zhang, P. Wang, J. Zhang, Z. Wang, S. Zhang, and R. Xu, "MapNav: A novel memory representation via annotated semantic maps for VLM-based vision-and-language navigation," *arXiv preprint arXiv:2502.13451*, 2025.

[17] "Science objetcs dataset." Available at: https://universe.roboflow.com/campus-74j4n/science-objetcs, Oct. 2023. Roboflow Universe. Accessed on 2025-04-29.

[18] "Big data dataset." Available at: https://universe.roboflow.com/final-project-roxvb/big-data-4r0y4, Dec. 2024. Roboflow Universe. Accessed on 2025-04-29.