IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 March 2026, accepted 11 April 2026, date of publication 16 April 2026, date of current version 6 May 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3684055

RESEARCH ARTICLE

A Hierarchical LLM-Based Framework for Heterogeneous Multi-Robot Orchestration in High-Risk Energy Facility Maintenance

JUNGI LEE[1,2], SEU-JAN KIM[1], GEONHYUP LEE[2], KANGMIN KIM[2], (Student Member, IEEE), JIMIN JEON[2], SEOK-KAP KO[1], AND KYOOBIN LEE[2,3,4], (Member, IEEE)

[1]Electronics and Telecommunications Research Institute, Buk-gu, Gwangju 61012, South Korea
[2]Department of AI Convergence, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea
[3]3GIST Institute for Artificial Intelligence, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea
[4]Graduate School of AI Policy and Strategy, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea

Corresponding author: Kyoobin Lee (kyoobinlee@gist.ac.kr)

This work was supported by Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government (MSIT), Development of a Foundation Model-Based Power Sector-Specialized Agent for Stable Power Generation and Operation under Grant RS-2025-02304557.

ABSTRACT Maintenance and operation of critical energy infrastructure require extreme precision and strict adherence to safety protocols to prevent catastrophic failures. While recent advancements in general-purpose Vision-Language-Action (VLA) foundation models have shown promise in robotics, their inherent stochasticity and lack of procedural precision often result in unacceptable safety risks in high-stakes industrial environments. To address these limitations, this paper proposes a novel Hierarchical Multi-Robot Orchestration Framework designed for the coordinated control of heterogeneous robots in energy facility maintenance. The proposed framework decouples high-level cognitive reasoning from low-level execution by utilizing a local Large Language Model (LLM) as a strategic orchestrator. Grounded in a Structured Environmental Representation ($\mathcal{M}$), the LLM employs Chain-of-Thought (CoT) reasoning to decompose complex maintenance missions into verifiable sub-tasks, effectively bridging the gap between linguistic intent and physical constraints. These tasks are then dispatched to specialized execution modules: a manipulation unit utilizing Action Chunking with Transformers (ACT) adapted for high-precision industrial tasks, an autonomous navigation unit for Simultaneous Localization and Mapping (SLAM)-based pathfinding, and a vision-guided logistics unit for retrieval. Experimental evaluations in a simulated power plant environment validate the efficacy of the framework. A comparative ablation study demonstrates that utilizing structured environmental grounding significantly enhances the reasoning reliability of large-scale models compared to unstructured text baselines. Specifically, the GPT-OSS (20B) model achieved a peak planning success rate of 91.7%, with a notable proficiency in handling long-horizon Strategic commands (73.3%), validating that structural clarity is essential for complex causal inference. Furthermore, the integrated execution layer demonstrated exceptional reliability, achieving a 90% success rate specifically in distribution panel maintenance tasks, confirming that the hierarchical decoupling of probabilistic reasoning and deterministic execution provides a reliable solution for autonomous maintenance.

INDEX TERMS Autonomous maintenance, hierarchical task planning, large language models, learning from demonstration, multi-robot orchestration.

I. INTRODUCTION

The maintenance and operational management of high-risk energy facilities, such as power plants and high-voltage substations, present some of the most formidable challenges in field robotics. These environments are characterized

The associate editor coordinating the review of this manuscript and approving it for publication was Rajesh Kumar Tripathy.

by extreme hazards, including high-voltage electricity, and complex structural layouts. Consequently, maintenance tasks are inherently safety-critical and multifaceted, ranging from long-range facility patrolling and parts logistics to high-precision manipulation of delicate electrical components. Given the life-threatening risks associated with human entry into these zones, there is an urgent industrial demand for autonomous systems capable of executing these diverse and dangerous tasks without human intervention.

The achievement of a full robotic maintenance system in such environments is complicated by the requirement that no single robotic platform can possess all the physical capabilities necessary for every maintenance scenario. A comprehensive maintenance mission typically requires the orchestration of heterogeneous robots, such as Automatic Guided Vehicles (AGVs) for mobility, bimanual manipulators for expert interaction, and collaborative arms for tool retrievals. Coordinating these diverse agents as a unified system requires not only high-level strategic reasoning but also the ability to handle long-horizon tasks that involve complex dependencies between different robotic skills.

Recent advances in Large Language Models (LLMs) have demonstrated remarkable efficacy in high-level task planning and semantic understanding [1], [2]. However, deploying these models for autonomous maintenance in energy facilities presents critical challenges that remain largely unaddressed. First, existing frameworks predominantly focus on single-agent scenarios, optimizing instruction following for individual robots. The application of LLMs to multi-robot orchestration remains significantly under-explored, particularly in heterogeneous fleets where tasks must be strategically assigned based on specific robotic capabilities. Second, current end-to-end models exhibit severe limitations in handling long-horizon tasks. Maintenance missions typically require prolonged sequences of interdependent actions. Finally, in terms of physical safety, the inherent stochasticity and 'black-box' nature of these models pose unacceptable risks. In high-voltage environments, a single reasoning error or a minor trajectory jitter can lead to catastrophic failure. Consequently, current approaches struggle to simultaneously achieve the strategic orchestration of multiple agents, the temporal consistency for long-horizon planning, and the procedural safety required for industrial operations.

In addition, the operational landscape of energy facilities requires diverse and complex physical interactions, such as toggling small switches, accessing recessed control panels, and maneuvering heavy levers. These environments are characterized by their unstructured layouts and restricted accessibility, which pose significant challenges for real-world robotic deployment. Although large-scale general-purpose robotic datasets, such as Open X-Embodiment [3], have advanced the field, they fundamentally fail to capture these domain-specific kinematic attributes. Consequently, relying solely on pre-existing datasets presents substantial limitations when training agents to perform precise manipulation tasks within the safety-critical context of energy infrastructure.

To overcome these limitations, a novel hierarchical LLM-based framework is proposed for the orchestration of heterogeneous robotic systems in high-risk maintenance tasks. The proposed system departs from risky end-to-end architectures by decoupling high-level cognitive orchestration from low-level execution expertise. The framework consists of a two-tier hierarchy: a **Strategic Task Planning Layer** and a **Skill-Based Execution Layer**. The planning layer leverages the reasoning power of an LLM grounded in a structured environmental representation model to decompose abstract human commands into a sequence of safety-verified sub-tasks. These tasks are then assigned to specialized modules, ensuring that navigation, manipulation, and retrieval are handled by controllers of expert-level that are tailored to the model of each specific robot.

The primary contributions of this work are as follows:

- A hierarchical orchestration framework is introduced that utilizes a local LLMs to enable the seamless coordination of heterogeneous robotic agents for complex, long-horizon maintenance missions in high-risk environments.
- A Structured Environmental Representation model is developed to ground the orchestrator's reasoning in strict facility metadata. This approach is empirically proven to suppress hallucinations and ensure that multi-robot plans adhere to industrial protocols compared to unstructured textual grounding.
- A set of specialized navigation and manipulation skills is proposed, including mission-aware pathfinding and deterministic action chunking, which provide the high-precision execution required for safety-critical tasks.
- Extensive experimental validation is provided in a simulated energy facility testbed, demonstrating that the proposed framework significantly outperforms baseline approaches in both strategic planning accuracy and physical execution reliability.

The remainder of this paper is organized as follows. Section II reviews related work in the fields of multi-robot orchestration and LLM-based robotic task planning and applications. Section III details the proposed hierarchical orchestration framework and the integration of the structured environmental representation model. Section IV describes the implementation of the specialized execution layer, focusing on the ACT-based manipulation policy and semantic navigation. Section V presents the experimental setup utilizing a high-fidelity power plant testbed and a heterogeneous robotic fleet. Section VI analyzes the quantitative and qualitative results, and Section VII concludes the paper with a summary of findings and directions for future research.

II. RELATED WORK

The orchestration of autonomous systems for high-risk industrial environments is an interdisciplinary challenge that integrates multi-robot coordination, semantic task planning, and specialized control architectures. This section provides a comprehensive review of the state-of-the-art in these fields and identifies the technical gaps addressed by the proposed framework.

A. MULTI-ROBOT ORCHESTRATION

Multi-robot orchestration serves as the operational backbone for autonomous fleet management, necessitating the precise synchronization of heterogeneous agents. Historically, this challenge has been predominantly framed through the lens of Multi-Robot Task Allocation (MRTA), focusing on the optimal distribution of tasks among a fleet [4]. Traditional methodologies have used primarily market-based mechanisms, such as auction algorithms [5], [6], [7] and optimization-based approaches such as Mixed-Integer Linear Programming (MILP) [8], [9] to minimize total energy consumption or completion time. Although these methods provide mathematical guaranties for task distribution, they often assume task independence and operate in structured environments with static constraints [10].

However, in the context of high-risk energy facilities, maintenance missions are characterized by intricate temporal and logical dependencies that extend beyond simple spatial optimization [11], [12]. For instance, a maintenance mission may require a robot to successfully retrieve a specialized voltage meter from a storage area before it can proceed to measure the electrical potential of a high-voltage switchboard. Classical MRTA methods often struggle to incorporate these high-level semantic prerequisites into low-level numerical optimization frameworks [13]. The failure to account for such procedural constraints can lead to deadlocks or safety violations in high-stakes environments.

Furthermore, effectively managing a diverse team—where a logistics unit performs retrieval and a bimanual manipulation unit executes the technical inspection—requires an orchestration framework capable of sophisticated reasoning over both the physical capabilities of the agents and the logical structure of the mission. To broaden the scope beyond traditional task allocation algorithms, recent paradigms have increasingly focused on embodied multi-robot cooperation [14]. These studies emphasize real-world, dynamic physical coordination and shared sensory context among agents, moving past mere abstract task distribution to solve joint physical tasks. While such research suggests that bridging the gap between abstract planning and physical execution is essential for the seamless operation of a heterogeneous fleet [15], scaling these embodied approaches to zero-tolerance energy facilities remains challenging. The proposed framework addresses these limitations by introducing a hierarchical structure that manages both complex semantic dependencies and agent heterogeneity within a unified system.

B. LLM-BASED STRATEGIC PLANNING AND THE GROUNDING PROBLEM

The emergence of Large Language Models (LLMs) has fundamentally shifted the paradigm of robotic task planning from symbolic logic to natural language reasoning. Frameworks such as SayCan [16] and Inner Monologue [17] have demonstrated that LLMs can effectively decompose abstract human intentions into a sequence of executable action primitives. Subsequent research has expanded these capabilities to include long-horizon planning in open-world environments [18] and real-time error recovery through multimodal feedback.

Specifically, recent developments like SayNav [19] have successfully integrated LLMs with 3D semantic maps to enable long-range navigation in unknown environments by generating sub-goals based on discovered objects. This approach highlights the potential of LLMs to act as high-level planners when grounded in structured spatial data. However, despite these advancements, the ''grounding problem''—mapping abstract linguistic tokens to physical world coordinates—remains a significant bottleneck [20], [21]. While some recent studies have attempted to use LLMs for multi-agent systems [22], they frequently operate in simplified simulation environments where the cost of failure is negligible. In high-risk energy infrastructure, however, the orchestrator must operate within a ''zero-tolerance'' safety regime. Current LLM-based planners often lack a structured representation of the environment, leading to potential hallucinations or procedurally unsafe plans [23]. The proposed system mitigates this risk by grounding the LLM in a structured environment, ensuring that every decomposed task is validated against actual facility metadata.

C. FROM GENERAL-PURPOSE VLAs TO HIERARCHICAL CONTROL

The landscape of robotic control has been transformed by the advent of Vision-Language-Action (VLA) foundation models. Models such as RT-2 [24] and OpenVLA [25] have demonstrated the capability to map perceptual inputs directly to motor commands, offering impressive cross-embodiment generalization. This trend has been further accelerated by initiatives like Project GR00T [26], which aim to ignite Vision-Language Models (VLMs) toward the embodied space, enabling robots to reason about physical properties without task-specific training.

However, a critical bottleneck in deploying these foundation models for facility maintenance is their limitation in maintaining temporal consistency during long-horizon missions. Unlike atomic interactions, maintenance tasks require prolonged sequences of interdependent actions (e.g., retrieval → navigation → manipulation). Standard VLA models are prone to ''catastrophic forgetting'' or context drift over these extended timelines, often losing track of the original high-level goal while focusing on immediate low-level actuation [27].

To bridge the gap between semantic reasoning and physical reliability, there is an increasing consensus toward hierarchical architectures. Rather than relying on a single end-to-end model, recent approaches advocate for separating high-level strategic planning from low-level execution. This paradigm allows for the integration of specialized imitation learning policies that are explicitly designed for high-precision and low-variance control, ensuring that critical operations are executed with the repeatability required by industrial safety standards.

D. ROBOTIC APPLICATIONS IN ENERGY INFRASTRUCTURE MAINTENANCE

Research specifically aimed at automating energy infrastructure has focused on the deployment of robotic systems for inspection and maintenance in hazardous zones [28]. Given the risks of radiation in nuclear facilities and high-voltage exposure in electrical substations, previous works have introduced specialized platforms for tasks such as nuclear power plant decommissioning [29] and the autonomous patrolling of power distribution lines [30], [31]. These systems have significantly reduced human exposure to extreme hazards by providing remote sensing and localized manipulation capabilities.

However, a critical limitation of existing systems is their design for singular, highly specific tasks, which results in a lack of flexibility to adapt to multifaceted and evolving maintenance scenarios [32], [33]. Most operational models in this domain rely on rigid, pre-programmed mission profiles that struggle to handle the dynamic complexity of a full-scale maintenance cycle—ranging from initial tool retrieval and navigation to high-precision component manipulation. Furthermore, the orchestration of a heterogeneous robotic fleet remains a significantly underexplored area in energy infrastructure, particularly regarding the integration. [34].

Existing literature lacks a unified approach that can bridge the gap between high-level cognitive reasoning and low-level deterministic execution across diverse robotic agents in a ''zero-tolerance'' safety environment. This paper fills this gap by presenting a hierarchical framework that enables the seamless orchestration of multiple robotic agents. By grounding the strategic orchestrator in facility metadata and utilizing specialized expert modules for execution, the proposed system enables the performance of complex, multi-agent missions within a unified, safety-critical maintenance testbed.

III. PROPOSED HIERARCHICAL FRAMEWORK

To achieve robust and safety-compliant orchestration of heterogeneous robotic agents in high-risk energy facilities, a hierarchical framework is proposed. The system architecture is designed to decouple high-level strategic reasoning from low-level deterministic execution, thereby mitigating the risks associated with the stochastic nature of end-to-end foundation models.

A. SYSTEM ARCHITECTURE OVERVIEW

As illustrated in Fig. 1, the proposed framework consists of a two-tier hierarchy: the Strategic Task Planning Layer and the Skill-Based Execution Layer. These layers are interconnected through a centralized Structured Environmental Representation Model, which serves as the unique source of truth for environmental grounding and safety verification.

1) **Strategic Task Planning Layer**: This upper layer functions as the cognitive orchestrator. It utilizes a local LLM to parse abstract human instructions and decompose them into a temporally and logically consistent sequence of sub-tasks.
2) **Skill-Based Execution Layer**: This layer bridges the gap between abstract planning and physical action through LLM-driven agent dispatch and skill selection. Based on the inferred requirements of each sub-task, the orchestrator identifies the optimal robotic platform from the heterogeneous fleet and invokes the corresponding primitive skill. Once allocated, specialized expert modules execute these skills with deterministic reliability, handling navigation, high-precision manipulation, and logistics retrieval.

B. STRUCTURED ENVIRONMENTAL REPRESENTATION

The core of the orchestration logic resides in the structured environmental representation model. Unlike raw occupancy grids or point clouds, this model represents the facility as a collection of semantic entities and their functional relationships. The structured environmental representaion stores metadata including object categories, physical properties (e.g., ''high-voltage,'' ''movable''), and spatial coordinates within a structured relational schema.

Formally, the structured environmental representation model $\mathcal{M}$ is defined as a set of entities E and relations R:

$$\mathcal{M} = \{E, R, \mathcal{P}\} \tag{1}$$

where E denotes the set of objects (e.g., ''switchgear,'' ''circuit breaker''), R defines the spatial and logical links between entities (e.g., *near(south_lounge)*), and $\mathcal{P}$ represents the set of robotic protocols. By querying $\mathcal{M}$, the orchestrator can verify the feasibility of a sub-task and identify the most suitable robotic agent based on its kinematic capabilities and current state.

C. STRATEGIC ORCHESTRATION AND TASK DECOMPOSITION

Upon receiving a natural language command (e.g., ''Turn off the switch of switchgear in Room 4''), the Strategic Task Planning Layer initiates a Chain-of-Thought (CoT) [35] reasoning process. The orchestrator queries the structured environmental representation to identify the target object's state and the prerequisites for the mission. The task decomposition process can be modeled as a function f_{plan}:

$$\mathcal{T} = f_{plan}(C, \mathcal{M}, \mathcal{A}) \tag{2}$$

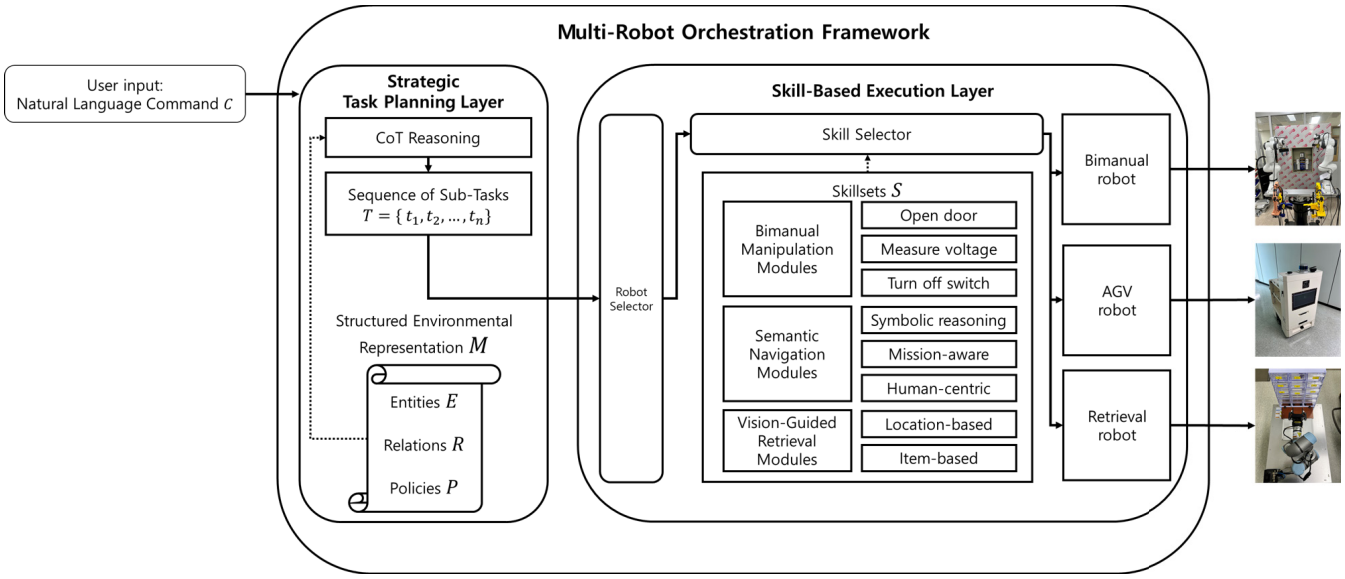


FIGURE 1. **Overall framework for hierarchical multi-robot orchestration with LLM.**

given a command C (e.g., "Cut off the power and check voltage."), the orchestrator queries $\mathcal{M}$ to verify prerequisites and constraints. And $\mathcal{A}$ denotes the set of available heterogeneous robots. The resulting mission sequence $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ ensures that all logical dependencies are satisfied, such as tool retrieval before manipulation.

D. AGENT-SPECIFIC SKILL EXECUTION

The Skill-Based Execution Layer translates high-level subtasks into motor commands. This layer leverages specialized foundation models and control algorithms that provide expert-level performance for specific domains:

- **Bimanual Manipulation Modules**: Employ deterministic models to perform high-precision interactions with electrical components, ensuring repeatability and safety that exceed the capabilities of general-purpose VLAs.
- **Semantic Navigation Modules**: Utilize semantic pathfinding and Simultaneous Localization and Mapping (SLAM) grounded in $\mathcal{M}$ to ensure collision-free movement in narrow industrial corridors.
- **Vision-Guided Retrieval Modules**: Handle logistics tasks, such as tool delivery, by coordinating with the mobile bases to maintain a continuous supply chain for the primary manipulation unit.

By separating these expert skills from the central orchestrator, the framework allows for the seamless integration of new robotic platforms without requiring a complete reconfiguration of the high-level planning logic.

Runtime Safety and Fail-Safe Mechanisms: While the Structured Environmental Representation ($\mathcal{M}$) acts as a cognitive invariant to prevent logically unsafe task sequences (e.g., operating a breaker before verifying voltage), formal safety compliance during physical execution relies on the deterministic low-level controllers of these expert modules. The execution layer integrates real-time runtime monitors independent of the LLM orchestrator. For the mobile bases, the SLAM navigation stack continuously processes Light Detection and Ranging (LiDAR) point clouds to detect dynamic anomalies, triggering an emergency stop (E-stop) fail-safe if an obstacle breaches the minimum safety radius. Similarly, the bimanual manipulation module operates within strictly bounded joint torque limits and pre-verified kinematic workspaces, ensuring that any unpredicted environmental deviations do not result in hardware damage. This strict decoupling ensures that probabilistic planning errors do not compromise the physical safety of the facility.

IV. TECHNICAL METHODOLOGY

The proposed framework enables autonomous maintenance of energy facilities by bridging high-level cognitive reasoning with low-level deterministic control. This section details the internal mechanisms of the LLM-based orchestrator, the formalization of the structured environmental representation model, and the specialized execution policies for heterogeneous agents.

A. LLM-BASED STRATEGIC TASK DECOMPOSITION

The Strategic Task Planning Layer serves as the cognitive brain of the framework, responsible for bridging the gap between abstract human intentions C and the physical capabilities of the heterogeneous fleet $\mathcal{A}$. To achieve high mission reliability in safety-critical environments, the orchestrator employs capability-based agent dispatch logic. The planning process is divided into three distinct phases: Decomposition, Agent Dispatch, and Skill Instantiation.

1) CHAIN-OF-THOUGHT TASK DECOMPOSITION

Upon receiving a complex command C, the orchestrator utilizes a CoT reasoning process to break down the mission

into a sequence of logical sub-tasks:

$$\mathcal{T} = \{t_1, t_2, \ldots, t_n\} \tag{3}$$

This reasoning ensures that prerequisite dependencies, such as retrieving a specific tool before initiating maintenance on an electrical distribution panel, are correctly identified and ordered in the temporal domain.

2) CONSTRAINT-AWARE AGENT DISPATCH

Unlike skill-first abstractions, this framework prioritizes the allocation of physical resources based on task constraints. For every sub-task t_i, the orchestrator first analyzes the operational requirements $Req(t_i)$ (e.g., workspace reachability, payload capacity, mobility type). Based on these constraints and the policy $\mathcal{P}$, the system identifies the optimal agent a_j from the heterogeneous fleet $\mathcal{A}$:

$$a_j = \text{Dispatch}(Req(t_i), \mathcal{A}, \mathcal{P}) \tag{4}$$

This step ensures that the task is assigned to a platform physically capable of entering the target zone and handling the target object (e.g., assigning the Bimanual Avatar for high-voltage switching vs. the AGV for transport), effectively filtering out infeasible assignments before execution planning begins.

3) SKILL INSTANTIATION AND PARAMETERIZATION

Once the agent a_j is selected, the orchestrator queries the agent's specific skillset $S(a_j)$ to select the precise executable primitive S_{exe}. This decoupling allows for heterogeneity; for instance, a "Move" task might trigger a wheel-based controller for an AGV but a gait-based controller for a quadruped. The final output is a structured task triplet:

$$t_i = \langle a_j, S_{exe}, \text{Target}_{obj} \rangle \tag{5}$$

which is dispatched to the specialized execution module of the selected robot a_j for deterministic performance.

B. IMPLEMENTATION OF THE STRUCTURED ENVIRONMENTAL REPRESENTATION MODEL

The Structured Environmental Representation Model ($\mathcal{M}$) serves as a semantic topology that grounds the orchestrator's decisions in physical reality, preventing hallucinations by constraining the search space. Instead of relying on abstract symbolic logic, $\mathcal{M}$ explicitly maps discrete operational zones to their constituent entities (E), spatial relations (R), and operational policies ($\mathcal{P}$).

1) CONSTRAINT-AWARE POLICY

The policy set $\mathcal{P}$ acts as a deterministic filter for agent-task assignment. Formally, a task t_i is dispatched to an agent a_j only if the agent's skillset $S(a_j)$ satisfies the technical requirements defined by the policy: $S(a_j) \supseteq Req(t_i)$. This ensures that high-risk operations are strictly decoupled from incapable platforms:

- **Manipulation** ($\mathcal{P}_{manip}$): This policy governs tasks requiring high-precision physical interaction. It dictates that complex maintenance operations, such as the measuring voltage of electrical panel, must be assigned to agents possessing bimanual coordination. The policy ensures that delicate electrical components are handled with the deterministic accuracy required for industrial safety.
- **Navigation** ($\mathcal{P}_{nav}$): This policy manages the assignment of autonomous movement tasks. It identifies agents equipped with advanced spatial awareness sensors (e.g., 3D LiDAR, RGB-D camera) to perform facility patrolling or to move between work zones. $\mathcal{P}_{nav}$ also incorporates safety constraints, ensuring that only agents with specific footprints or kinematic agility are assigned to navigate through the narrow corridors of the electrical room.
- **Retrieval** ($\mathcal{P}_{retr}$): This policy is dedicated to logistics and tool-preparation tasks. It routes operations such as label-based drawer opening and object retrieval to specialized agents, such as the UR3 robotic arm.

2) JSON INSTANTIATION AND INTEROPERABILITY

To ensure direct interoperability with the LLM orchestrator, the theoretical model $\mathcal{M}$ is instantiated as a structured JSON schema. Listing 1 presents the instantiated node for the electrical room. Crucially, the policy field is embedded directly within the metadata, serving as a hard constraint that allows the orchestrator to validate agent capabilities (Req) against the environment's safety rules before plan generation.

This structured approach allows the LLM to efficiently parse the environment by querying a location key and retrieving a comprehensive context snapshot. By treating relations as a verifiable dataset of spatial-semantic mappings, the system effectively bridges the gap between linguistic intent and robotic execution. The complete operational flow, orchestrating this world model with the strategic planner, is formalized in Algorithm 1.

C. BIMANUAL MANIPULATION SKILL FOR ELECTRICAL DISTRIBUTION PANELS

To address the precision requirements of high-voltage maintenance tasks—such as inserting a voltmeter probe into a specific terminal—we employ an adaptation of Action Chunking with Transformers (ACT) [36] as our low-level manipulation policy. Unlike VLA models that output a single action step per time step, ACT predicts a sequence of actions (a "chunk") based on current observations, significantly reducing the compounding errors typical in long-horizon manipulation. This module focuses on providing deterministic and repeatable trajectories for safety-critical operations, such as operating switches or handling internal components of the panel. The manipulation skill is integrated as a specialized execution primitive. When the orchestrator

```
{
  "node_id": 45,
  "canonical_name": "Electrical Room",
  "aliases": [
    "Electrical Area", "Electrical Equipment Room", "Power Equipment Room",
    "Electrical Maintenance Room", "Electrical Service Room"
  ],
  "entity": [
    "electrical_control_panel", "cable_tray", "lockout_tagout_board",
    "warning_signage", "insulated_gloves_box"
  ],
  "relation": {
    "near": [46, 40],
    "across_from": [43], // Semantic Link: Directly opposite the Circuit Breaker Room
    "inside": null
  },
  "policy": {
    "open_door": "bimanual", // Requires dual-arm manipulation
    "voltage_check": "bimanual",
    "switch_off": "bimanual",
    "retrieval": false, // Logistic tasks prohibited in high-voltage zone
    "navigation": "mobile"
  },
  "description": "High-voltage electrical room located across from the Circuit Breaker Room."
}
```

LISTING 1. Structured Environmental Representation for the "Electrical Room." The schema explicitly structures the environment into three semantic attributes: entity lists interactable assets, relation defines the topological connectivity for navigation, and policy imposes capability-based constraints (e.g., 'bimanual') to ensure safe agent dispatch.

dispatches a task t_i involving panel maintenance, the system loads the corresponding pre-trained policy. This policy maps real-time visual feedback and joint states directly to a sequence of coordinated bimanual actions. By utilizing the transformer-based architecture's ability to handle temporal dependencies, the robot performs smooth, synchronized movements required to navigate the narrow tolerances and complex layouts of electrical distribution panels.

D. NATURAL LANGUAGE-BASED SEMANTIC NAVIGATION

To facilitate seamless human-robot interaction in complex industrial layouts, the framework incorporates a Natural Language-Based Semantic Navigation module. This module enables the fleet to navigate through the facility without the need for manual coordinate inputs, instead relying on the semantic labels stored in the structured environmental representation $\mathcal{M}$. When the orchestrator dispatches a navigation task, the module translates natural language references (e.g., "Retrieve the tape measure and go to the Electrical Department.") into a sequence of topological waypoints. The semantic navigator performs Grounding and Spatial Reasoning by querying the structured environmental representation to identify the target entity's global pose and its surrounding operational constraints. By integrating a LiDAR-based SLAM system with semantic graph-matching, the robot can identify and navigate to specific assets even in dynamic environments. This approach ensures that the navigation is "mission-aware," meaning the robot does not just reach a coordinate but node itself in the optimal workspace configuration required for the subsequent manipulation task.

E. VISION-BASED RETRIEVAL MODULE FOR STORAGE

The logistics and tool-preparation phase of the mission is managed by a specialized Vision-Based Retrieval Module. This module utilizes a manipulator equipped with a parallel two-finger gripper that specializes in autonomous identification and retrieval of maintenance instruments from structured storage units.The retrieval process is driven by an eye-in-hand perception pipeline. When a task t_i specifies a target tool, the manipulator performs the following functional steps:

1) **Object Detection and Label Association**: The agent scans the storage drawer surfaces using an integrated RGB camera. A YOLOv8-based object detection model is employed to identify the semantic label associated with the required tool. This perception layer maps the visual bounding box information to the specific storage node defined in the structured environmental representation $\mathcal{M}$.
2) **Target Pose Estimation**: Upon identifying the correct drawer label, the module extracts the centroid coordinates of the target handle from the detection bounding box. These 2D pixel coordinates are converted into a 3D target pose relative to the robot base frame using depth estimation and extrinsic camera calibration, providing a static goal for the grasping maneuver.
3) **Cartesian Linear Retraction**: Once the gripper successfully grasps the handle, the module executes a Cartesian linear motion command. Instead of complex dynamic control, the manipulator performs a strictly defined retraction path along the end-effector's negative Z-axis. This linear trajectory is designed to align with the mechanical sliding axis of the drawer,

Algorithm 1 Hierarchical Multi-Robot Orchestration and Execution Framework

Require:
- Abstract natural language instruction C
- Structured Environmental Representation Model $\mathcal{M}$ (Entities E, Relations R, Policies $\mathcal{P}$)
- Fleet of heterogeneous robotic agents $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$

Ensure: Mission success status or failure analysis $\mathcal{F}$

Step 1: Strategic Task Decomposition
$\mathcal{T} \leftarrow f_{plan}(C, \mathcal{M}, \mathcal{A})$ ▷ LLM-based decomposition into sub-tasks $t_1, \ldots, t_n$
Initialize operational context $S_0 \leftarrow \mathcal{M}.get\_current\_state()$

Step 2: Agent Dispatch and Skill Execution
while $\mathcal{T}$ is not empty **do**
  $t_i \leftarrow \mathcal{T}.pop\_front()$ ▷ Extract current maintenance task
  // 1. Optimal Robot Dispatching
  $Req(t_i) \leftarrow$ ParseRequirements(t_i) ▷ Identify task constraints (e.g., location)
  $a_j \leftarrow$ DispatchAgent$(Req(t_i), \mathcal{A}, \mathcal{P})$ ▷ Select best-suited agent based on policy $\mathcal{P}$
  if a_j is `None` **then**
    return `Failure`: No available agent satisfies requirements $Req(t_i)$
  end if
  // 2. Context-Aware Skill Selection
  $S_{exe} \leftarrow$ SelectSkill$(t_i, S(a_j))$ ▷ Choose specific primitive from agent's library
  // 3. Execution and Monitoring
  Command agent a_j to execute S_{exe} on target Target_{obj}
  $h_i \leftarrow a_j.execute(t_i, S_{exe})$ ▷ Execution via specialized module
  // 4. Feedback and Reactive Re-planning
  if $h_i ==$ `Success` **then**
    $\mathcal{M}.update\_state(t_i,$ `Success`$)$ ▷ Update relational metadata in structured environmental representation
  else
    $\mathcal{M}.update\_state(t_i,$ `Failure`$)$ ▷ Update context for re-planning
    $\mathcal{T} \leftarrow f_{plan}(C, \mathcal{M}, \mathcal{A})$ ▷ Trigger reactive re-planning loop
    continue
  end if
end while
return `Success`: Entire mission sequence T finalized safely

minimizing lateral forces that could cause jamming during the opening process.

By abstracting the retrieval task into vision-guided targeting and deterministic linear execution, this module allows the retrieval agent to serve as a reliable logistics hub, ensuring that the primary maintenance fleet is consistently supplied with the necessary equipment.

V. EXPERIMENTAL SETUP

This section describes the physical testbed, the specifications of the heterogeneous robotic fleet, and the data acquisition protocols used to validate the proposed hierarchical orchestration framework.

A. INDUSTRIAL TESTBED AND ENVIRONMENT MODELING

Since performing robotic experiments in operational power plants is prohibited due to safety protocols and high-voltage risks, we established a high-fidelity experimental environment within a single-floor building. The facility was architected to mirror the topological layout of a standard energy facility, with various rooms and transit zones represented in the digital world modeling.

1) INDUSTRIAL TESTBED AND ENVIRONMENTAL CONFIGURATION

The experimental environment is established within a single-floor facility designed to validate cross-room autonomous navigation and multi-agent orchestration in an integrated manner. The workspace is structured to replicate the operational flow of a power plant maintenance zone, allowing robots to traverse between patrol, transit, and maintenance areas.

The testbed is established within a single-floor facility structured to evaluate the robustness of cross-room autonomous navigation and multi-robot coordination. The workspace is digitized into a high-fidelity spatial representation to support the grounding of strategic plans.

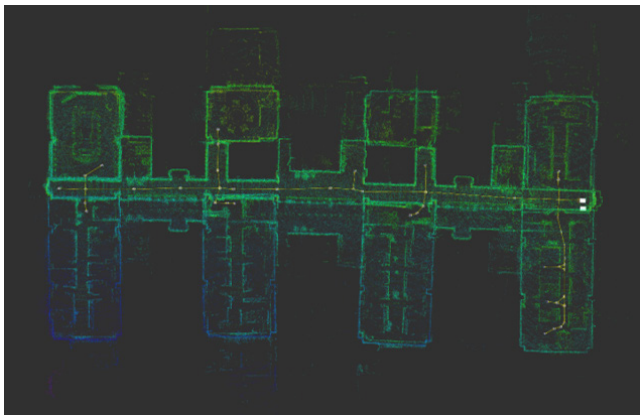
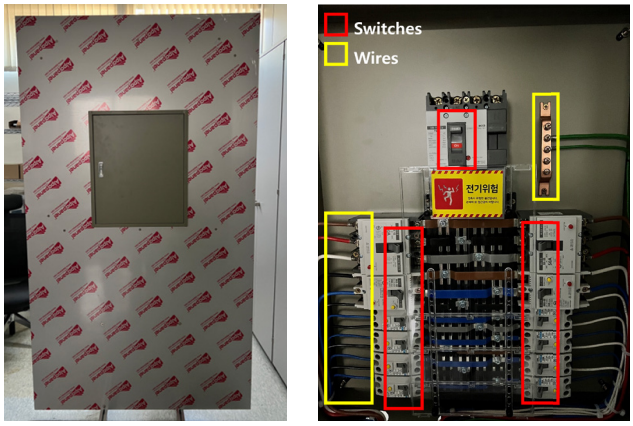


FIGURE 2. The dense 3D point cloud map constructed via LiDAR SLAM. This geometric representation captures vertical structural details, enabling real-time obstacle avoidance and precise localization in cluttered maintenance zones.



(a) Overall view of the panel (b) Internal layout of the panel

FIGURE 3. The industrial electrical distribution panel testbed: (a) illustrates the exterior appearance and initial state with the door closed; (b) highlights the internal operational interfaces, including switches and designated locations for voltage measurement.

- **Environmental Mapping**: To enable precise semantic navigation, the facility was digitized using a LiDAR-based SLAM framework. This process yielded a dual-layer representation: a 2D occupancy grid that functions as the topological backbone for global path planning, and a dense 3D point cloud, as shown in Fig. 2, that captures vertical geometric details essential for real-time obstacle avoidance and precise localization in cluttered maintenance zones.
- **Industrial Electrical Distribution Panel**: The primary maintenance target is a full-scale industrial electrical distribution panel. As shown in Fig. 3, the panel is equipped with various operational interfaces, including switches and wires. The environment incorporates realistic visual complexities such as metallic reflections and cluttered semantic labels, requiring high-precision perception and manipulation.
- **Modular Storage**: For retrieval and logistics tasks, a specialized storage unit is utilized with a 3 × 3 grid configuration. This unit serves as the physical grounding for the retrieval module, where tools and parts are stored in drawers. These targets are identified via a YOLOv8 [37]-based object detection model, which allows the robots to locate specific drawers and tools even in cluttered environments.

B. HETEROGENEOUS ROBOT FLEET

The experimental validation involves a fleet of three specialized robotic agents, each possessing distinct kinematic configurations and operational roles. These robots are registered within the global orchestrator with unique skillsets S, allowing for the seamless execution of multi-agent maintenance workflows.

1) BIMANUAL MAINTENANCE ROBOT

The bimanual platform, illustrated in Fig. 4(a), is specifically architected for high-precision manipulation tasks that require coordinated dual-arm control within the distribution panel workspace. The hardware integration features two Franka Emika Panda arms, providing a combined 14 degrees of freedom (DOF) for complex manipulation. Each arm is equipped with a parallel gripper and an Osmo Action 5 camera mounted on the end-effector. These cameras are configured in wide-angle mode to capture a broad field of view, ensuring that the visual perception system can monitor both the intricate details of the distribution panel and the surrounding workspace.

To develop the necessary autonomous skills, a maintenance dataset was constructed using the GELLO [38] teleoperation system. GELLO is a leader–follower bimanual teleoperation framework in which an operator directly manipulates the GELLO leader device, and the corresponding joint angles are transmitted in real time to a Franka Panda follower robot, enabling it to execute the same motions. This Learning from Demonstration (LfD) [39] approach involved collecting expert demonstrations for three representative tasks: opening the distribution panel door, performing voltage measurements, and toggling switches.

The dataset was recorded at a high temporal resolution of 30 frames per second (fps). The data collection process creates a structured HDF5 file for each demonstration episode, as illustrated in Fig. 5. Each episode consists of an average of 500 time steps, and contains synchronized streams of proprioceptive data and visual observations. The action space consists of the target joint positions for the 6-DoF manipulator and the gripper state (T, 14). The observation group aggregates real-time sensory inputs, including joint positions (*qpos*), velocities (*qvel*). Crucially, visual data is captured from dual wrist-mounted cameras (*wrist_1*, *wrist_2*) to minimize occlusion during fine manipulation tasks, stored as RGB sequences with a resolution of 640 × 480. This rich dataset enables the ACT model to learn the complex spatio-temporal dependencies required for safe and effective power plant maintenance.

(a) Bimanual Maintenance Robot (b) AGV Robot (c) Retrieval Manipulator

FIGURE 4. Heterogeneous robotic fleet utilized for power plant maintenance tasks: (a) bimanual robot equipped with dual Franka Panda arms for complex manipulation, (b) patrol AGV with Velodyne LiDAR for semantic navigation, and (c) retrieval unit based on a UR3 arm for vision-guided tool acquisition.

```
episode_x
├── action ..................................... (T, 14)
└── observation/
    ├── images/
    │   ├── wrist_1/ ..................... (T, 480, 640, 3)
    │   └── wrist_2/ ..................... (T, 480, 640, 3)
    ├── qpos ..................................... (T, 14)
    └── qvel ..................................... (T, 14)
```

FIGURE 5. Hierarchical structure of the collected HDF5 dataset episode. *T* denotes the episode length, and image dimensions represent height, width, and RGB channels respectively.

2) AGV FOR SEMANTIC NAVIGATION

The AGV robot, shown in Fig. 4(b), serves as the primary mobile agent for autonomous transport and semantic grounding throughout the facility. To ensure reliable localization and high-fidelity spatial awareness, the platform is integrated with a comprehensive sensor suite, including a primary LiDAR sensor and a depth camera. For safety-critical operations, a secondary 2D LiDAR sensor is dedicated to real-time obstacle detection and collision avoidance, providing a redundant safety layer during complex maneuvers. All computational tasks related to SLAM-based autonomous driving and multi-sensor fusion are processed on an onboard dedicated high-performance PC, which ensures low-latency execution of navigational algorithms.

The navigation framework utilizes a graph-based representation of the environment, where operational nodes are meticulously defined based on the generated 3D point cloud. These nodes serve as discrete topological targets that are interconnected to facilitate global path planning across the facility. The system is architected to execute movement commands through ROS (Robot Operating System) topics, which transmit target node sequences directly to the navigation stack for deterministic execution. By integrating this physical navigation layer with the structured environmental representation, the framework enables robust LLM-driven semantic navigation. This allows the AGV to interpret high-level natural language instructions and resolve them into specific graph nodes through CoT reasoning, effectively bridging the gap between human-centric commands and low-level robotic locomotion.

3) VISION-GUIDED RETRIEVAL ROBOT FOR TOOL MANAGEMENT

The retrieval robot, depicted in Fig. 4(c), is responsible for the autonomous identification and acquisition of maintenance instruments from a 3×3 grid storage unit. This agent utilizes a UR3 manipulator integrated with a 2-finger parallel gripper, providing the dexterity necessary for vision-guided drawer operations. The retrieval process relies on a perception system powered by the YOLOv8 object detection model, which is trained to identify labels and tool identifiers on each drawer of the grid. The workflow begins with a natural language request for a specific tool. The system cross-references the detected labels with the structured environmental representation's relational data to determine the exact coordinates of the target drawer. Once the drawer is identified, the robot executes a cartesian linear retraction to open the drawer and retrieve the required item. This integration of YOLOv8-based perception and deterministic kinematic control allows the retrieval robot to manage maintenance logistics effectively, bridging the gap between high-level task planning and physical tool acquisition.

C. EVALUATION METHODOLOGY

This section defines the quantitative metrics and operational scenarios designed to evaluate the hierarchical orchestration framework. By establishing standardized success criteria for each agent and the strategic orchestrator, we ensure a rigorous assessment of the system's reliability in safety-critical power plant environments.

1) EVALUATION CRITERIA FOR BIMANUAL MANIPULATION

The performance of the bimanual maintenance robot is quantified using the Manipulation Success Rate (SR_{manip}), which evaluates the model's ability to execute learned skills on the distribution panel. Each trial is considered successful only if the robot completes the entire action chunk—such as toggling a breaker or rotating a switch—without causing

mechanical collisions or exceeding predefined torque safety limits. The success rate is calculated as:

$$SR_{manip} = \frac{n_{success}}{N} \times 100\% \quad (6)$$

where $n_{success}$ is the number of successfully completed trials and $N = 20$ represents the total attempts per task. This metric accounts for the precision of the Action Chunking with Transformers (ACT) model in replicating the complex spatio-temporal trajectories required for delicate electronic components.

2) EVALUATION CRITERIA FOR SEMANTIC NAVIGATION

The evaluation of the semantic navigation module focuses on Semantic Goal Reachability, a multi-layered metric designed to assess the system's proficiency in translating high-level human intent into functionally viable physical coordinates. Unlike standard navigation benchmarks that emphasize geometric precision, this criteria prioritizes the logical transition from human language to spatial reasoning. The primary metric for evaluating the orchestrator's cognitive performance is the Semantic Grounding Success (SR_{ground}), which quantifies the model's accuracy in mapping natural language commands to the correct topological node G within the structured environmental representation model M. This success rate is formally expressed as:

$$SR_{ground} = \frac{n_{ground}}{N} \times 100\% \quad (7)$$

where n_{ground} represents the number of trials in which the LLM correctly identified the intended semantic target from the environment metadata, and N denotes the total number of evaluation attempts. A grounding failure is recorded if the orchestrator selects an incorrect node or a non-existent spatial identifier, effectively decoupling the model's reasoning errors from the physical navigation performance.In addition to logical correctness, the module is evaluated on Operational Reachability, which ensures that the final pose of the AGV provides an optimal workspace configuration for the subsequent bimanual manipulation phase. A trial is classified as a total success only when the correct semantic node is grounded and the robot's final position (x_{pos}, y_{pos}) falls within the functional boundary required to interface with the distribution panel. This physical threshold is verified using the Euclidean distance error e_{dist} between the target coordinates (x_{goal}, y_{goal}) and the robot's actual stopping point, satisfying the condition:

$$e_{dist} = \sqrt{(x_{goal} - x_{pos})^2 + (y_{goal} - y_{pos})^2} \leq 0.1 \text{ m} \quad (8)$$

By integrating these two distinct layers—cognitive grounding and physical reachability—the evaluation framework provides a holistic view of the system's ability to facilitate a continuous maintenance workflow. Any instance involving a collision, mechanical timeout, or incorrect semantic mapping is recorded as a system failure, ensuring that the overall success rate reflects the deterministic reliability required for safety-critical energy facilities.

3) EVALUATION CRITERIA FOR VISION-GUIDED RETRIEVAL

The evaluation of the vision-guided retrieval module assesses the synergy between a detection-based perception layer and a prismatic manipulation layer within a 3×3 grid storage unit. The performance is quantified by the Retrieval Success Rate (SR_{ret}), which represents the system's reliability in identifying and accessing specific maintenance instruments. A trial is recorded as a total success only if the YOLOv8 model correctly identifies the target drawer label and the UR3 manipulator executes a complete prismatic opening maneuver that provides sufficient clearance for tool access. The success rate is formally defined as:

$$SR_{ret} = \frac{n_{success}}{N} \times 100\% \quad (9)$$

where $n_{success}$ is the number of trials meeting the operational success criteria and $N = 50$ is the total number of attempts. To ensure that the ''opening'' action is functionally meaningful, we define a displacement threshold d_{min}, which represents the minimum distance the drawer must be pulled to expose the tool compartment. The execution is validated by the condition:

$$d_{actual} \geq d_{min} \quad (10)$$

In this context, d_{actual} serves as the quantitative verification that the prismatic control attained the required stroke length for practical maintenance logistics. Failures are categorized into perception-level errors, such as the misidentification of a drawer label, and execution-level errors, including mechanical jamming or the parallel gripper losing its friction-based grasp on the handle. This bifurcated evaluation approach allows for a precise analysis of whether a failure originated from vision-based targeting or low-level kinematic control, ensuring rigorous validation of the retrieval module's operational robustness.

4) EVALUATION CRITERIA FOR HIERARCHICAL TASK PLANNING

To isolate and evaluate the reasoning capabilities of the proposed orchestrator, planning accuracy (SR_{plan}) was utilized as the primary performance metric. This metric assesses the system's ability to translate natural language instructions into logically valid and executable task sequences, independent of the subsequent physical execution. To ensure transparency and reproducibility, a planning trial is recorded as a success only if the generated task sequence strictly satisfies two rigorous conditions:

- **Parameter Validity:** The exact verification that all generated arguments strictly match the predefined entities and capabilities within the structured environmental representation ($\mathcal{M}$). Specifically, the orchestrator must accurately ground semantic concepts into valid system parameters across agent (e.g., assigning a bimanual robot instead of a simple retrieval arm for complex manipulation), entity (e.g., mapping the ''circuit breaker room'' to the exact topological ID *node_43*), and skill

TABLE 1. Classification of command complexity with illustrative examples and expected sequences.

Command Class	Definition	Example Query	Expected Sequence (Ground Truth)
1. Atomic *(Complexity: None)*	Explicit, single or two step instruction for a single agent.	"Go to the Circuit Breaker Room."	`nav_to(agv, node_43)`
2. Coordinated *(Complexity: Low)*	Explicit multi-step instruction requiring agent collaboration.	"Go to Ryu's location and open the drawer that contains the tape measure and the drawer that contains the cutter knife."	`nav_to(agv, node_25)` → `skill_retrieve(ur3, tool_6)` → `skill_retrieve(ur3, tool_4)`
3. Abstract *(Complexity: High)*	Ambiguous intent requiring semantic goal inference.	"Go to place across from circuit breaker room."	`nav_to(agv, node_45)`
4. Strategic *(Complexity: Very High)*	Complex mission combining ambiguity and multi-agent orchestration.	"Go to the room containing the arc flash warning sign and flip the switch handle. Once completed, perform a voltage check on the electrical control panel."	`nav_to(agv, node_43)` → `skill_toggle(bimanual, switch)` → `skill_probe(bimanual, voltage)`

(e.g., selecting the appropriate *skill_toggle* for a switch). Any generation of out-of-vocabulary or physically mismatched parameters is strictly penalized.

- **Semantic Correctness:** The logical verification that the generated sequence of primitive skills accurately reflects the physical and topological constraints required to fulfill the user's ambiguous intent, without omitting necessary intermediate steps.

To ensure absolute objectivity and eliminate the need for inter-rater agreement metrics, the ground-truth templates for all test queries were manually curated and rigorously verified by domain experts. Because the evaluation strictly requires deterministic functional and parametric matches against these ground-truth sequences (as illustrated in Table 1), the grading process yields identical and unambiguous outcomes regardless of whether it is performed programmatically or via manual inspection.

The evaluation methodology is structured around the classification of command complexity outlined in Table 1, which categorizes user queries into four distinct levels of cognitive load. To ensure a comprehensive assessment of the orchestrator's generalization capabilities, we constructed a diverse command set consisting of 60 unique test queries, with 15 distinct commands allocated to each of the four complexity classes.

First, for Atomic and Coordinated commands (Classes 1 and 2), the evaluation focuses on precise keyword mapping and multi-agent assignment. It verifies whether the system correctly maps explicit targets (e.g., 'Ryu's location' or specific tools) to their nodes and sequentially dispatches agents. Conversely, Abstract and Strategic commands (Classes 3 and 4) assess semantic goal inference and long-horizon causal reasoning under ambiguity. The orchestrator must autonomously deduce operational nodes from unstructured cues (e.g., "arc flash warning sign") and logically sequence prerequisite skills (e.g., navigation → toggling) without step-by-step guidance.

Furthermore, to systematically diagnose these limitations and quantify the robustness of the hierarchical framework, any trial that fails to meet the above criteria is classified into one of three distinct failure taxonomies:

1) **Hallucinated Entity:** Generating invalid arguments (i.e., *fabricated entities or parameters*), such as non-existent tools, unregistered agents, or unmapped locations.
2) **Topological Violation:** Attempting navigation between unconnected nodes or ignoring the physical spatial hierarchy defined in $\mathcal{M}$ (i.e., *wrong relations*).
3) **Causal Sequence Error:** Violating strict operational rules (i.e., *policy violations*, such as failing to allocate an appropriate agent capable of executing the required skill). Crucially, this also includes degenerate planning loops, where the orchestrator repeatedly visits the same location or redundantly executes the same valid action in a failed attempt to substitute an unresolvable intent.

TABLE 2. Success rate of ACT-based bimanual manipulation for electrical distribution panel tasks.

Task Category	Number of Trials	Success Rate
Open Panel Door	20	**100%**
Voltage Measurement	20	90%
Switch-off Operation	20	80%
Total Average	60	**90%**

VI. RESULTS AND DISCUSSION

A. EVALUATION 1: BIMANUAL MANIPULATION VIA ACT

The autonomous performance of the bimanual maintenance robot was rigorously evaluated across three representative operational tasks within the distribution panel mock-up. A total of 20 trials were conducted—20 for each task—to determine the reliability of the ACT model in replicating expert skills. The quantitative results of these experiments are summarized in Table 2.

The experimental results indicate that the opening panel door task achieved a success rate of 100%. Fig. 6(a) demonstrates the bimanual robot agent executing the door-opening process. This high reliability is attributed to the relatively large operational workspace and the model's

(a) Bimanual manipulation for opening the distribution panel door.

(b) Precision voltage measurement using coordinated bimanual control.

(c) Single-arm operation for switch control.

FIGURE 6. Autonomous maintenance tasks performed by the bimanual robot using the ACT model. The framework demonstrates (a) high-torque door manipulation, (b) precision contact for voltage sensing, and (c) task-specific single-arm dexterity for switch operations.

ability to precisely replicate the high-dimensional trajectories gathered during Gello-based teleoperation.

In contrast, the Voltage Measurement task yielded a 90% success rate. Fig. 6(b) illustrates the bimanual agent executing the voltage measurement task. Unlike the door-opening task, this operation is significantly more challenging due to the need for high-precision control to contact the thin probe tips with the target points. Analysis of the failure cases revealed that most errors stemmed from microscopic contact misalignments between the voltage probes and the measurement points, which occasionally failed to trigger a successful sensor reading despite the arms reaching the target coordinates.

The Switch-off Operation recorded the lowest success rate at 80%. As illustrated in Fig. 6(c), this unimanual task involves overcoming significant mechanical resistance and torque, which occasionally caused grasp slips or incomplete rotations. However, the ACT policy demonstrated robustness through its visual feedback loop. By continuously monitoring the switch's state via RGB observations, the agent autonomously recognized incomplete actuations and persisted with corrective actions until the switch was successfully deactivated. Despite these physical challenges, the cumulative average success rate of 90% confirms that the proposed bimanual system is capable of effectively replacing human operators in high-risk, high-voltage environments.

B. EVALUATION 2: SEMANTIC GOAL REACHABILITY

The experimental evaluation of the semantic navigation module focused on the agent's ability to seamlessly bridge the gap between high-level linguistic intent and physical spatial constraints. Over 100 randomized trials, the system demonstrated exceptional reliability, achieving robust success rates in both cognitive and physical domains as summarized in Table 3. The Semantic Grounding Success (SR_{ground}) recorded a rate of 95%, validating that the strategic orchestrator, enhanced by Chain-of-Thought prompting, could effectively resolve ambiguous maintenance instructions into the correct topological nodes within the structured environmental representation. This high-fidelity grounding performance indicates that the hierarchical constraints significantly mitigate the potential for LLM hallucinations by restricting the reasoning space to the validated environmental metadata.

TABLE 3. Performance of semantic goal reachability.

Evaluation Criterion	Number of Trials	Success Rate
Semantic Grounding	100	**95%**
Operational Reachability	100	**98%**

Furthermore, the system achieved a 98% Operational Reachability rate, with the AGV successfully navigating to grounded coordinates in the vast majority of instances. Analysis of locomotion data showed that the Euclidean distance error e_{dist} was maintained significantly below the 0.1 m threshold, ensuring a precise starting configuration for subsequent manipulation tasks. This seamless integration between high-level LLM-based reasoning and low-level SLAM-based navigation proves that the proposed framework is not only intelligent in its interpretation of human commands but also deterministic in its physical execution. The

(a) Real-time object detection from the egocentric gripper camera.

(b) Third-person view of the successful prismatic drawer opening.

FIGURE 7. Qualitative results of the vision-guided retrieval experiment. (a) The YOLOv8 model achieves robust bounding box regression even with the camera in motion. (b) Based on the precise target localization, the UR3 manipulator executes a stable grasp and successfully pulls the drawer to the required extension.

TABLE 4. Performance of the YOLOv8-based label recognition model.

Class	Precision	Recall	mAP@50	mAP@50-95
Drawer Label	0.968	0.957	0.981	0.843

TABLE 5. Performance metrics for vision-guided tool retrieval.

Task Category	Trials	Successes	Success Rate
Drawer Retrieval	50	46	**92%**

results confirm that the use of a graph-based representation derived from 3D point cloud data provides the structured environment necessary for LLMs to function as reliable planners in safety-critical energy facilities.

C. EVALUATION 3: VISION-GUIDED GRID STORAGE RETRIEVAL

The operational effectiveness of the vision-guided retrieval system was validated through 50 autonomous trials within the 3×3 grid storage unit. This evaluation tested the integrated performance of the YOLOv8-based perception layer and the cartesian linear retraction control for drawer manipulation. The quantitative results of these experiments are summarized in Table 4 and Table 5.

Prior to physical manipulation, the reliability of the perception layer was assessed. As detailed in Table 4, the YOLOv8 model demonstrated robust detection capabilities, achieving a precision of 0.968 and a recall of 0.957. Notably, the model maintained a high mean Average Precision (mAP) of 0.981 at an Intersection over Union (IoU) threshold of 0.5. More critically for robotic grasping, the mAP@50-95 score of 0.843 indicates superior localization accuracy. This high geometric fidelity ensures that the predicted bounding boxes are tightly aligned with the physical drawer labels, minimizing coordinate errors during the hand-eye calibration process and enabling precise end-effector alignment.

Building on this accurate perception, the physical retrieval module achieved a success rate of 92% across 50 trials, as shown in Table 5. The system exhibited exceptional robustness in identifying target drawer handles even under varying ambient lighting conditions. The qualitative performance is further illustrated in Fig. 7. As demonstrated in the egocentric view (Fig. 7(a)), the perception layer maintained stable detection despite camera motion. Furthermore, Fig. 7(b) captures the physical execution phase, where the manipulator executed a Cartesian linear retraction to match the drawer's sliding axis, ensuring full extension beyond the minimum displacement threshold (d_{min}) in successful trials.

Analysis of the four recorded failures revealed that the errors were primarily concentrated in the mechanical execution phase. Specifically, failures occurred when the drawer's mechanical resistance exceeded the gripper's friction threshold, causing slippage. Crucially, the current framework employs a decoupled "perceive-then-act" architecture, where the manipulation trajectory is generated based on the initial state estimation from YOLOv8 rather than continuous visual servoing. Consequently, the system lacked the real-time feedback loop required to detect and compensate for this slippage during the pulling motion. In these instances, the displacement d_{actual} failed to meet the d_{min} threshold. Despite this limitation, the modular approach proved highly effective for structured tasks.

D. EVALUATION 4: HIERARCHICAL TASK PLANNING

1) BASELINE SETUP: UNSTRUCTURED VS. STRUCTURED GROUNDING

To rigorously validate the efficacy of the proposed Structured Environmental Representation ($\mathcal{M}$), we established a control baseline denoted as Plain Text. Crucially, to ensure a fair comparison of reasoning capabilities, the Plain Text baseline was constructed to contain the **exact same informational content** as $\mathcal{M}$.

While $\mathcal{M}$ organizes facility data into a rigid graph topology (e.g., Node IDs, adjacency lists, and entity attributes), the Plain Text baseline presents this identical data as a continuous

narrative description (e.g., ''The Electrical Room (Node 45) is located across from the Circuit Breaker Room and contains a control panel...''). This experimental design isolates the impact of **data structure** on the LLM's planning performance, eliminating information asymmetry as a confounding variable. Both formats were fed into the LLMs' context window to assess whether the structural constraints of $\mathcal{M}$ effectively reduce search space and hallucination compared to the unstructured prose.

The core contribution of this research is the orchestration of a heterogeneous robotic fleet through a hierarchical task planning layer grounded in a structured knowledge base. To validate the efficacy of the proposed **Structured Environmental Representation** ($\mathcal{M}$), a comprehensive ablation study was conducted comparing it against a baseline Plain Text grounding approach. The evaluation utilized four local LLMs: Llama3 (8B) [40], Gemma2 (9B) [41], Qwen 2.5 (14B) [42], and GPT-OSS (20B) [43]. The quantitative results across four command complexity categories—Atomic, Coordinated, Abstract, and Strategic—are summarized in Table 6.

2) IMPACT OF STRUCTURED ENVIRONMENTAL REPRESENTATION

The most significant finding is that the structured grounding source ($\mathcal{M}$) functions as a critical cognitive scaffold, substantially enhancing the reasoning reliability of larger models compared to unstructured Plain Text. As evidenced in Table 6, the **GPT-OSS (20B)** model achieved a peak mission success rate of **91.7%** when utilizing $\mathcal{M}$, marking a notable improvement over the 85.0% baseline with Plain Text.

This performance leap in the **Strategic** category (53.3% $\rightarrow$ **73.3%**) confirms that the explicit definition of entity E and relation R attributes within $\mathcal{M}$ is crucial for high-level planning. Unlike unstructured text, where asset availability and spatial connectivity are buried in narrative descriptions, the structured node topology allows the orchestrator to rigorously link required tools E with their target locations via defined spatial edges R. This structural clarity enables the planner to generate valid multi-hop sequences—such as retrieving a 'voltage tester' from a specific storage node before navigating to a connected electrical room—thereby ensuring functionally viable execution plans that implicit text processing often fails to construct.

Similarly, **Qwen 2.5 (14B)** exhibited a clear benefit from the structured representation, improving its Strategic score from 40.0% to 53.3% and its Abstract command understanding from 86.7% to **93.3%**. This suggests that $\mathcal{M}$ effectively reduces the search space for the LLM, allowing mid-to-large scale models to disambiguate vague instructions (e.g., ''Go to the breaker room'') with higher precision than when processing verbose textual descriptions.

3) MODEL SCALABILITY AND SCHEMA ADHERENCE

The evaluation also highlighted a correlation between model parameter scale and the ability to leverage structured data. While the larger models (14B, 20B) successfully utilized $\mathcal{M}$ to boost performance, the smaller-scale **Llama3 (8B)** showed no performance variation (58.3%) between the two grounding sources. This indicates that despite the availability of structured constraints, smaller models struggle with the cognitive load required to parse complex JSON schemas alongside long-horizon planning logic.

Gemma2 (9B) demonstrated a marginal improvement (70.0% $\rightarrow$ 71.7%), showing potential in Abstract tasks (73.3%) but remaining limited in Strategic scenarios (33.3%). Conversely, GPT-OSS (20B) not only achieved the highest overall average but also maintained a robust 100% success rate in Atomic and Coordinated tasks regardless of the grounding source, proving its foundational stability.

4) QUANTITATIVE ANALYSIS OF HALLUCINATION REDUCTION

To empirically validate the claim that the Structured Environmental Representation ($\mathcal{M}$) mitigates reasoning hallucinations, we conducted a quantitative analysis of the unsuccessful planning trials. To ensure statistical significance, we aggregated the failure logs of the two high-capacity models capable of leveraging the schema constraints: Qwen 2.5 (14B) and GPT-OSS (20B). The error distribution across these 120 combined test queries is summarized in Table 7.

As shown in Table 7, the unstructured text baseline suffered evenly across all failure types. In contrast, integrating the structured constraints of $\mathcal{M}$ yielded a consistent reduction across every dimension of planning hallucination, decreasing the overall error rate from 16.7% (20 errors) to 10.8% (13 errors).

Specifically, the rate of *Hallucinated Entities* (i.e., generating fabricated tools, unregistered nodes, or invalid parameters) decreased from 5.8% to 4.2%. This indicates that the explicit schema serves as an effective vocabulary boundary, although the models still occasionally misalign parameters when attempting to resolve highly ambiguous user intents.

Furthermore, the explicit definition of spatial and functional attributes in $\mathcal{M}$ noticeably improved logical flow. The per-type rate of *Topological Violations* (wrong relations, such as navigating unconnected paths) was reduced from 5.8% to 3.3%, demonstrating improved spatial awareness. Similarly, *Causal Sequence Errors* decreased from 5.0% to 3.3%. While the structured policies successfully prevented the skipping of prerequisite safety steps in most cases, the detailed logs revealed an interesting behavioral characteristic: the remaining causal errors often manifested as degenerate planning loops. When the orchestrator failed to resolve a complex strategic intent, rather than fabricating out-of-schema parameters, it occasionally defaulted to redundantly repeating valid actions or repeatedly navigating back to the same node. Consequently, while $\mathcal{M}$ effectively suppresses generalized hallucinations and structural violations, resolving these persistent logical loops remains a key challenge for purely open-loop LLM planners.

TABLE 6. Comparison of task planning success rates: plain text vs. structured environmental representation ($\mathcal{M}$).

Model	Grounding Source	Atomic	Coordinated	Abstract	Strategic	Average
Llama3 (8B)	Plain Text	53.3%	100%	40.0%	40.0%	58.3%
	Ours ($\mathcal{M}$)	53.3%	100%	40.0%	40.0%	58.3%
Gemma2 (9B)	Plain Text	80.0%	100%	66.7%	33.3%	70.0%
	Ours ($\mathcal{M}$)	80.0%	100%	73.3%	33.3%	71.7%
Qwen 2.5 (14B)	Plain Text	100%	100%	86.7%	40.0%	81.7%
	Ours ($\mathcal{M}$)	100%	100%	**93.3%**	53.3%	86.7%
GPT-OSS (20B)	Plain Text	100%	100%	86.7%	53.3%	85.0%
	Ours ($\mathcal{M}$)	100%	100%	**93.3%**	**73.3%**	**91.7%**

TABLE 7. Quantification and per-type rates of hallucination in failed planning trials (aggregated for qwen 2.5 and GPT-OSS, total 120 queries).

Failure Taxonomy	Plain Text	Proposed ($\mathcal{M}$)
1. Hallucinated Entity *- Fabricated entities or parameters*	7 (5.8%)	**5 (4.2%)**
2. Topological Violation *- Wrong relations*	7 (5.8%)	**4 (3.3%)**
3. Causal Sequence Error *- Policy violations or logical loops*	6 (5.0%)	**4 (3.3%)**
Total Errors (Overall Error Rate)	20 (16.7%)	**13 (10.8%)**

5) DISCUSSION

The experiments validate that reliability in safety-critical O&M tasks is not achieved by model size alone, but by the synergy between **sufficient reasoning capacity (20B+)** and **deterministic environmental grounding** ($\mathcal{M}$). The proposed framework effectively bridges the gap between probabilistic LLM outputs and physical execution constraints, ensuring that high-level intents are translated into logically sound and safe robotic actions.

6) ROBUSTNESS IN DYNAMIC ENVIRONMENTS AND LIMITATIONS

While the proposed structured environmental representation ($\mathcal{M}$) significantly enhances the reasoning reliability of the LLM orchestrator, its reliance on a static prior map introduces challenges regarding robustness to missing or incorrect metadata.

In terms of dynamic facility changes (e.g., temporary obstacles, slight misalignments of target objects), our framework achieves local robustness by delegating these disturbances to the ''Skill-Based Execution Layer.'' Even if the exact spatial metadata is slightly inaccurate, the lower-level execution modules compensate for these dynamic changes. For example, the AGV utilizes LiDAR-based local planners for dynamic obstacle avoidance, and the retrieval manipulator utilizes real-time YOLOv8 bounding box regression to dynamically grasp tools regardless of minor positional shifts. Furthermore, the ACT-based LfD model employed for bimanual manipulation processes continuous visual feedback, enabling real-time verification of the electrical distribution panel's status and adapting to slight structural variations during physical interaction.

However, the framework remains highly vulnerable to severe metadata inaccuracies. If the prior JSON schema contains fundamentally false information (e.g., a designated tool is completely missing from its storage node), the LLM orchestrator, operating in an open-loop cognitive manner, will successfully generate a logically valid plan that ultimately fails during physical execution.

Addressing this discrepancy between the static cognitive map and the dynamic physical world requires a closed-loop semantic feedback mechanism. In future work, we plan to integrate lightweight Vision-Language Models (VLMs) on the patrol AGV. By leveraging the AGV's egocentric vision, the system could autonomously detect semantic discrepancies and dynamically update the JSON schema $\mathcal{M}$ in real-time, ultimately bridging the gap between static orchestration and highly dynamic, real-world maintenance operations.

VII. CONCLUSION

This study proposed and validated a hierarchical multi-robot orchestration framework tailored for the demanding requirements of safety-critical maintenance in energy facilities. By strategically decoupling high-level linguistic reasoning from low-level physical execution, the framework addresses the dual challenges of operational reliability and cognitive flexibility. Central to this approach is the Structured Environmental Representation ($\mathcal{M}$), which bridges the gap between probabilistic LLM outputs and deterministic robotic skills by explicitly mapping semantic intent to physical entities and topological relations.

The experimental results obtained from a high-fidelity power plant testbed underscore the efficacy of this synergistic approach. A comprehensive ablation study identified GPT-OSS (20B) as the optimal orchestrator, achieving a peak planning success rate of 91.7% when grounded in $\mathcal{M}$. Crucially, the evaluation revealed that performance in complex strategic tasks is driven not by model scale alone, but by the interplay between sufficient reasoning capacity and structured environmental grounding. The use of $\mathcal{M}$ enabled the 20B model to resolve long-horizon

causal dependencies—such as multi-hop navigation and tool retrieval—achieving a 73.3% success rate and significantly outperforming the unstructured plain text baseline.

On the physical execution front, the framework demonstrated robust adaptability. The bimanual agent, integrated with ACT, achieved a 90% success rate in handling variable distribution panel tasks, validating the system's capability to translate abstract plans into precise physical actions. In conclusion, this research provides a proven pathway for deploying autonomous systems in industrial environments where safety is non-negotiable. Future work will investigate the scalability of this framework for multi-agent swarm coordination and the implementation of adaptive learning strategies to handle unexpected hardware degradations. Furthermore, to address the inherent limitations of relying on static prior maps, we aim to integrate Vision-Language Models to establish a closed-loop semantic feedback mechanism. This will enable the orchestrator to dynamically update the structured representation ($\mathcal{M}$) in real-time, ensuring robust adaptation to unmapped environmental changes and metadata inaccuracies in real-world power plants.

REFERENCES

[1] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023.

[2] J.-W. Choi, Y. Yoon, H. Ong, J. Kim, and M. Jang, "LoTa-bench: Benchmarking language-oriented task planners for embodied agents," 2024, *arXiv:2402.08178*.

[3] A. O'Neill et al., "Open X-embodiment: Robotic learning datasets and RT-X models," 2023, *arXiv:2310.08864*.

[4] B. P. Gerkey and M. J. Matarić, "A formal analysis and taxonomy of task allocation in multi-robot systems," *Int. J. Robot. Res.*, vol. 23, no. 9, pp. 939–954, Sep. 2004.

[5] M. B. Dias, R. Zlot, N. Kalra, and A. Stentz, "Market-based multirobot coordination: A survey and analysis," *Proc. IEEE*, vol. 94, no. 7, pp. 1257–1270, Jul. 2006.

[6] S. Sariel, T. Balch, and N. Erdogan, "Incremental multi-robot task selection for resource constrained and dynamic environments," *Auto. Robots*, vol. 30, no. 2, pp. 163–188, 2011.

[7] L. E. Parker, "Alliance: An architecture for fault-tolerant multirobot control," *IEEE Trans. Robot. Autom.*, vol. 14, no. 2, pp. 220–240, Feb. 1998.

[8] G. A. Korsah, A. Stentz, and M. B. Dias, "A comprehensive taxonomy for multi-robot task allocation," *Int. J. Robot. Res.*, vol. 32, no. 12, pp. 1495–1512, Oct. 2013.

[9] M. Gombolay, R. Wilcox, and J. Shah, "Fast scheduling of multi-robot teams with temporospatial constraints," in *Proc. Robot., Sci. Syst.*, 2013, pp. 1–8, doi: 10.15607/RSS.2013.IX.049.

[10] A. Farinelli, L. Iocchi, and D. Nardi, "Multirobot systems: A classification focused on coordination," *IEEE Trans. Syst., Man Cybern., B, Cybern.*, vol. 34, no. 5, pp. 2015–2028, Oct. 2004.

[11] C. R. Kube and H. Zhang, "Collective robotics: From social insects to robots," *Adapt. Behav.*, vol. 2, no. 2, pp. 189–218, Sep. 1993.

[12] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: Algorithms and theory," *IEEE Trans. Autom. Control*, vol. 51, no. 3, pp. 401–420, Mar. 2006.

[13] L. Vig and J. A. Adams, "Multi-robot coalition formation," *IEEE Trans. Robot.*, vol. 22, no. 4, pp. 637–649, Apr. 2006.

[14] H. Hu, X. Liao, W. Du, and F. Qian, "Multi-robot connection towards collective obstacle field traversal," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Aug. 2024, pp. 1–7.

[15] L. P. Kaelbling and T. Lozano-Pérez, "Hierarchical task and motion planning in the now," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 1470–1477.

[16] M. Ahn et al., "Do as i can, not as i say: Grounding language in robotic affordances," 2022, *arXiv:2204.01691*.

[17] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson,I. Mordatch, Y. Chebotar, P. Sermanet, T. Jackson, N. Brown, L. Luu, S. Levine, K. Hausman, and B. Ichter, "Inner monologue: Embodied reasoning through planning with language models," in *Proc. Conf. Robot Learn. (CoRL)*, 2022, pp. 1769–1782.

[18] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Programmatic prompt generation for harmonious robot control," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 11523–11530.

[19] A. Rajvanshi, K. Sikka, X. Lin, B. Lee, H.-P. Chiu, and A. Velasquez, "SayNav: Grounding large language models for dynamic planning to navigation in new environments," 2023, *arXiv:2309.04077*.

[20] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and manipulation," in *Proc. AAAI Conf. Artif. Intell.*, 2011, pp. 1507–1514.

[21] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 9493–9500.

[22] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen, "Multi-agent collaboration mechanisms: A survey of LLMs," 2025, *arXiv:2501.06322*.

[23] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," 2023, *arXiv:2305.16291*.

[24] A. Brohan et al., "RT-2: Vision-language-action models transferred to real-world robotics," 2023, *arXiv:2307.15818*.

[25] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An open-source vision-language-action model," 2024, *arXiv:2406.09246*.

[26] NVIDIA et al., "GR00T N1: An open foundation model for generalist humanoid robots," 2025, *arXiv:2503.14734*

[27] S. Zhang, Z. Xu, P. Liu, X. Yu, Y. Li, Q. Gao, Z. Fei, Z. Yin, Z. Wu, Y.-G. Jiang, and X. Qiu, "VLABench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025.

[28] C. Canali, A. Pistone, P. Guardiani, D. Ludovico, S. Leggieri, C. Gloriani, and D. G. Caldwell, "Inspection robotics for harsh environments, industrial applications and infrastructures," in *Proc. I-RIM 3D*, Rome, Italy, Dec. 2020, pp. 1–3.

[29] D. Hyun, I. Kim, S. Joo, J. Ha, and J. Lee, "Remote dismantling system using a digital manufacturing system and workpiece localization for nuclear facility decommissioning," *Ann. Nucl. Energy*, vol. 195, Jan. 2024, Art. no. 110182.

[30] X. Qin, B. Jia, J. Lei, J. Zhang, H. Li, B. Li, and Z. Li, "A novel flying–walking power line inspection robot and stability analysis hanging on the line under wind loads," *Mech. Sci.*, vol. 13, no. 1, pp. 257–273, Mar. 2022.

[31] J. Katrasnik, F. Pernus, and B. Likar, "A survey of mobile robots for distribution power line inspection," *IEEE Trans. Power Del.*, vol. 25, no. 1, pp. 485–493, Jan. 2010.

[32] D. Gitardi, M. Giardini, and A. Valente, "Autonomous robotic platform for inspection and repairing operations in harsh environments," *Int. J. Comput. Integr. Manuf.*, vol. 34, no. 6, pp. 666–684, Jun. 2021.

[33] D. Lattanzi and G. Miller, "Review of robotic infrastructure inspection systems," *J. Infrastructure Syst.*, vol. 23, no. 3, Sep. 2017, Art. no. 04017004.

[34] P. H. C. Morais, K. C. T. Vivaldini, E. R. R. Kato, and R. S. Inoue, "A review of robot fleet management," *IEEE Access*, vol. 13, pp. 1–22, 2025.

[35] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2022, *arXiv:2201.11903*.

[36] T. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proc. Robot., Sci. Syst. (RSS)*, 2023, pp. 1–19, doi: 10.15607/RSS.2023.XIX.016.

[37] R. Varghese and M. Sambath, "YOLOv8: A novel object detection algorithm with enhanced performance and robustness," in *Proc. Int. Conf. Adv. Data Eng. Intell. Comput. Syst. (ADICS)*, Apr. 2024, pp. 1–6.

[38] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "GELLO: A general, low-cost, and intuitive teleoperation framework for robot manipulators," 2023, *arXiv:2309.13037*.

[39] C. G. Atkeson and S. Schaal, "Robot learning from demonstration," in *Proc. 14th Int. Conf. Mach. Learn. (ICML)*, 1997, pp. 12–20.
[40] A. Grattafiori et al., "The llama 3 herd of models," 2024, *arXiv:2407.21783*.
[41] G. Team et al., "Gemma 2: Improving open language models at a practical size," 2024, *arXiv:2408.00118*.
[42] B. Hui et al., "Qwen2.5 technical report," 2024, *arXiv:2412.15115*.
[43] S. Agarwal et al., "GPT-OSS-120b & GPT-OSS-20b model card," 2025, *arXiv:2508.10925*.

JUNGI LEE received the B.S. and M.S. degrees in electrical engineering and computer science (EECS) from Gwangju Institute of Science and Technology (GIST) College, South Korea, in 2014 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the Department of AI Convergence. He is also a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI), South Korea. His research interests include embodied AI for robotics, multi-robot orchestration, and large language models.

SEU-JAN KIM received the B.S. degree in business administration and the M.S. degree in data science from Chonnam National University, South Korea, in 2021 and 2025, respectively. She is currently a Research Associate with the Electronics and Telecommunications Research Institute (ETRI), South Korea. Her research interests include embodied AI for robotics, multimodal learning, and large language models.

GEONHYUP LEE received the B.S. degree in mechanical engineering from Pukyong National University, South Korea, in 2019, and the M.S. degree from Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2021, where he is currently pursuing the Ph.D. degree with the Department of AI Convergence. His research focuses on robotic manipulation for assembly and the development of foundational models for general manipulation. His research interests include multimodal learning for robotics, particularly the integration of vision and force information for contact-rich tasks.

KANGMIN KIM (Student Member, IEEE) received the B.S. degree in mechanical engineering and the M.S. degree from Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2022 and 2024, respectively, where he is currently pursuing the Ph.D. degree in AI convergence. His research interests include robotic intelligence, humanoid manipulation and control, physical AI, vision-language-action (VLA) models, and policy learning.

JIMIN JEON is currently pursuing the B.S. degree in electrical engineering and computer science (EECS) with Gwangju Institute of Science and Technology (GIST), South Korea. He is an Undergraduate Research Intern with the GIST AI Laboratory, South Korea. His research interests include robot learning, particularly imitation learning and video-to-robot manipulation.

SEOK-KAP KO received the B.S., M.S., and Ph.D. degrees in information telecommunication engineering from Soongsil University, South Korea, in 1997, 2002, and 2009, respectively. Since 2008, he has been a Principal Research Engineer with the Electronics and Telecommunications Research Institute (ETRI), South Korea. His research interests include machine learning for energy management systems.

KYOOBIN LEE (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mechanical engineering from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2008. From 2008 to 2010, he was a Postdoctoral Scholar with the Center for Neuroscience, KIST. From 2012 to 2017, he was a Principal Researcher with the Samsung Advanced Institute of Technology. Since 2017, he has been a Professor with the Department of AI Convergence, Gwangju Institute of Science and Technology (GIST). His research interests include vision recognition using deep learning, robot control applications using computer vision, and split learning for neural networks of cloud computing applications. He is an Editor of Korea Robotics Society Review.