ODYSSEY: Open-World Quadrupeds Exploration and Manipulation for Long-Horizon Tasks

Kaijun Wang[1*], Liqin Lu[2*], Mingyu Liu[1], Jianuo Jiang[3], Zeju Li[1], Bolin Zhang[1], Wancai Zheng[2], Xinyi Yu[2†], Hao Chen[1†], Chunhua Shen[1,4†]

[1]Zhejiang University
[2]Zhejiang University of Technology
[3]The Chinese University of Hong Kong, Shenzhen
[4]Ant Group

Abstract

Language-guided long-horizon mobile manipulation has long been a grand challenge in embodied semantic reasoning, generalizable manipulation, and adaptive locomotion. Three fundamental limitations hinder progress: First, although large language models have shown promise in enhancing spatial reasoning and task planning through learned semantic priors, existing implementations remain confined to tabletop scenarios, failing to address the constrained perception and limited actuation ranges characteristic of mobile platforms. Second, current manipulation strategies exhibit insufficient generalization when confronted with the diverse object configurations encountered in open-world environments. Third, while crucial for practical deployment, the dual requirement of maintaining high platform maneuverability alongside precise end-effector control in unstructured settings remains understudied in the literature.

In this work, we present ODYSSEY, a unified mobile manipulation framework for agile quadruped robots equipped with manipulators, which seamlessly integrates high-level task planning with low-level whole-body control. To address the challenge of egocentric perception in language-conditioned tasks, we introduce a hierarchical planner powered by a vision-language model, enabling long-horizon instruction decomposition and precise action execution. At the control level, our novel whole-body policy achieves robust coordination of locomotion and manipulation across challenging terrains. We further present the first comprehensive benchmark for long-horizon mobile manipulation, evaluating diverse indoor and outdoor scenarios. Through successful sim-to-real transfer, we demonstrate the system's generalization and robustness in real-world deployments, underscoring the practicality of legged manipulators in unstructured environments. Our work advances the feasibility of generalized robotic assistants capable of complex, dynamic tasks.

1 Introduction

Open-world mobile manipulation enables robots to autonomously navigate and interact in dynamic, unstructured environments by integrating mobility, manipulation, and perception. Unlike traditional pipelines that decouple navigation and manipulation, this unified design fosters emergent behaviors such as active perception—e.g., adjusting pose for a better grasp—crucial for real-world robustness.

Previous studies have achieved strong results in navigation (Grandia et al. 2023; Zhuang et al. 2023; Liu et al. 2025) and manipulation (Kim et al. 2024; Brohan et al. 2023; Cheang et al. 2024), but recent whole-body frameworks (Pan et al. 2025a; Fu, Cheng, and Pathak 2023; Liu et al. 2024; Zhang et al. 2025; Wang et al. 2025) still struggle with scalability due to simplified settings and short-horizon evaluations. We introduce ODYSSEY, a reinforcement learning–based whole-body control system unifying quadruped locomotion and precise manipulation through an integrated vision-language framework. ODYSSEY achieves state-of-the-art accuracy under out-of-distribution conditions and diverse terrains for real-world deployment.

Recent works (Qi et al. 2025; Pan et al. 2025b) show that large language models enhance robotic task planning via spatial reasoning. We extend their use to whole-body navigation and manipulation, grounding execution at two levels: (1) task-level semantic planning and (2) fine-grained guidance through geometry-constrained pose estimation.

To close the evaluation gap, we propose the first long-horizon mobile manipulation benchmark with eight diverse daily tasks across indoor and outdoor environments. It holistically assesses reasoning, planning, navigation, and manipulation, incorporating the standardized Arnold framework for precise evaluation. Through extensive experiments, our system shows strong ability for sim2real transfer, showcasing exceptional generalization where both control and planning modules maintain consistent performance across diverse real-world scenarios. Our contributions are four folds:

(i) We introduce a hierarchical vision-language planner that bridges egocentric perception and language-conditioned tasks, decomposing long-horizon instructions into executable actions.

(ii) We propose the first whole-body control policy that generalizes to challenging terrains while jointly coordinating locomobility and manipulation.

(iii) We introduce the first long-horizon mobile manipulation benchmark, covering a wide range of realistic indoor

*These authors contributed equally.
†Corresponding authors.

Figure 1: We present ODYSSEY, a unified mobile manipulation framework for agile quadruped robots equipped with manipulators, which seamlessly integrate high-level tasks planning with low-level whole-body control.

and outdoor scenarios.

(iv) We further demonstrate successful sim-to-real transfer of both high-level planners and low-level control policies, showing strong generalization and robustness in real-world deployments.

Our results highlight the feasibility and practicality of deploying a legged mobile manipulator in unstructured environments, paving the way toward generalized robotic assistants.

2 Related Work

2.1 Open-world mobile manipulation

Previous research has made significant advances in both navigation and manipulation as separate domains, with robust solutions developed for mobile robot path planning in dynamic environments (Grandia et al. 2023; Zhuang et al. 2023) and sophisticated manipulation techniques for object interaction in controlled settings (Kim et al. 2024; Brohan et al. 2023; Cheang et al. 2024). While pioneering works (Pan et al. 2025a; Fu, Cheng, and Pathak 2023; Liu et al. 2024; Zhang et al. 2025; Wang et al. 2025; Fu, Zhao, and Finn 2024; Jiang et al. 2025b) have developed initial whole-body control frameworks, they face scalability limitations in open-world scenarios due to oversimplified environmental assumptions and evaluations limited to short-horizon pick-and-place tasks. Some works (Ha et al. 2024; Qiu et al. 2025b) have attempted to attach more complex interactions by learning per-action policies from human demonstrations. However, these methods lack compositionality and scalability, needing task-specific data for each scenario. ODYSSEY overcomes these limitations by unifying terrain-aware locomotion with hierarchical planning, enabling robust mobile manipulation in unstructured environments.

2.2 Foundation Models for Embodied Tasks

Vision-language models (VLMs) have shown promise in enhancing robotic reasoning (Qi et al. 2025; Pan et al. 2025b; Zhi et al. 2025; Qiu et al. 2025a; Yating Wang 2025), but their evaluation has been restricted to tabletop settings with fixed cameras. For navigation, foundation models improve spatial understanding (Gu et al. 2024; Jatavallabhula et al. 2023; Jiang et al. 2025a), yet they lack fine-grained manipulation support. ODYSSEY advances this by grounding hierarchical planning in egocentric perception, using VLMs to decompose tasks via scene graphs while generating precise end-effector trajectories. This contrasts with modular approaches (Zhang et al. 2025) that struggle with compositional reasoning under uncertainty.

2.3 Benchmarks for Real-World Deployment

Existing benchmarks for mobile manipulation (Qiu et al. 2025a) focus narrowly on navigation or short-term interactions, lacking standardized metrics for long-horizon tasks. Simulation frameworks like IsaacSim have enabled progress in locomotion (Zhi et al. 2025), but manipulation-centric evaluations remain sparse. ODYSSEY introduces a comprehensive benchmark with diverse indoor/outdoor scenarios, addressing multi-stage reasoning, object-aware navigation, and precision manipulation. This complements prior work (Qiu et al. 2025b) by enabling scalable testing of sim-to-real transfer for integrated control and planning.

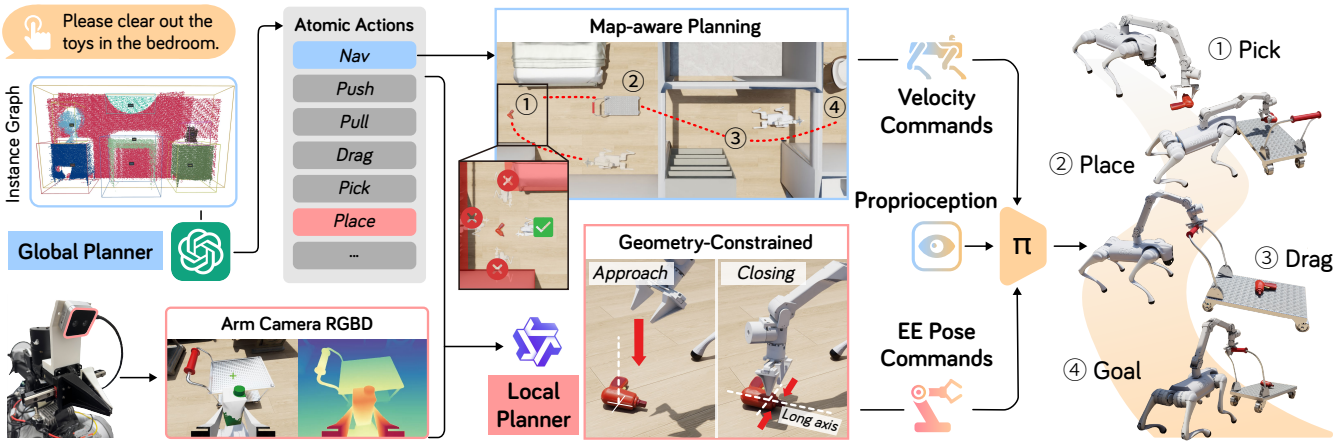

Figure 2: ODYSSEY pipeline spans the entire process of a long-horizon task, including multi-modal semantic perception, map-aware global planning, geometry-constrained action grounding, and step-wise execution by a learned low-level policy.

3 Method

In this section, we present ODYSSEY, a unified framework that spans long-horizon task planning, whole-body control, and standardized evaluation for mobile manipulation. It comprises three key components:

1. **Coarse-to-Fine Task Planner** (Section 3.1): a hierarchical planner that orchestrates top-down task execution under the guidance of foundation models.
2. **Quadruped Whole-Body Policy** (Section 3.2): a reinforcement learning-based whole-body controller that generalizes to diverse terrains and overcomes the sim-to-real gap.
3. **Mobile Manipulation Benchmark** (Section 3.3): the first scalable evaluation suite for assessing long-term task performance across versatile real-world scenarios.

3.1 Long-horizon Task Planner

To bridge the gap left by prior work in modeling the intricate dependencies between semantic reasoning-based navigation and fine-grained, generalizable manipulation, our hierarchical framework is explicitly designed to ensure the reliability of both components while reinforcing their mutual dependencies for coherent long-horizon task execution.

Map-Aware Task-Level Planning To support long-horizon task planning grounded in egocentric observations, we first build a global planner that integrates a lightweight multi-modal perception module as a plug-in component. Concretely, we fuse on-board RGB and LiDAR streams to form a unified spatial-semantic representation of the scene. Leveraging a suite of pre-trained foundation models, we map an instance graph that encodes object geometry and semantics for symbolic task reasoning. The pipeline of this module is detailed in Appendix A.1.

As illustrated in Fig. 2, given the instance-level semantic map, GPT-4.1 (Achiam et al. 2023) is used to break down template-free natural language instructions into a sequence of atomic actions from a predefined set: `navigate`, `pick`, `place`, and `push/pull/drag`. Each action is paired with a language description that tracks task progress and provides guidance for local planning.

For actions involving spatial displacement (`navigate`, `drag`), the model is further prompted to output a coarse target waypoint to guide planning. We project this target onto a 2D occupancy map built via online SLAM from accumulated LiDAR scans. A local search is then performed around the projected waypoint to identify a collision-free goal pose, avoiding object bounding boxes and structural obstacles. This process yields a globally grounded task plan that aligns with the scene context and is feasible under physical constraints.

Geometry-Constrained Local Manipulation For atomic actions requiring close-range manipulation, we use wrist-mounted depth observations to guide a vision-language model for precise end-effector pose generation. Despite the diverse physical nature of different actions, we unify their execution through a single visuomotor interface, eliminating the need for per-action heuristics.

Specifically, given an RGB observation and the corresponding textual description of the current atomic action, we employ Qwen2.5-VL-72B-Instruct (Bai et al. 2025), a model enhanced with pixel-level grounding capabilities, to infer a task-relevant contact point $p^* \in \mathbb{R}^2$ in the image space.

The contact point is projected onto the aligned depth image to recover its corresponding 3D position in the robot coordinate frame, denoted as $\mathbf{P}_{ee} \in \mathbb{R}^3$. We further prompt the model to generate the orientation $\mathbf{R}_{ee} \in SO(3)$ of the end-effector by determining the gripper's closing direction(x-axis) and approaching direction(z-axis), subject to the following geometric constraints:

- **Axis-alignment constraint:** When the target object or contact region exhibits a dominant axis $\mathbf{a} \in \mathbb{R}^3$, both the x-axis and z-axis of the end-effector should be orthogonal to it:

$$\mathbf{r}_x^\top \mathbf{a} = 0, \quad \mathbf{r}_z^\top \mathbf{a} = 0. \tag{1}$$

- **Surface-normal constraint:** If the object is attached to a planar surface with normal vector $\mathbf{n} \in \mathbb{R}^3$, then the z-axis of the end-effector should align with the surface normal without violating the first constraint:

$$\mathbf{r}_z \parallel \mathbf{n}, \quad \text{s.t.} \quad \mathbf{r}_z^\top \mathbf{a} = 0. \tag{2}$$

By leveraging the expressive grounding capacity of Qwen-VL and constraining the output pose with interpretable geometric conditions, our system achieves reliable local guidance for interaction-intensive manipulation primitives. To the best of our knowledge, this constitutes the first fine-grained manipulation planning system without third-person observation or scripted policies, marking a significant step toward scalable deployment in mobile, in-the-wild environments.

3.2 Policy for Whole-body Control

To effectively execute commands from the high-level planner and adapt to diverse terrains, a whole-body control policy is essential. This work proposes a two-stage, learning-based policy that utilizes a neural network to generate desired joint positions from a set of observations. To enhance the policy's robustness, the training process incorporates a carefully designed, terrain-invariant end-effector sampling strategy and comprehensive domain randomization. The resulting controller is resilient to varied environmental interactions and achieves direct and reliable deployment on physical robots. In this section, we first define the policy and subsequently discuss the training methodology.

Mobile Manipulation Policy The mobile manipulation policy π is formulated as a single network that maps a comprehensive observation vector to a target action $\mathbf{a}_t \in \mathbb{R}^{18}$ as shown in Eq.4. The observation includes the locomotive command $\mathbf{c}_t = (\hat{x}, \hat{y}, \hat{\omega})$, the 6-D end-effect target $\mathbf{e}_t$, a local ground height map $\mathbf{m}_t$, the projected gravity vector $\mathbf{g}_t$, the previous timestep $\mathbf{a}_{t-1} \in \mathbb{R}^{18}$ and the proprioceptive state $\mathbf{s}_t \in \mathbb{R}^{36}$ (joint positions $\mathbf{q}_t$ and velocities $\dot{\mathbf{q}}_t$). All commands and targets are expressed in the robot's base frame. To stabilize the policy output and reduce the simulation-to-reality gap (Fu, Cheng, and Pathak 2023), the action $\mathbf{a}_t$ is formulated as the offsets to the default joint configuration $\mathbf{q}^{default} \in \mathbb{R}^{18}$. The final target, $\mathbf{q}_t^{target} = \mathbf{q}^{default} + \mathbf{a}_t$, is then converted to torques by a Proportional-Derivative (PD) controller.

$$\mathbf{s}_t = (\mathbf{q}_t, \dot{\mathbf{q}}_\mathbf{t}) \tag{3}$$

$$\mathbf{a}_t = \pi(\mathbf{c}_t, \mathbf{e}_t, \mathbf{s}_t, \mathbf{g}_t, \mathbf{m}_t, \mathbf{a}_{t-1}) \tag{4}$$

To ensure the robust policy (Pan et al. 2025a), we employ a two-stage learning approach, as shown in Fig. 3.

Stage 1 In this stage, the arm joints are fixed to focus training on locomotion under a static load, improving exploration efficiency. Inspired by (Mittal et al. 2023), we introduce a gait reward incorporated alongside the base tracking reward to structure the robot's gait. Furthermore, a novel frequency reward is introduced to regulate the gait's cadence. The gait reward r_{gait} encourages specific synchronous (e.g., diagonal) and asynchronous (e.g., lateral) foot contact patterns, with reward functions r_s and r_a detailed in the Appendix B.1.

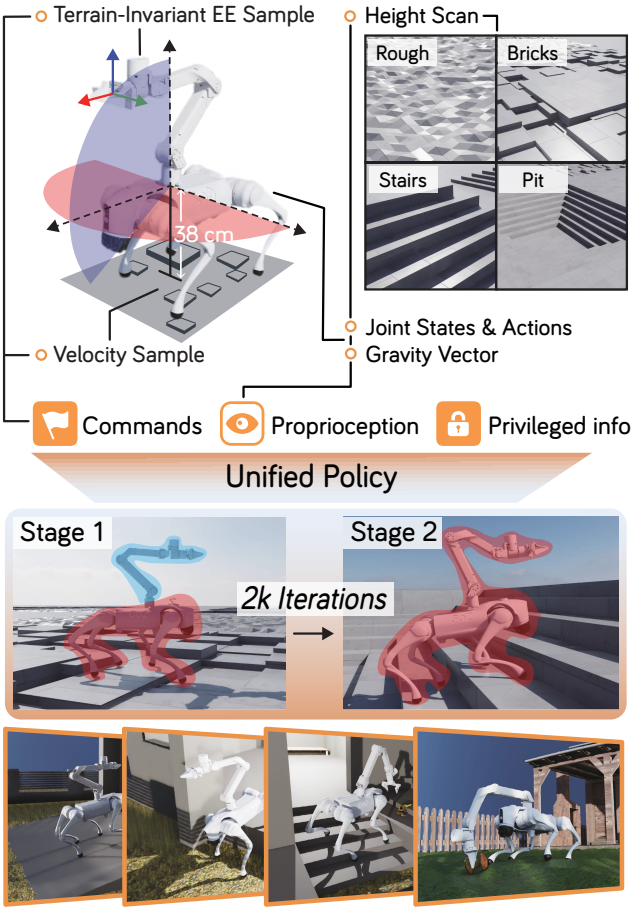

Figure 3: An overview of the mobile manipulator policy and its two-stage training framework.

$$r_{gait} = \prod_{i,j \in \text{sync pairs}} r_s(i,j) \cdot \prod_{k,l \in \text{async pairs}} r_a(k,l) \tag{5}$$

The frequency reward r_{fre} regulates the gait's cadence based on the error from a target frequency f_{target}. The gait frequency $f(leg)$ is the inverse of the time between consecutive ground contacts $(t_k^{cont} - t_{k-1}^{cont})$. The reward is then:

$$err(\text{leg}) = (f(\text{leg}) - f_{target})^2 \tag{6}$$

$$r_f(\text{leg}) = \exp(-0.5 \cdot err(\text{leg})) \tag{7}$$

$$r_{fre} = \prod_{\text{leg} \in \text{FL,FR,RL,RR}} r_f(\text{leg}) \tag{8}$$

Stage 2 Following 2k training iterations, the process transitions to the second stage. In this stage, the policy controls all 18 joints, encompassing both the manipulator and the four legs. Consequently, the reward function is expanded to include end-effector tracking terms, r_{arm}, in addition to the previously described locomotion rewards.

Terrain-Invariant End-Effector Sampling To ensure robust performance across varied topographies, our method employs a terrain-invariant end-effector sampling strategy. The process begins by sampling a target position from a spherical volume defined in the world coordinate system, centered at the robot's arm base. A key aspect of this strategy is that the target's z-axis height is fixed within the world frame before the coordinates are transformed into a Cartesian target position relative to the robot's moving base frame. This approach offers a significant advantage over sampling directly in the arm's local frame, as it effectively decouples the end-effector target from disturbances caused by changes in the robot's base pitch or the underlying terrain height. Consequently, this decoupling improves interaction accuracy during task execution.

Domain Randomizatoin To bridge the simulation-to-reality gap, domain randomization is employed throughout the training process, a strategy supported by recent research (Fu, Cheng, and Pathak 2023; Pan et al. 2025a). To ensure adaptability to different payloads, the end-effector's mass is also randomized during training, improving the policy's ability to handle objects of unknown weight. A detailed breakdown of all randomization parameters and the key reward components is provided in the Appendix B.3.

3.3 Simulation Benchmark

To evaluate navigation, manipulation, and whole-body control as a unified system, we present the first simulation benchmark tailored for long-horizon mobile manipulation in both indoor and outdoor environments.

Asset and Scene Library To support realistic and versatile evaluation environments, we curate a diverse set of assets encompassing both object instances and full-scale 3D scenes. The object assets are sourced from a combination of prior open-source datasets (Wang et al. 2024; Nasiriany et al. 2024), publicly available object repositories, and manually created models.

Object Assets: We curate a diverse set of interactive objects categorized into four types: 50 rigid objects (e.g., common graspable items), 15 containers (e.g., bowls and bins with annotated containment region), 30 articulated structures (e.g., cabinets and doors), and 10 draggable items (e.g., carts and chairs).

Environments: Our benchmark includes 10 realistic scenes, with 5 indoor homes, 2 supermarkets, 1 restaurant, and 2 outdoor courtyards featuring slopes and stairs. All environments are designed for full traversability by legged robots and support multiple initialization zones to allow sampling and spatial variation of large-scale tasks.

Rich Domain-style Variation To ensure generalization, we incorporate variability across four dimensions during simulation rollout: (1) *Object layouts* are varied within semantic constraints across episodes, promoting diversity in interaction. (2) *Physical attributes*, including mass, friction, and articulation limits, are resampled per episode to induce dynamic variability. (3) *Environment conditions* such as lighting, material textures, and clutter elements are randomized to simulate perceptual noise. (4) *Terrain complexity* is varied across outdoor scenes to assess locomotion robustness.

Multi-stage Task Suite Our benchmark includes two categories of tasks: short-horizon manipulation skills merged from ARNOLD (Gong et al. 2023), and long-horizon mobile manipulation tasks designed to reflect practical daily scenarios.

Short-Horizon Arnold Tasks. We integrate four single-step manipulation tasks from the ARNOLD benchmark: PICKUPOBJECT, REORIENTOBJECT, OPENCABINET, and CLOSECABINET. While retaining their original goal state definitions and scene configurations, we adjust the spatial layout and object positioning to accommodate the kinematics and workspace of our quadruped robot platform, ensuring fair and consistent evaluation.

Long-Horizon Mobile Manipulation. To assess the system's embodied reasoning, navigation, and sequential manipulation capabilities, we construct 8 multi-stage tasks spanning diverse indoor and outdoor scenarios. Each task consists of 2–3 subgoals, with a total of 246 indoor and 58 outdoor variations spanning object types, spatial layouts, and interaction modes.

Our task pool emphasizes a broad range of skills spanning grasping, reorientation, container placement, articulated manipulation, and long-term navigation over complex terrain. The combination of short and long-horizon tasks enables benchmarking at both low-level manipulation and high-level planning. Detailed task configurations are elaborated in Appendix C.

Modular Evaluation Protocol We evaluate both overall task success and per-action success rate. For instance, in the CARTDELIVERY task, we define subtasks such as *nav_to_object*, *pick_object*, *nav_to_cart*, *place_object*, *drag_cart*, and *nav_to_goal*. A subtask is considered complete if its corresponding goal condition is met during the task horizon. This protocol captures both execution precision and planning consistency.

4 Experiment

4.1 High-level Planner Performance

To evaluate the performance of our high-level planner in a modular and scalable manner, we conducted experiments based on the benchmark discussed in Section 3.3. Firstly, we tested our local planner on thousands of single-step test cases, focusing on precision and consistency. Secondly, we integrated the global planner and evaluated our proposed approach on several hundred long-horizon mobile manipulation tasks. Additionally, we performed a detailed analysis of the completion rates for the decomposed atomic actions within each task.

ARNOLD Short-horizon Tasks Before moving on to the long-horizon evaluation, we first conducted experiments in relatively confined spaces to demonstrate the fine-grained operation precision and generalization capabilities of our framework. We migrated four short-horizon tasks from ARNOLD and faithfully replicated their continuous monitoring system for goal states.

Their evaluation protocol divides five splits into two categories: 1) *Seen* includes shuffled seen data; 2) *Novel* features one of unseen components (objects, scenes, or goal states). We compare against their strongest baseline model PerAct (Shridhar, Manuelli, and Fox 2022), an end-to-end imitation learning paradigm trained on large-scale human trajectories, which leverages observations from five external cameras to achieve accurate spatial perception. As shown in Table 1, our method achieves substantial overall improvements, demonstrating superior fine-grained manipulation capabilities while relying solely on a single egocentric camera. Moreover, while their performance declined dramatically on novel splits, our method maintains stable performance across all datasets, showcasing generalized ability to handle O.O.D object configurations. Further experiment details are provided in Appendix D.1.

	Seen		**Novel**	
	PerAct	Ours	PerAct	Ours
P.OBJECT	94.03	60.45	25.70	**45.24**
R.OBJECT	19.48	**51.32**	8.23	**52.09**
O.CABINET	31.09	**56.30**	16.62	**51.09**
C.CABINET	60.81	**74.32**	41.32	**79.50**

Table 1: Performance comparison between our approach and PerAct on 4 ARNOLD tasks, evaluated on both seen and generalized splits by success rate (%).

ODYSSEY Long-horizon Tasks Table 2 summarizes the performance of our system across eight long-horizon mobile manipulation tasks, reporting both overall task success rates and success rates for decomposed atomic actions. Notably, ODYSSEY consistently achieves 40% or higher overall success across all tasks, and maintains over 60% success in each atomic skill category, demonstrating robust coordination in a generalized long-horizon task. Building on this, we highlight several key findings from different perspectives of system performance:

Low-level ability: The consistent success rates across indoor and outdoor settings, despite irregular terrains, validate the reliable locomotion and effective pose tracking enabled by our terrain-adaptive whole-body control policy. Most control failures occur from interactions with objects beyond the robot's reachable range.

Fine-grained action: VLM-guided grounding enables high `pick` and `place` success rates, demonstrating strong capability in identifying and localizing semantic targets. Failures primarily stem from suboptimal gripper alignment, indicating limitations in spatial reasoning over object geometry. Tasks involving more complex interactions, such as `drag` and `pull`, occasionally fail due to inaccurate localization, especially with slim handles or occluded items.

Task-level planning: Our global task planner demonstrates strong symbolic reasoning over instance graphs, enabling reliable multi-stage decomposition. In conjunction, our SLAM-based path planner ensures safe and consistent navigation. These components together lead to high `navigate` success rates.

4.2 Low-level Policy Performance

We evaluated the proposed whole-body control policy against RoboDuet (Pan et al. 2025a), a baseline that also uses a two-stage training process. In contrast to RoboDuet's dual-policy (locomotion and manipulation) approach with base-centric sampling, our method employs a single, unified policy trained with a novel terrain-invariant end-effector sampling strategy. For the evaluation, 4096 parallel agents were instantiated with an average of five data samples collected from each agent.

Metric To quantitatively evaluate performance in the simulator, the following metrics are defined:

- **Base Tracking Error**: The error between commanded and actual base velocities, comprising linear (e_x, e_y) and angular e_ω components.
- **End-Effector Position Error**: The Euclidean distance (D_{pos}) between the current and commanded end-effector positions in the world frame.
- **End-Effector Orientation Error**: The quaternion geodesic distance (D_{ori}) between the current $\mathbf{q}_{curr}$ and target $\mathbf{q}_{tar}$ orientations, calculated by $D_{ori} = 2 * arcos(|\mathbf{q}_{curr} \cdot \mathbf{q}_{tar}|)$.

Simulation result Our method was compared under both static (standing) and dynamic (moving) evaluation conditions. Owing to the structural limitations of the robot, sampling unreachable targets may cause self-collisions, thereby degrading the training performance. To alleviate this problem, the sampling space was intentionally reduced during training. For a fair comparison and to evaluate generalization performance, both methods were tested in the same, larger workspace, as detailed in Appendix B.2, with the comparative results summarized in Table 3.

The evaluation results indicate that our policy achieves better performance in base velocity tracking (rows 1-3), an improvement we attribute to the inclusion of terrain data in the policy's observation, which enhances the robot's state estimation. End-effector pose tracking performance remains comparable to the baseline (rows 4-5). Notably, a key aspect of this evaluation is that our policy was trained in an intentionally smaller end-effector workspace, and our policy is adaptive to different topographies (e.g., steps). This demonstrates strong generalization capabilities from a more constrained training domain of our approach.

4.3 Sim-to-real Performance

We conducted real-world experiments to validate the sim-to-real performance of our framework, which integrates the high-level coarse-to-fine task planner with the low-level whole-body control policy.

Hardware The high-level planner was deployed on a PC equipped with an Intel i7-12700KF CPU and an RTX 3060 GPU, communicating with the robot via Ethernet. The low-level policy was deployed on the Jetson Orin NX 16GB mounted on the Go2 platform. Our robot platform, as illustrated in Fig. 4, integrates a 12-DoF Unitree Go2 quadruped with a 6-DoF Arx5 manipulator. The Go2 (15kg weight,

	I.COLLECT	R.NAVIGATE	C.DELIVERY	C.STORAGE	RESTOCKING	SHOPPING	O.COLLECT	O.DELIVERY
Navigate	97.4	86.6	98.3	97.7	98.2	98.3	98.4	95.6
Pick	72.7	/	84.6	79.6	83.3	85.0	69.0	72.7
Place	96.8	/	72.7	83.8	79.2	76.5	95.0	80.0
Push/Pull	/	94.1	/	71.0	/	/	/	85.7
Drag	/	/	69.2	/	/	79.2	/	/
Overall	**66.7**	**69.8**	**41.0**	**44.9**	**56.7**	**47.5**	**63.3**	**46.4**

Table 2: Overall success rates (%) of 8 ODYSSEY long-horizon tasks, along with per-action success rates for each task.

	Stand still		**Move**	
	Roboduet	ours	Roboduet	Ours
$e_x \downarrow$	0.32	**0.08**	9.70	**0.36**
$e_y \downarrow$	**0.34**	2.69	15.42	**2.31**
$e_\omega \downarrow$	0.32	**0.26**	60.59	**0.79**
$D_{pos} \downarrow$	**11.08**	11.48	10.75	**10.57**
$D_{ori} \downarrow$	47.14	**46.93**	47.53	**47.15**

Table 3: The quantitative result under static (stand still) and dynamic (move) conditions.

8kg payload) includes a built-in Unitree L1 LiDAR, and the 3.35kg Arx5 arm is mounted on its back, similar to (Ha et al. 2024). For high-level perception, the platform is equipped with a MID-360 LiDAR for localization and two RealSense cameras: a head-mounted D435i for RGB imagery and a gripper-mounted D405 for RGB-D data. The control policy operates at 50 Hz, with a PD controller issuing motor commands at 200 Hz.

Real-world experiments The ODYSSEY framework was evaluated on two indoor long-horizon tasks: COLLECT (navigate to an object and pick it up) and REARRANGE (pick up an object and place it into a container). These tasks were tested using five different objects, with each object executed ten times. The entire system demonstrated successful sim-to-real transfer on task planning and execution as illusioned in Fig. 4. The quantitative results, including the task success rates, are summarized in Table 4.

	Toy	Can	Bottle	Orange	Banana
COLLECT	4/10	5/10	4/10	3/10	3/10
REARRANGE	5/10	4/10	2/10	2/10	1/10

Table 4: Success rates in real-world experiments.

Despite this success, some sim-to-real gaps persist. For instance, the robot occasionally failed to grasp small objects due to inaccuracies in end-effector tracking and visual perception. These issues are mainly caused by cumulative errors introduced by sensor noise, imperfect camera calibration, and mismatches between simulated and real-world dynamics, all of which degrade the precision required for fine manipulation. Experimental observations further revealed that the grasping success rate was strongly influenced by the robot's base orientation: the success rate was notably higher when the target object was positioned directly in front of the robot, and lower when it was located toward the sides. This suggests that autonomously adjusting the robot's base pose during tasking could serve as a promising strategy to enhance overall manipulation performance.

Through these experiments, our approach has demonstrated strong potential for addressing long-horizon mobile exploration and manipulation tasks, while also highlighting the remaining challenges—robust perception and high-precision control—that must be overcome for seamless real-world deployment.

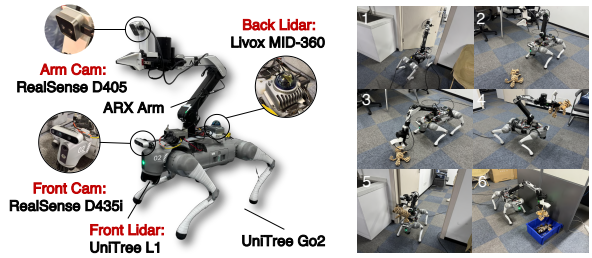

Figure 4: The robot system and real-world experiments

5 Conclusions and Future Work

We present ODYSSEY, a unified framework for open-world mobile manipulation that integrates hierarchical task planning with terrain-adaptive whole-body control. Our approach demonstrates robust sim-to-real transfer and generalization across diverse environments and long-horizon tasks. Future work will extend our benchmark into a comprehensive evaluation paradigm for vision-language models (VLMs) and mobile manipulators, enabling cross-embodiment assessment of semantic reasoning and locomotion-manipulation coordination. Additionally, we aim to explore the emergent capabilities of active perception, where dynamic scene understanding and adaptive motion synergize for more efficient real-world interaction. This direction could unlock new behaviors in cluttered, unstructured environments, further bridging the gap between high-level planning and low-level control.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62576315, No. 62373329), the R&D Program of Zhejiang (No. 2025C01011), the Zhejiang Natural Science Foundation (No. LZ25F030003), and the Baima Lake Laboratory Joint Funds of the Zhejiang Provincial Natural Science Foundation (No. LBMHD24F030002). We would also like to thank Shijia Hu for her assistance in creating the figures for this paper.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; et al. 2023. Rt-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*.

Cheang, C.-L.; Chen, G.; Jing, Y.; Kong, T.; Li, H.; Li, Y.; Liu, Y.; Wu, H.; Xu, J.; Yang, Y.; et al. 2024. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*.

Fu, Z.; Cheng, X.; and Pathak, D. 2023. Deep whole-body control: learning a unified policy for manipulation and locomotion. In *Conference on Robot Learning*, 138–149. PMLR.

Fu, Z.; Zhao, T. Z.; and Finn, C. 2024. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, 4066–4083. PMLR.

Gong, R.; Huang, J.; Zhao, Y.; Geng, H.; Gao, X.; Wu, Q.; Ai, W.; Zhou, Z.; Terzopoulos, D.; Zhu, S.-C.; et al. 2023. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20483–20495.

Grandia, R.; Jenelten, F.; Yang, S.; Farshidian, F.; and Hutter, M. 2023. Perceptive locomotion through nonlinear model-predictive control. *IEEE Transactions on Robotics*, 39(5): 3402–3421.

Gu, Q.; Kuwajerwala, A.; Morin, S.; Jatavallabhula, K. M.; Sen, B.; Agarwal, A.; Rivera, C.; Paul, W.; Ellis, K.; Chellappa, R.; et al. 2024. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 5021–5028. IEEE.

Ha, H.; Gao, Y.; Fu, Z.; Tan, J.; and Song, S. 2024. Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, 5254–5270. PMLR.

Huang, X.; Huang, Y.-J.; Zhang, Y.; Tian, W.; Feng, R.; Zhang, Y.; Xie, Y.; Li, Y.; and Zhang, L. 2023. Open-set image tagging with multi-grained text supervision. *arXiv preprint arXiv:2310.15200*.

Jatavallabhula, K. M.; Kuwajerwala, A.; Gu, Q.; Omama, M.; Chen, T.; Maalouf, A.; Li, S.; Iyer, G.; Saryazdi, S.; Keetha, N.; et al. 2023. Conceptfusion: Open-set multimodal 3d mapping. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*.

Jiang, J.; Zhu, Y.; Wu, Z.; and Song, J. 2025a. DualMap: Online Open-Vocabulary Semantic Mapping for Natural Language Navigation in Dynamic Changing Scenes. *IEEE Robotics and Automation Letters*, 10(12): 12612–12619.

Jiang, Y.; Zhang, R.; Wong, J.; Wang, C.; Ze, Y.; Yin, H.; Gokmen, C.; Song, S.; Wu, J.; and Fei-Fei, L. 2025b. BEHAVIOR Robot Suite: Streamlining Real-World Whole-Body Manipulation for Everyday Household Activities. In *9th Annual Conference on Robot Learning*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, 2679–2713. PMLR.

Liu, M.; Chen, Z.; Cheng, X.; Ji, Y.; Qiu, R.-Z.; Yang, R.; and Wang, X. 2024. Visual whole-body control for legged loco-manipulation. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, 234–257. PMLR.

Liu, P.; Guo, Z.; Warke, M.; Chintala, S.; Paxton, C.; Shafiullah, N. M. M.; and Pinto, L. 2025. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. In *IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025*, 13346–13355. IEEE.

Mittal, M.; Yu, C.; Yu, Q.; Liu, J.; Rudin, N.; Hoeller, D.; Yuan, J. L.; Singh, R.; Guo, Y.; Mazhar, H.; Mandlekar, A.; Babich, B.; State, G.; Hutter, M.; and Garg, A. 2023. Orbit: A Unified Simulation Framework for Interactive Robot Learning Environments. *IEEE Robotics and Automation Letters*, 8(6): 3740–3747.

Nasiriany, S.; Maddukuri, A.; Zhang, L.; Parikh, A.; Lo, A.; Joshi, A.; Mandlekar, A.; and Zhu, Y. 2024. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*.

Pan, G.; Ben, Q.; Yuan, Z.; Jiang, G.; Ji, Y.; Li, S.; Pang, J.; Liu, H.; and Xu, H. 2025a. RoboDuet: Learning a Cooperative Policy for Whole-body Legged Loco-Manipulation. *IEEE Robotics and Automation Letters*.

Pan, M.; Zhang, J.; Wu, T.; Zhao, Y.; Gao, W.; and Dong, H. 2025b. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17359–17369.

Qi, Z.; Zhang, W.; Ding, Y.; Dong, R.; Yu, X.; Li, J.; Xu, L.; Li, B.; He, X.; Fan, G.; et al. 2025. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Qiu, R.-Z.; Hu, Y.; Song, Y.; Yang, G.; Fu, Y.; Ye, J.; Mu, J.; Yang, R.; Atanasov, N.; Scherer, S.; et al. 2025a. Learning generalizable feature fields for mobile manipulation. In *International Conference on Intelligent Robots and Systems (IROS)*.

Qiu, R.-Z.; Song, Y.; Peng, X.; Suryadevara, S. A.; Yang, G.; Liu, M.; Ji, M.; Jia, C.; Yang, R.; Zou, X.; et al. 2025b. Wildlma: Long horizon loco-manipulation in the wild. In *IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025*, 10011–10019. IEEE.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.

Shridhar, M.; Manuelli, L.; and Fox, D. 2022. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*.

Wang, H.; Chen, J.; Huang, W.; Ben, Q.; Wang, T.; Mi, B.; Huang, T.; Zhao, S.; Chen, Y.; Yang, S.; et al. 2024. Grutopia: Dream general robots in a city at scale. *arXiv preprint arXiv:2407.10943*.

Wang, J.; Rajabov, J.; Xu, C.; Zheng, Y.; and Wang, H. 2025. Quadwbg: Generalizable quadrupedal whole-body grasping. In *IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025*, 11675–11682. IEEE.

Yating Wang, M. L. J. Y. H.-S. F. T. H., Haoyi Zhu. 2025. VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhang, C.; Han, D.; Zheng, S.; Choi, J.; Kim, T.-H.; and Hong, C. S. 2023. MobileSAMv2: Faster Segment Anything to Everything. arXiv:2312.09579.

Zhang, H.; Yu, H.; Zhao, L.; Choi, A.; Bai, Q.; Yang, Y.; and Xu, W. 2025. Learning Multi-Stage Pick-and-Place with a Legged Mobile Manipulator. *IEEE Robotics and Automation Letters (RA-L)*.

Zhi, P.; Zhang, Z.; Zhao, Y.; Han, M.; Zhang, Z.; Li, Z.; Jiao, Z.; Jia, B.; and Huang, S. 2025. Closed-loop open-vocabulary mobile manipulation with gpt-4v. In *IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025*, 4761–4767. IEEE.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *European conference on computer vision*, 696–712. Springer.

Zhuang, Z.; Fu, Z.; Wang, J.; Atkeson, C.; Schwertfeger, S.; Finn, C.; and Zhao, H. 2023. Robot Parkour Learning. In *Conference on Robot Learning (CoRL)*.