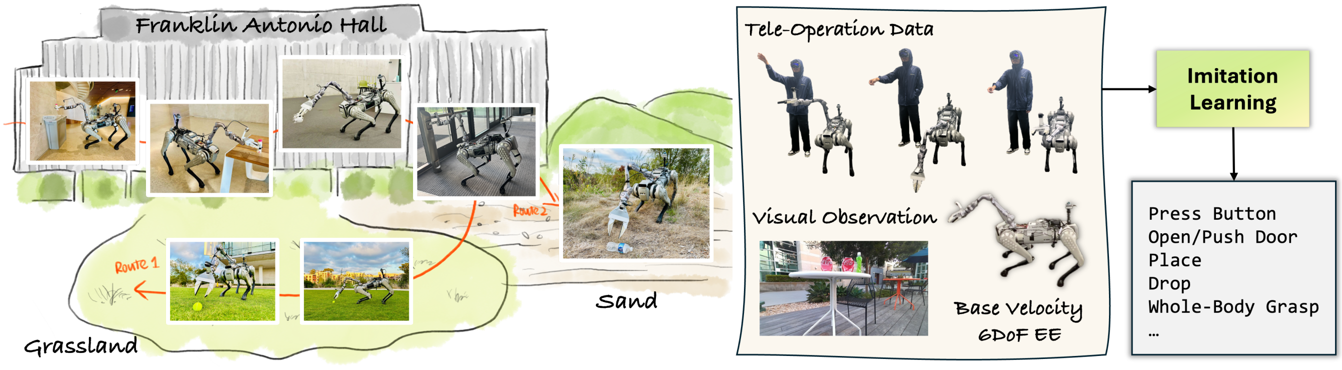
# WildLMa: Long Horizon Loco-Manipulation in the Wild

Ri-Zhao Qiu*[1], Yuchen Song*[1], Xuanbin Peng*[1], Sai Aneesh Suryadevara[1], Ge Yang[2], Minghuan Liu[1]
Mazeyu Ji[1], Chengzhe Jia[1], Ruihan Yang[1], Xueyan Zou[1], Xiaolong Wang[1,3]
*equal contribution
[1]UC San Diego [2]MIT [3]NVIDIA
https://wildlma.github.io



(a) In-The-Wild Quadruped Mobile Manipulation (b) Whole-body VR Teleoperation (c) Skill Learning

Fig. 1: WildLMa implements a framework for in-the-wild manipulation with a quadruped robot, which combines a whole-body controller and imitation learning for effective single-skill learning. (a) Long Horizon Loco-Manipulation in indoor as well as outdoor settings. (b) Teleoperation demonstration for collecting training data for imitation learning. (c) The constructed skill library with various skills, which can be composed by LLM planner to complete complex tasks.

***Abstract*— ‘In-the-wild’ mobile manipulation aims to deploy robots in diverse real-world environments, which requires the robot to (1) have skills that generalize across object configurations; (2) be capable of long-horizon task execution in diverse environments; and (3) perform complex manipulation beyond pick-and-place. Quadruped robots with manipulators hold promise for extending the workspace and enabling robust locomotion, but existing results do not investigate such a capability. This paper proposes *WildLMa* with three components to address these issues: (1) adaptation of learned low-level controller for VR-enabled whole-body teleoperation and traversability; (2) *WildLMa-Skill* — a library of generalizable visuomotor skills acquired via imitation learning or heuristics and (3) *WildLMa-Planner* — an interface of learned skills that allow LLM planners to coordinate skills for long-horizon tasks. We demonstrate the importance of high-quality training data by achieving higher grasping success rate over existing RL baselines using only tens of demonstrations. WildLMa exploits CLIP for language-conditioned imitation learning that empirically generalizes to objects unseen in training demonstrations. Besides extensive quantitative evaluation, we qualitatively demonstrate practical robot applications, such as cleaning up trash in university hallways or outdoor terrains, operating articulated objects, and rearranging items on a bookshelf.**

## I. Introduction

Practical robot mobile manipulation requires generalizable skills and long-horizon task execution. Consider a scenario where a mobile robot is deployed out-of-box at a family house. The robot is tasked with daily chores including collecting the trash around the house and grabbing something for human. To accomplish these tasks, the robot needs skills that generalize to unseen objects and a planner capable of compositing skills over a long horizon.

Existing methods [17, 20, 31, 32, 44, 61, 71] have approached mobile manipulation from two primary directions. Modular methods [32, 44, 71] aim at designing decoupled perception-planning modules. With advances in large-scale vision models [28, 34, 45], recent modular methods [32, 44] exhibit strong generalizability in perception to an open set of language-specified objects. However, their planning modules [6, 20, 32, 44] often rely on heuristic-based motion planning, limiting tasks to mostly simple pick-and-place. End-to-end approaches [11, 12, 17, 22, 31, 69], on the other hand, use learned policies to enable robot with complex actions. They, however, often hold a strong assumption of the small training-testing distribution gap (*e.g.,* sim2real [31] or intra-class variation [17]) and thus do not show strong generalizability. In addition, policies learned via imitation learning are prone to compounding error [26, 69] over long-horizon execution. Thus, these learned skills should be designed to be as atomic as possible for both generalizability and accuracy.

This paper investigates *in-the-wild mobile manipulation* that addresses these issues for real-world deployment. Specifically, in-the-wild manipulation requires the robot to have skills that (1) generalize across texture, lighting, and diverse environments; (2) are capable of long-horizon execution; and

(3) perform complex manipulation beyond pick-and-place.

To this end, we propose WildLMa. For generalizability, WildLMa enables language-conditioned imitation learning (WildLMa-Skill). Building upon ACT [17, 69], WildLMa-Skill improves generalizability via pre-trained CLIP and composable skills. Instead of simply using CLIP features [12, 22], we apply a reparameterization trick [73] to CLIP to compute probability maps given object text query as an auxiliary input. We then use VR teleoperation [9, 13] to collect human demonstrations to acquire complex skills such as non-prehensile manipulation. We adapt a learned low-level controller [31] for VR-based whole-body teleoperation, which significantly increases the robot workspace and reduces the demonstration cost by 26.9% compared to the decoupled strategy. Finally, based on a library of acquired generalizable and atomic skills, WildLMa provides a language interface (WildLMa-Planner) that allows interfacing with LLMs to composite skills for long-horizon execution.

In summary, our contributions are:

- A generic framework with techniques that allow generalizable language-condition imitation learning (*WildLMa-Skill*) with interfacing to the LLM planner (*WildLMa-Planner*).
- Demonstrations of in-the-wild mobile manipulation tasks with full-stack and systematic deployment of the proposed framework.
- Comprehensive evaluation and ablation for the proposed technique, which paves the way for future study.

## II. Related Work

*a)* ***Mobile Manipulation****:* Mobile manipulation has gained increasing attention for its vision of enabling robots to perform diverse practical tasks. In terms of hardware configurations, wheeled robots have made substantial strides [1, 29, 32, 55, 58, 61, 71] for its reliable base movement [54], while recently, legged robots have also gained more interest for its robust locomotion [10, 63] and the extended workspace via whole-body arm-base coordination [15, 22, 31, 42].

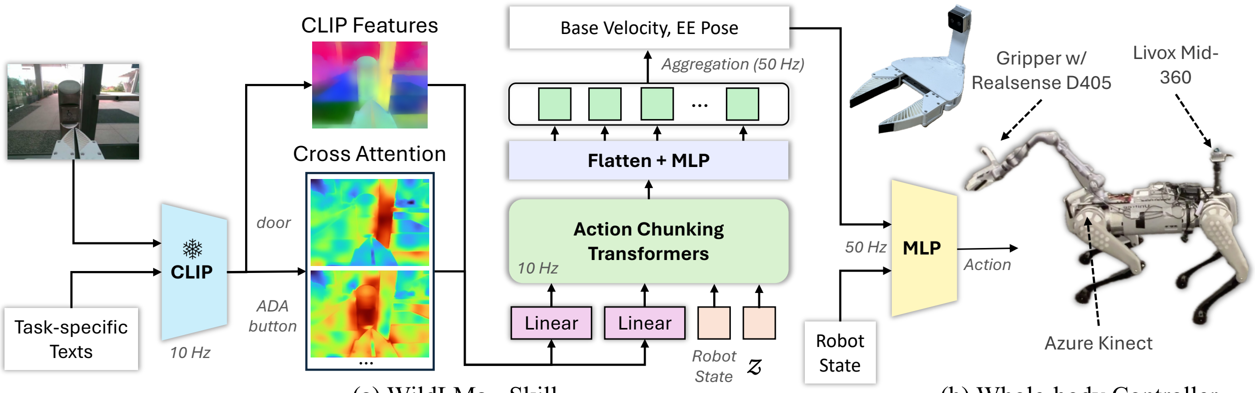
Categorized by methodology, existing methods can be divided into modular methods and end-to-end methods. Recent modular approaches [2, 27, 32, 33, 44, 66, 67, 68, 71] design decoupled perception-planning strategy. In particular, perception [32, 44, 71] are often done by applications of vision foundation models [21, 28, 34, 45]; whereas grasping are done by off-the-shelf pose prediction models (*e.g.,* GraspNet [14]) and IK solver [47]. Despite strong perception designs, modular methods are mostly limited to simple pick-and-place tasks. On the other hand, end-to-end approaches use Reinforcement Learning (RL) [19, 31, 40, 60, 61] or Imitation Learning (IL) [17, 22, 48] to enable complex tasks beyond pick-and-place such as articulated manipulation [4, 61] and non-prehensile manipulation [17, 22]. However, these work often fall short when training-testing distributions mismatch.

Most closely related to our work, Yokoyama *et al.* [66] proposed to use sim2real RL for in-the-wild mobile manipulation. However, they do not investigate manipulation tasks other than simple pick-and-place. WildLMa uses imitation learning to learn diverse skills with generalizability, task complexity, and long-horizon run for in-the-wild execution.

*b)* ***Long-horizon Mobile Manipulation****:* For robots to assist with real-world tasks such as cleaning up home, they need to be capable of dealing with long-horizon mobile manipulation, where independent skills are planned and triggered to complete given goals. Existing methods rely on sampling-based planning [18, 53], RL [19, 30, 65, 66], and Large Language Models (LLMs) [20, 24, 46, 51] to coordinate skill primitives for long-horizon task execution. Recent work [20, 24, 46, 51] have found that LLM-based methods, especially Large-Mutlimodal Models (LMMs) [8, 37], are promising to serve as effective planners for embodied agents, where the research efforts are centered around hierarchical search [46] and re-planning [71]. WildLMa is intended to be orthogonal to these existing work in LLM planner. Instead of studying the planning capability, we investigate the potential of interfacing LMMs with skills acquired via imitation learning for practical applications.

*c)* ***Imitation Learning****:* Imitation learning has demonstrated promising results through learning from real-world expert demonstrations [9, 11, 13, 17, 22, 48, 57, 64]. Investigated for decades since the 80s [41], behavior cloning [5, 41] is one of the most commonly used imitation learning approach that learns an end-to-end mapping from observations to actions. Recently, researchers have shown that this classic approach not only allows complex manipulations [9, 13, 17, 70], but also holds the potential that scaling up training data with low-cost hardware [12, 17, 22, 48, 59, 64, 69] will lead to generalizable policies. Similar to existing work [9, 13, 23, 43, 49], we also use VR devices to collect expert demonstrations that minimize the expert-agent observation gap [69]. To reduce the cross-embodiment gap between the human operator and the quadruped robot, we combine the VR demonstration with learned whole-body controller. In addition, we also improve the vanilla ACT model [17, 69] to support language-conditioned imitation learning that is more generalizable with autonomous termination.

*d)* ***Whole-body Control****:* Quadruped Whole-Body Control (WBC) draws inspiration from the natural motions of animals to extend the robot workspace via arm-base coordination. The WBC capability is usually achieved via model-based hierarchical trajectory optimization [3, 35, 50, 72, 74] or sim2real RL [15, 22, 31]. Our work is based on the low-level controller proposed by VBC [31], which designed a bi-level RL paradigm with a low-level whole-body controller. Notably, some existing work has also attempted to combine teleoperation with whole-body control for quadruped robots [42, 72] but does not investigate learning skills from teleoperation. Most related to our work, Ha *et al.* [22] demonstrated whole-body imitation learning with handheld data collection hardware. The main differences between our work and Ha *et al.* [22] are (1) the data collected without teleoperating the robot can include only wrist camera observation, which may lead to worse performance than multi-camera setup as we empirically verify and (2) Ha *et al.* [22]



(a) WildLMa - Skill (b) Whole-body Controller

Fig. 2: **Overview of WildLMa models and robot setups.** (a) WildLMa takes a frozen CLIP model to encode task-specific texts and visual observations; (b) Our robot platform is a Unitree B1 quadruped combined with a Unitree Z1 arm and a 3D-printed gripper, with two RGBD cameras and one lidar mounted on.

focus on execution of short tasks; while we investigate in-the-wild mobile manipulation with long horizon task execution.

## III. Method

WildLMa designs three components to address challenges for in-the-wild mobile manipulation. Sec. III-A describes adapting a whole-body controller to support efficient teleoperation and more diverse real-world tasks. In Sec. III-B, we propose *WildLMa-Skill*, which modifies the pre-trained CLIP model [45] for generalizable imitation learning. WildLMa-Skill then constructs a skill library consisting of learned skills and analytical skills (*e.g.,* navigation). Finally, *WildLMa-Planner* (Sec. III-C) interfaces WildLMa-Skill with an LLM planner to carry out long-horizon execution.

### A. *Whole-body VR Teleoperation*

Recent imitation learning methods have benefited from improved data collection methods via VR/AR-based teleoperation [9, 13, 23, 43]. However, though human operators can naturally tele-operate bipedal humanoid robots [9, 16, 23], it is non-trivial to tele-operate quadruped robots due to the embodiment gap [39] between two-legged human and quadruped robots inspired by four-legged animals.

To reduce the need for the tele-operator to consider both the base movement and the arm movement, we propose to use a whole-body controller [31] that allows smooth arm-base coordination for the robot. In particular, we use the low-level whole-body policy developed by Liu *et al.* [31]. Trained with RL, the learned whole-body controller takes in *base commands* (linear velocity and angular velocity) and *6DOF end effector pose* w.r.t. the arm base. The policy outputs arm and base joint commands for coordinated movement that extends the workspace (illustrated in Fig. 1).

Based on the pre-trained low-level controller, we then design an interface for human users to teleoperate the robot. We use the OpenTV framework [9] with Apple Vision Pro, which allows real-time video streaming, tracking of 6DOF poses of head and hands, and 3D gesture keypoints. To minimize the expert-agent observation gap [69], the tele-operator gets real-time streams of the robot's head camera views and wrist camera views.

To translate tele-operator movement to robot movement, we linearly transform the operator's right wrist pose (relative to their initial hand pose) $T_{right} \in \mathrm{SE}(3)$ into the relative end effector pose $T_{ee} \in \mathrm{SE}(3)$. We scale the translations with a constant $s_c$, as we find that the workspace of the Z1 arm is slightly larger than average human arms. More concretely, let $\mathbf{R}_{right}$ be the rotational component and $\mathbf{t}_{right}$ be the translational component of $T_{right}$, $T_{ee}$ is given by,

$$T_{ee} = \begin{bmatrix} \mathbf{R}_{right} & s_c \cdot \mathbf{t}_{right} \\ \mathbf{0}^\intercal & 1 \end{bmatrix}. \quad (1)$$

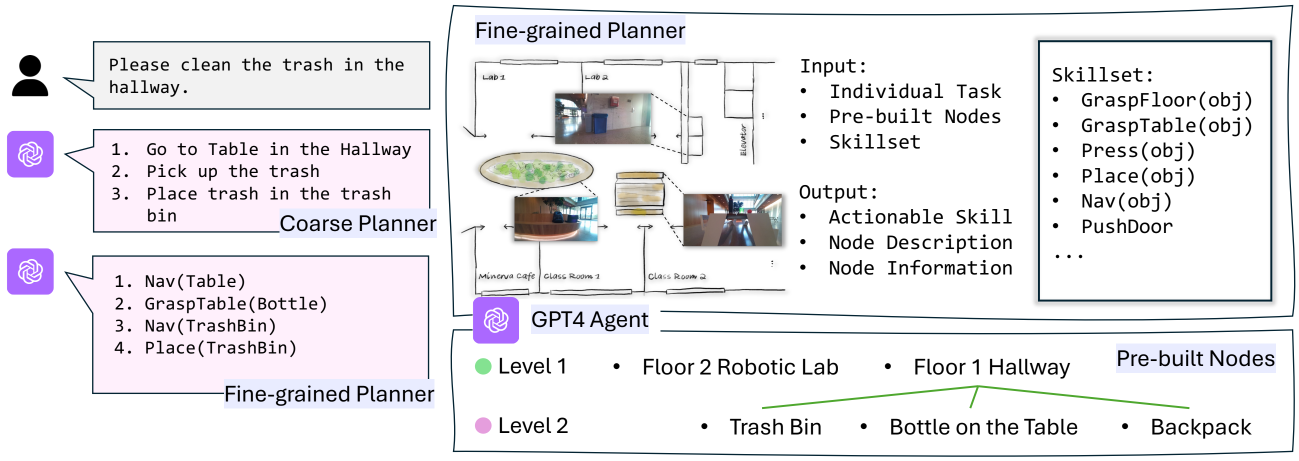
The gripper open-close actions are then naturally mapped from the pinching of the thumb and the index finger (via 3D keypoints). The whole-body controller automatically controls the base rotation to coordinate with the arm. In turn, the teleoeprator's left wrist governs planar base movements (*e.g.,* angular and linear velocities). When the tele-operator pinches their left fingers, VR tracks the pose $T_{left}$ as a virtual joystick with deadzone ($x_{th} = 5cm$). We find this simple base command mapping sufficient for the tasks involved.

### B. *WildLMa-Skill*

*WildLMa-Skill* contains skills from two categories: skills acquired via imitation learning and with analytical planners.

*a) WildLMa-Skill - Imitation Learning:* The collected real-world demonstrations can be turned into autonomous skills with existing behavior cloning methods [9, 17]. Many existing methods, however, struggle to generalize to novel environments [17]. To improve the generalizability of learned skills without expensive demonstration collection cost, we propose to adapt pre-trained CLIP [45] to ACT [69] for imitation learning of individual skills.

**Improving Generalizability with CLIP.** We encode camera observations with a frozen CLIP visual backbone. Instead of simply using intermediate CLIP features as in [12, 22], we apply MaskCLIP [73], a reparameterization trick to generate image-text cross attention map. More concretely, let $\Omega$ be



Please clean the trash in the hallway.

1. Go to Table in the Hallway
2. Pick up the trash
3. Place trash in the trash bin

1. Nav(Table)
2. GraspTable(Bottle)
3. Nav(TrashBin)
4. Place(TrashBin)

Input:
- Individual Task
- Pre-built Nodes
- Skillset

Output:
- Actionable Skill
- Node Description
- Node Information

Skillset:
- GraspFloor(obj)
- GraspTable(obj)
- Press(obj)
- Place(obj)
- Nav(obj)
- PushDoor

...

Fig. 3: Overview of WildLMa-planner. Given a constructed hierarchical scene graph, WildLMa-planner adopts a coarse-to-fine searching mechanism to determine node traversal and structured actions to take.

the space of RGB images. The original CLIP [45] is a mapping function $f_{visual}(\cdot) : \Omega \mapsto \mathbb{R}^C$, where $\mathbb{R}^C$ is the image-text embedding space learned from contrastive learning [45]. MaskCLIP modifies the network architecture to a new mapping function $g_{visual}(\cdot) : \Omega \mapsto \mathbf{H} \times \mathbf{W} \times \mathbb{R}^C$, which is a feature map aligned to the CLIP embedding space $\mathbb{R}^C$ (illustrated in Fig. 2).

**Image-text Cross Attention.** Consistent with findings by Chi *et al.* [12], we found that the adaptation of CLIP [45] improves the performance. However, when tested with objects unseen in the training demonstrations, the success rate is still unsatisfactory. Thus, we propose to make the acquired skills *language-conditioned* by introducing cross-attention. We provide task-specific texts during both the training and testing time (*e.g.,* for the ADA door button-pressing task, we use ‘door’ and ‘ADA button’) with CLIP text embedding $f_{text}(\cdot) : Text \mapsto \mathbb{R}^C$. With slight abuse of notation, the text vector can then be compared with the CLIP feature map via cosine similarity

$$\textsc{CrossAtt}(\cdot,\cdot) = \frac{g_{visual}(\cdot) f_{text}(\cdot)}{||g_{visual}(\cdot)||\,||f_{text}(\cdot)||}, \tag{2}$$

where the comparison is done independently on the pixel level. The resulting similarity is comparable to the probability map of text queries. We apply dropout [52] to cross-attention during training to avoid over-reliance on attention.

**Autonomous Termination.** To autonomously terminate skills, in order to hand control back to high-level planners, we add a virtual *‘end’* action signal prediction. Empirically, adding the end signal to only the end of the episode does not work, as the supervision is too sparse. Our proposed solution is to implement a buffer of end signal for every skill such that the last $n = 10$ frames of the demonstrations carry the end signal. During deployment, we use a sliding window detector to terminate task execution if the end signal is greater than $\tau = 0.8$ for 10 consecutive predictions.

*b) WildLMa-Skill - Analytical Planning:* In this paper, we learn all manipulation-related skills with imitation learning. For the base-only skill (*i.e.,* navigating from a known location to another known location), we implement it with analytical planning.

### C. WildLMa-Planner

The *WildLMa-Skill* module provides skills that can be composed for long-horizon execution, which is intentionally designed to be agnostic of the high-level planner. Here, we propose *WildLMa-Planner*, a simple LLM-based planner to show how learned skills can be composed.

**Initial Mapping.** We implement a LiDAR-based SLAM system using FAST-LIO [62] and DLO [7] to obtain consistent robot pose estimation in the world frame. We manually annotate pose-level waypoints (*e.g.,* stand in front of receptacles) and connectivity for task execution. The robot stands at every waypoint to capture images with its head camera. To automatically annotate the semantics of each waypoint, GPT4-V [37] provides high-level descriptions of images and lists of objects of each waypoint. We manually create abstract nodes (*e.g.,* a room with multiple pose-level waypoints) to construct a hierarchical graph for searching. Note that off-the-shelf scene graph construction methods [20, 25, 36] can potentially replace this step.

**Hierarchical Long-horizon Planning.** We adopt a hierarchical coarse-to-fine approach to translate template-free commands into detailed, actionable robot skills.

*Coarse Planner.* Using CoT [56], the coarse planner receives template-free instructions and decomposes them into individual tasks. For instance, the command ‘clean the trash in the hallway’ can be decomposed into tasks ‘navigate to hallway’, ‘pick up the trash’, and ‘place trash in the trash bin’. We will release the detailed prompts.

*Fine-grained Planner.* The fine-grained planner invokes actionable skills at particular nodes given individual tasks generated by the coarse planner. The fine-grained planner has prior knowledge of the robot’s skill library (shown in Fig. 3) and nodes constructed in the initial mapping stage. For each task, the agent uses a breadth-first search (BFS) approach to search nodes and identify the optimal goal node. During this stage the LLM acts as a heuristic evaluator, estimating the likelihood of a node being the most likely location related to

|  | Tabletop Grasping |  | Button Pressing |  | Ground Grasping |  |  |
|---|---|---|---|---|---|---|---|
| Method | I.D. | O.O.D. | I.D. | O.O.D | I.D. | O.O.D | Avg. Succ. |
| WildLMa (Ours) | **94.4%** | 75% | **80%** | **57.5%** | **60%** | **60%** | **71.2%** |
| ACT (Mobile ALOHA) [17] | 77.8% | 19.4% | 55% | 25% | 60% | 30% | 40.8% |
| OpenTV [9] | 88.9% | **77.8%** | 75% | 25% | 50% | 50% | 64.4% |
| VBC [31] | 50%* | 50%* | NA† | NA† | 43.8%* | 43.8%* | 46.9% |
| GeFF [44] | 55.6%* | 55.6%* | NA† | NA† | NA† | NA† | 55.6% |

TABLE I: **Success rate of autonomous skill execution**. Imitation learning methods outperform RL [31] and zero-shot method [44] on comparable tasks. Both OpenTV and *WildLMa* achieve noticeably higher success rates in the challenging O.O.D. setting. †: methods involve learned/manual policies that are not trivially applicable to the task settings. *: Method does not differentiate object sets and success rates are averaged on I.D. and O.O.D. object sets.

| Pipeline | Collect & Drop Trash | Shelf Rearrangement |
|---|---|---|
| WildLMa (Ours) | **7/10** | **3/10** |
| ACT [17, 69] | 0/10 | 0/10 |

TABLE II: **Evaluation of long-horizon execution**. Given a few training demonstrations (10), WildLMa improves long-horizon task success rate via (1) improved generalizability of single skill and (2) divide-and-conquer.

| Backbone | In Dist. | Out of Dist. | Avg. Succ. |
|---|---|---|---|
| CLIP [45] | 83.3% | 69.4% | 76.4% |
| ResNet [69]* | 77.8% | 19.4% | 48.6% |
| DinoV2 [38] | **88.9%** | **77.8%** | **83.3%** |

TABLE III: **Ablation of different visual encoders** pre-trained with different objectives. The evaluation is done on the object-grasping tasks. *: we followed ACT [17, 69] to use ImageNet-pretrained ResNet-18 as the encoder, which has fewer parameters.

the task, based on the semantic context and objects present at the node. Once the target node is identified, the planner constructs a plan detailing the navigation and manipulation sequence drawn from the pre-defined skill library.

## IV. Experiments

**Hardware Platforms.** We use the Unitree B1 quadruped robot with a Unitree Z1 arm. We replace the beak-like default Z1 end effector with a 3D-printed parallel soft gripper, which was adapted from UMI-Gripper [12] to directly operate the gripper with gear rotations. For perception, an Azure Kinect camera is mounted on the robot's head, and an Intel Realsense D405 is used as the in-wrist camera. A LIVOX MID-360 LiDAR is installed at the robot's tail for enhanced localization during navigation.

**Implementation Details.** The WildLMa-Skill module independently trains weights for each skill (with 30-60 demonstrations each acquired via tele-operation). The head/wrist RGB observations are processed through a CLIP [45] ViT-B/16 encoder with MaskCLIP [73] re-parameterization. Task-specific texts are then compared with the feature maps to generate cross-attention, where texts may differ in training sequences and testing run. For navigation between given waypoints, we implement a PD-based waypoint follower. WildLMa-Planner requires geometric annotations of nodes and edges. For efficiency, the spatial locations of nodes are annotated by operating the robot to turn 360 degrees during the initial scene scanning, and the edges are made between physically adjacent nodes with no obstacle in between.

**Experimental Protocol.** We define two experiment settings to investigate the generalizability of skills learned via imitation learning [9, 13, 17]. The *in-distribution (I.D.)* setting tests the learned skills with backgrounds and object arrangements approximately similar to the training demonstrations. Note that, to make the setting more realistic, we do not enforce identical robot positioning and lighting conditions even in I.D. settings. The *Out-of-distribution (O.O.D.)* setting permutes the testing objects (placement/texture), receptacles, and background environments for learned skills. Illustrations of the differences between these two settings can be found on the website.

**Baseline Implementation.** Besides ablating design choices of our components, we implement several baselines to validate the efficacy of WildLMa. To compare with existing imitation learning methods, we choose Mobile ALOHA [17] which uses ACT [69] with ResNet-18 and OpenTV [9] using ACT and DinoV2 [38]. Unless specifically noticed, these baselines use the same training data as WildLMa. In addition, we compare two recent works in quadruped loco-manipulation [31, 44] to compare WildLMa against RL-based and zero-shot grasping methods. Note that both VBC [31] and GeFF [44] were designed for grasping, so they are not trivially applicable to non-prehensile manipulation such as button pressing.

### A. Evaluation

We address important research questions in our evaluation:

- What advantages does WildLMa-Skill have compared to existing baselines in quadruped manipulation? [A1, A2]
- How does WildLMa-Planner perform in long-horizon execution? [A3]
- Are the design choices (*e.g.,* visual backbone and cross-attention) effective? [A4, A5]
- Does whole-body control improve teleoperation? [A6]
- What are the real-world applications of WildLMa? [A7]

**A1. WildLMa outperforms recent imitation learning baselines.** From Tab. I, we can see that WildLMa achieves best overall success rate. Compared to vanilla ACT [17, 69], WildLMa achieves slightly better success rates on I.D. setting

| | Whole-body (Ours) | | Decoupled Control | | W/o Whole-body (Arm Only) | |
|---|---|---|---|---|---|---|
| Metric | Ground Grasping | Rearrange Shelf | Ground Grasping | Rearrange Shelf | Ground Grasping | Rearrange Shelf |
| Average Time | 21.87s | 27.25s | 37.35s | 29.81s | - | 27.88s |
| Success Rate | 95% | 70% | 80% | 40% | 0% | 70% |

TABLE IV: **Comparison of success rate and completion time** for our whole-body controller, decoupled control with manual base pitching and arm control implemented via Unitree SDK, and arm-only policies. Four teleoperators are tasked to manipulate objects at various heights for three trials in each task.

| Camera | Tabletop Grasping | Button Pressing | Door Opening |
|---|---|---|---|
| Head + Wrist | **94.4%** | 80% | **70%** |
| Head Only | 27.8% | 75% | 30% |
| Wrist Only | 83.3% | **85%** | 10% |

TABLE V: **Ablation of input visual modality**. Tasks that involve occlusion significantly benefit from multi-view setup.

| Backbone | In Dist. | Out of Dist. | Avg. Succ. |
|---|---|---|---|
| w/ cross-attention (Ours) | **94.4%** | **75%** | **84.7%** |
| w/o cross-attention | 83.3% | 69.4% | 76.4% |

TABLE VI: **Ablation of cross-attention** on the object-grasping tasks. Cross-attention improves both I.D. and O.O.D. setting by using additional task-specific information.

and significantly better success rate on the O.O.D. setting. We reason this is because ResNet is vulnerable to changes in lighting and texture. OpenTV [9], on the other hand, shows more robustness to these adversarial conditions due to its use of the recent DinoV2 backbone [38], but slightly underperforms our method.

**A2. WildLMa outperforms RL and zero-shot baselines.** Due to less reliance on real-world demonstrations, RL and zero-shot baselines demonstrate less performance gap between I.D. and the O.O.D. settings in Tab. I. As an RL-based method, VBC [31] suffers from sim2real gaps such as inaccurate contact modeling and cumulative sensor latencies. Therefore, VBC performs worse in real-world settings than its simulation counterpart. On the other hand, zero-shot modular methods such as GeFF [44] do not naturally exhibit corrective behavior like learning-based methods, which are vulnerable to errors compounding from different modules.

**A3. WildLMa is capable of handling long-horizon manipulation under perturbations.** Tab. V validates the efficacy of WildLMa in handling long-horizon tasks under certain perturbations. We include videos on the website. Our experiments include 20 training sequences with variations in robot positioning, lighting, and object placement in both the training and testing time. ACT fails entirely for long-horizon tasks when trained directly on a few sequences of demonstrations. On the other hand, WildLMa successfully learns generalizable skills from a limited number of demonstrations to achieve better success rates for long-horizon execution.

**A4. Pre-trained Visual Backbones improve skill generalizability.** We ablate the choice of visual backbones in Tab. III. CLIP [45] is the simple application of CLIP features without cross-attentions. While different backbones perform similarly in the I.D. setting, we see that frozen large models [38, 45] perform much better in the O.O.D. setting.

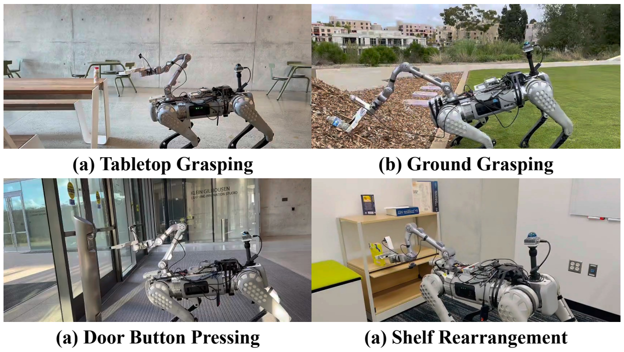


Fig. 4: Qualitative illustrations of some evaluated tasks.

**A5. Cross-attention significantly improves O.O.D. imitation learning.** Tab. VI shows the proposed cross-attention improves both the I.D. and O.O.D. performance of CLIP [45] module by introducing additional task-specific text prompts.

**A6. Whole-body controllers enable efficient VR teleoperation of quadruped robots.** The motivation for combining whole-body control and teleoperation is to improve teleoperation efficiency. To validate this point, we report the statistics of teleoperation in Tab. IV, which shows our learning-based controller outperforms the decoupled analytical controller from Unitree SDK. Since these tasks require reaching objects at various heights (toys and books at different levels of storage), teleoperation without whole-body control fails to grasp from the ground due to the limited workspace.

**A7. WildLMa allows the robot to learn diverse tasks.** Besides qualitative samples in Fig. 1, we provide more videos of our robot working on different practical tasks in the supplementary video and the website.

## V. Conclusion

In this paper, we present WildLMa, a modular framework that includes (1) *WildLMa-Skill*, which implements a library of generalizable visuomotor skills that improve ACT [69] for learning generalizable imitation learning skills; and (2) *WildLMa-Planner*, an interface that enables interactions between imitation learning skills and LLM planner to support long-horizon task execution. Furthermore, we deploy this framework on a quadruped robot controlled by a whole-body controller, which allows us to efficiently collect demonstration data and support extended workspace for diverse tasks. In summary, WildLMa implements practical, generalizable skills, and long-horizon manipulation, which we hope will motivate future research toward in-the-wild mobile manipulation that facilitates real-world deployment of robots.

## References

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

[2] E. Arcari *et al.*, "Bayesian multi-task learning mpc for robotic mobile manipulation," *IEEE Robotics and Automation Letters*, 2023.

[3] C. D. Bellicoso *et al.*, "Alma-articulated locomotion and manipulation for a torque-controllable robot," in *2019 International conference on robotics and automation (ICRA)*, 2019.

[4] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation," in *ECCV*, 2024.

[5] M. Bojarski, "End to end learning for self-driving cars," in *arXiv preprint arXiv:1604.07316*, 2016.

[6] M. Chang *et al.*, "Goat: Go to any thing," in *RSS*, 2024.

[7] K. Chen, B. T. Lopez, A.-a. Agha-mohammadi, and A. Mehta, "Direct lidar odometry: Fast localization with dense point clouds," *IEEE Robotics and Automation Letters*, 2022.

[8] A.-C. Cheng *et al.*, "Spatialrgpt: Grounded spatial reasoning in vision language model," *arXiv preprint arXiv:2406.01584*, 2024.

[9] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," in *CoRL*, 2024.

[10] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," in *ICRA*, 2024.

[11] C. Chi *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," in *RSS*, 2023.

[12] C. Chi *et al.*, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *RSS*, 2024.

[13] R. Ding *et al.*, "Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning," *arXiv preprint arXiv:2407.03162*, 2024.

[14] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020.

[15] Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: Learning a unified policy for manipulation and locomotion," in *Conference on Robot Learning*, 2023.

[16] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," in *CoRL*, 2024.

[17] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv*, 2024.

[18] C. R. Garrett, T. Lozano-Pérez, and L. P. Kaelbling, "Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning," *arXiv*, 2020.

[19] J. Gu, D. S. Chaplot, H. Su, and J. Malik, "Multiskill mobile manipulation for object rearrangement," in *The Eleventh International Conference on Learning Representations*, 2023.

[20] Q. Gu *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *ICRA*, 2024.

[21] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[22] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song, "Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers," in *CoRL*, 2024.

[23] T. He *et al.*, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *CoRL*, 2024.

[24] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *International conference on machine learning*, 2022.

[25] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3d scene graph construction and optimization," *arXiv preprint arXiv:2201.13360*, 2022.

[26] A. Iyer *et al.*, "Open teach: A versatile teleoperation system for robotic manipulation," *arXiv preprint arXiv:2403.07870*, 2024.

[27] M. Ji, R.-Z. Qiu, X. Zou, and X. Wang, "Graspsplats: Efficient manipulation with 3d feature splatting," in *CoRL*, 2024.

[28] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[29] T. Lew *et al.*, "Robotic table wiping via reinforcement learning and whole-body trajectory optimization," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[30] Z. Liang, Y. Mu, H. Ma, M. Tomizuka, M. Ding, and P. Luo, "Skilldiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution," in *CVPR*, 2024.

[31] M. Liu *et al.*, "Visual whole-body control for legged loco-manipulation," in *CoRL*, 2024.

[32] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Ok-robot: What really matters in integrating open-knowledge models for robotics," *arXiv preprint arXiv:2401.12202*, 2024.

[33] P. Liu *et al.*, "Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation," *arXiv preprint arXiv:2411.04999*, 2024.

[34] S. Liu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[35] Y. Ma, F. Farshidian, T. Miki, J. Lee, and M. Hutter, "Combining learning-based locomotion policy with

model-based manipulation for legged mobile manipulators," *IEEE Robotics and Automation Letters*, 2022.

[36] D. Maggio *et al.*, "Clio: Real-time task-driven open-set 3d scene graphs," *arXiv preprint arXiv:2404.13696*, 2024.

[37] OpenAI, "Gpt-4 technical report," OpenAI, Tech. Rep., 2023.

[38] M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[39] A. Padalkar *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

[40] G. Pan *et al.*, "Roboduet: A framework affording mobile-manipulation and cross-embodiment," *arXiv preprint arXiv:2403.17367*, 2024.

[41] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *Advances in neural information processing systems*, 1988.

[42] T. Portela, G. B. Margolis, Y. Ji, and P. Agrawal, "Learning force control for legged manipulation," in *ICRA*, 2024.

[43] Y. Qin *et al.*, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," *arXiv preprint arXiv:2307.04577*, 2023.

[44] R.-Z. Qiu *et al.*, "Learning generalizable feature fields for mobile manipulation," *arXiv preprint arXiv:2403.07563*, 2024.

[45] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, PMLR, 2021.

[46] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," in *CoRL*, 2023.

[47] *Ros moveit motion planning framework*, https://moveit.ros.org/, Accessed: 2024-09-13.

[48] N. M. M. Shafiullah *et al.*, "On bringing robots home," *arXiv preprint arXiv:2311.16098*, 2023.

[49] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola, "Distilled feature fields enable few-shot language-guided manipulation," in *CoRL*, 2023.

[50] J.-P. Sleiman, F. Farshidian, and M. Hutter, "Versatile multicontact planning and control for legged loco-manipulation," *Science Robotics*, 2023.

[51] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," in *ICCV*, 2023.

[52] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, 2014.

[53] S. Srivastava, E. Fang, L. Riano, R. Chitnis, S. Russell, and P. Abbeel, "Combined task and motion planning through an extensible planner-independent interface layer," in *ICRA*, 2014.

[54] *Stretch open source mobile manipulator - hello robot*, https://hello-robot.com/stretch-3-product, Accessed: 2024-09-01.

[55] C. Sun *et al.*, "Fully autonomous real-world reinforcement learning with applications to mobile manipulation," in *Conference on Robot Learning*, 2022.

[56] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *NeurIPS*, 2022.

[57] J. Wong *et al.*, "Error-aware imitation learning from teleoperation data for mobile manipulation," in *Conference on Robot Learning*, 2022.

[58] J. Wu *et al.*, "Tidybot: Personalized robot assistance with large language models," *Autonomous Robots*, 2023.

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

[60] F. Xia, C. Li, R. Martín-Martín, O. Litany, A. Toshev, and S. Savarese, "Relmogen: Integrating motion generation in reinforcement learning for mobile manipulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[61] H. Xiong, R. Mendonca, K. Shaw, and D. Pathak, "Adaptive mobile manipulation for articulated objects in the open world," *arXiv preprint arXiv:2401.14403*, 2024.

[62] W. Xu and F. Zhang, "Fast-lio: A fast, robust lidar-inertial odometry package by tightly-coupled iterated kalman filter," *IEEE Robotics and Automation Letters*, 2021.

[63] R. Yang *et al.*, "Generalized animal imitator: Agile locomotion with versatile motion prior," *arXiv preprint arXiv:2310.01408*, 2023.

[64] S. Yang *et al.*, "Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation," in *CoRL*, 2024.

[65] S. Yenamandra *et al.*, "Homerobot: Open-vocabulary mobile manipulation," *arXiv preprint arXiv:2306.11565*, 2023.

[66] N. Yokoyama *et al.*, "Asc: Adaptive skill coordination for robotic mobile manipulation," *IEEE Robotics and Automation Letters*, 2023.

[67] J. Zhang *et al.*, "Gamma: Graspability-aware mobile manipulation policy learning based on online grasping pose fusion," *arXiv preprint arXiv:2309.15459*, 2023.

[68] K. Zhang, B. Li, K. Hauser, and Y. Li, "Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation," in *RSS*, 2024.

[69] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *arXiv preprint arXiv:2304.13705*, 2023.

[70] T. Z. Zhao *et al.*, "Aloha unleashed: A simple recipe for robot dexterity," in *CoRL*, 2024.

[71] P. Zhi *et al.*, "Closed-loop open-vocabulary mobile manipulation with gpt-4v," *arXiv preprint arXiv:2404.10220*, 2024.

[72] C. Zhou, C. Peers, Y. Wan, R. Richardson, and D. Kanoulas, “Teleman: Teleoperation for legged robot loco-manipulation using wearable imu-based motion capture,” *arXiv preprint arXiv:2209.10314*, 2022.

[73] C. Zhou, C. C. Loy, and B. Dai, “Extract free dense labels from clip,” in *ECCV*, 2022.

[74] S. Zimmermann, R. Poranne, and S. Coros, “Go fetch!-dynamic grasps using boston dynamics spot with external robotic arm,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.