Humanoid Whole-Body Manipulation via Active Spatial Brain and Generalizable Action Cerebellum

Zhizhao Liang[1*], Yi-Lin Wei[1*], Xuhang Chen[1*], Mu Lin[1], Yi-Xiang He[1], Zhexi Luo[1], Jun-Hui Liu[1], Kun-Yu Lin[1], and Wei-Shi Zheng[1,2,3†]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China
[2] Shenzhen Loop Area Institute, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
https://leungchaos.github.io/Humanoid-WMAG/
liangzhzh26@mail2.sysu.edu.cn, wszheng@ieee.org

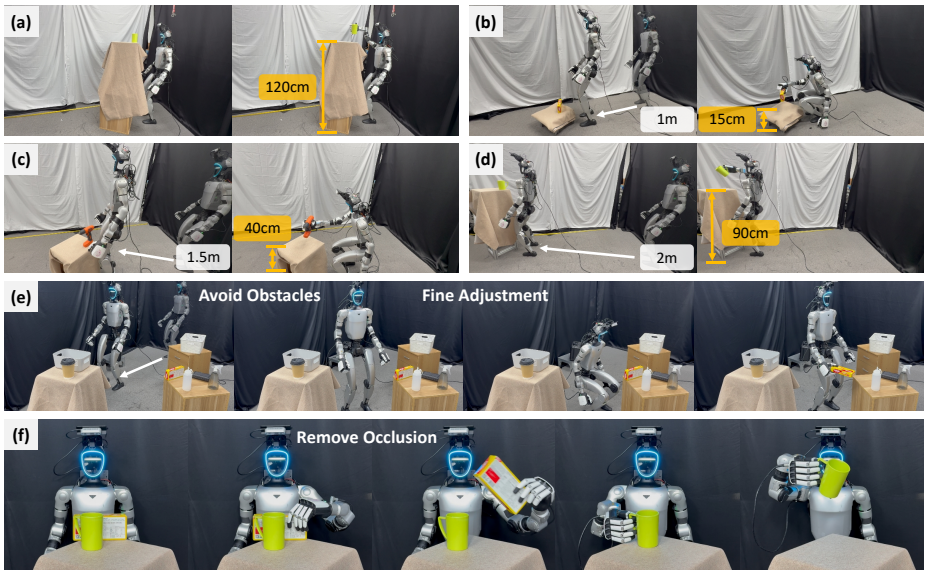


Fig. 1: This work enables generalizable humanoid whole-body manipulation in complex spatial environments through a multi-agent multimodal framework composed of an Active Spatial Brain and a Generalizable Action Cerebellum, without relying on task-specific data.

Abstract. In this paper, we explore spatial-aware humanoid whole-body manipulation task. Compared with tabletop settings, this task poses two key challenges: 1) Spatial understanding is challenging in complex 3D environments with diverse spatial relations. 2) Action generation

⋆ Equal contribution.
† Corresponding author.

is difficult to generalize, as limited and costly real-robot data restricts data-driven models generalization. To address these challenges, we propose a generalizable humanoid loco-manipulation framework that leverages the spatial perception and action generation capabilities of multi-agent large models. Specifically, our framework includes two components: Active Spatial Brain for active spatial perception and decision-making, and Generalizable Action Cerebellum for executable robot action generation. The first component actively perceives the spatial scene and makes decisions on task planning and subtask decomposition. The second component generates executable robot actions based on the decisions made by the first module without needs of task-specific real robot data. To benchmark our framework, we design a set of spatial manipulation tasks from two perspectives: evaluating spatial perception and understanding, and assessing real-robot task performance. The results demonstrate strong performance on both aspects across diverse tasks and environments.

Keywords: Humanoid · Loco-Manipulation · Active Perception · Embodied AI

1 Introduction

Enabling robots to autonomously execute complex tasks in spatial environments stands as a foundational yet challenging problem in robotics and computer vision communities. Among various embodiments, humanoid robots are particularly promising due to their whole-body mobility and dexterous manipulation capabilities, which allow them to operate in complex 3D spaces. Therefore humanoid whole-body manipulation is essential for real-world applications, including home assistance and industrial scenarios.

With the development of deep learning, data-driven works have achieved remarkable success in tabletop scenarios, evolving from early parallel-jaw grasping and manipulation [9, 24, 47, 53], to dexterous-hand grasping and manipulation [31, 50, 51, 55], and more recently to tabletop manipulation on humanoid embodiments [2, 4, 28, 41, 59], typically assuming fixed camera positions and constrained workspaces. Recent research extends these works to mobile manipulation platforms, including wheeled robots [11, 54, 60], quadrupeds [22, 38, 48], and humanoids [1, 7, 32]. Compared to tabletop settings and other robotic platforms, humanoid whole-body manipulation significantly enlarges both the observation and action spaces due to complex spatial environments and whole-body dexterity, resulting in much higher demands for large-scale, high-quality training data. Although prior efforts [12, 16, 29, 37, 58, 68, 70] attempt to mitigate this issue by leveraging human motion capture data, simulation data, or tabletop data, as well as editing existing data to augment training for better generalization [25, 57], these strategies still suffer from sim-to-real gaps, embodiment discrepancies, and spatial distribution shifts. As a result, learned policies exhibit limited robustness and poor generalization in diverse and unstructured real-world environments,

and even VLA models trained on diverse datasets are shown to struggle with cross-task generalization to unseen tasks [69].

To address these challenges, we propose a real-robot-data-free humanoid whole-body manipulation framework by leveraging the spatial intelligence and generalizable knowledge of current vision-language models (VLMs). Fig. 1 highlights the diverse tasks our framework can accomplish. The framework comprises two coupled components: the Active Spatial Brain and the Generalizable Action Cerebellum. Together, they form a closed-loop system where the Brain decomposes the goal into manageable sub-tasks and continuously invokes proper agents in Cerebellum to solve them step-by-step.

Specifically, the Active Spatial Brain operates fundamentally as an active spatial observer and dynamic planner. To enable active spatial perception, the humanoid is equipped with a camera capable of two degrees of freedom, whose viewpoint is dynamically controlled by the VLM for active spatial understanding. We further design a memory bank to archive visual and historical context. On top of these, we introduce an adaptive task planning module that integrates closed-loop spatial reasoning with long-horizon planning, decomposing high-level instructions into specific sub-task commands.

Building on the sub-task commands produced from the brain, the Generalizable Action Cerebellum addresses the challenge of how to generate executable humanoid action in a generalizable manner. To this end, we employ several robotic action agents by leveraging the ability of foundation models without the need of task-specific robot data. Specifically, we decouple the action into lower-body and upper-body. For the lower body, two agents are provided to move the robot to where the target object lies within the dexterous hand's workspace. For the upper body, we develop another two VLM-driven action agents that predict feasible dexterous grasp and corresponding arm motions to complete the manipulation.

To benchmark spatial-aware humanoid whole-body manipulation, we curate a set of real-world, whole-body tasks that explicitly require active spatial perception (e.g., resolving occlusions via viewpoint changes), 3D spatial reasoning (e.g., inferring relative object locations), and coordinated loco-manipulation (e.g., re-positioning the base to make targets reachable).On this benchmark,we first systematically evaluate representative VLMs in terms of complex spatial understanding. We then assess our proposed framework through real-robot experiments. The results show that our system combines reliable spatial perception and decision-making with generalizable, physically feasible whole-body action generation. Our framework achieves consistently higher success rates than existing data-driven baselines while requiring no real-robot data for training or fine-tuning.

2 Related Work

2.1 Whole-Body Manipulation

Humanoid whole-body manipulation is an active and challenging research area in embodied AI. To overcome the data bottleneck in learning-based methods, recent works utilize human teleoperation [1, 13, 29, 37, 63] for Imitation Learning [23, 56, 61, 64]. TrajBooster [32] retargets trajectories from wheeled robots to humanoids and integrates the teleoperation data for whole-body control. ResMimic [67] and R^2S^2 [66] employ task-specific residual learning or structural skill priors on top of human motion tracking. In the context of Reinforcement Learning, alongside unified controllers [45] and force-adaptive frameworks [65], AMO [27] combines sim-to-real RL with trajectory optimization, and VIRAL [14] uses large-scale visual distillation for zero-shot RL policy deployment in the real-world. However, most methods are only applied in simplified settings, lacking active environmental interaction and reasoning by the robot. Also, they are constrained by the costly real robot data or synthetic data, making it difficult to adapt to environmental changes and limit the generalization capabilities. In contrast, our approach leverages large models for active spatial perception. By dynamically understanding 3D scenes, our multi-agent architecture autonomously decomposes tasks into executable whole-body actions, achieving superior generalization across diverse environments.

2.2 LLMs-driven Manipulation

The emergent capabilities of Large Language Models (LLMs) in sematic reasoning and zero-shot generalization [3, 8, 26, 36] have driven their extensive application in robotic manipulation. Early works in this domain primarily utilized LLMs to translate semantic instructions into structured sub-goals sequences [19, 21]. To further bridge the semantic gap between planning and execution, researchers explored the synthesis of manipulation policies directly through LLMs' code generation ability [30, 43]. Upon this, recent works seek to enhance physical grounding by extracting geometric representations from the tabletop workspace [17,18,39,44]. These methods effectively convert language into spatial constraints and cost maps to guide the motion planning. While these approaches achieve promising manipulation capabilities, they are primarily designed for static tabletop scenarios with fixed camera perspectives. They inherently lack the active spatial reasoning ability and whole-body coordination mechanisms required to manage complex 3D scene-level environments and integrate locomotion and manipulation. By decoupling the problem into active spatial perception and multi-agent action generation, our system enables humanoids to dynamically reason the current scenarios and execute the whole-body action.

2.3 Spatial Intelligence for Robotics

Recent advances in robotic manipulation have increasingly focused on spatial intelligence to enable robust 3D perception, spatial reasoning, and precise inter-

action under complex real-world conditions. Existing methods can be generally divided into three categories. The first line [15, 40, 42] focuses on explicit spatial modeling and geometric-semantic disentanglement, which decouples and fuses geometric and semantic information to infer actionable 3D structures and poses for accurate manipulation. The second [5, 33, 35] adopts spatial-aware vision-language-action (VLA) frameworks, unifying spatial grounding, language understanding, and action generation in an end-to-end pipeline to improve generalization via embedded spatial reasoning. The third [6, 10, 62] exploits spatial affordance and keypoint prediction, often using lightweight augmentation or vision-language model fine-tuning to localize operable regions with low computational overhead. Although these efforts lay the foundation for advanced embodied intelligence, they are mostly limited to tabletop scenarios or simple loco-manipulation tasks. In contrast, this work targets spatial-aware humanoid whole-body manipulation, which integrates active spatial perception and complex decision-making by fully exploiting the strong spatial reasoning and decision-making capabilities of multimodal large language models.

3 Methods

3.1 Framework Overview

Problem Formulation In this paper, we focus on the humanoid whole-body manipulation task. Given user language commands, RGB-D observations, and the proprioceptive state, the framework aims to generate the whole-body action of the humanoid, denoted as $\mathcal{A} = \{a_{cam}, a_{dex}, a_{upper}, a_{lower}\}$, each referring to the active camera action a_{cam}, dexterous hand poses a_{dex}, upper-body joint poses a_{upper}, and lower-body joint poses a_{lower}. Finally, $\mathcal{A}$ is sent to PD controllers to output joint torques.

Architecture Our framework adopts a hierarchical multi-agent architecture for humanoid whole-body manipulation in complex environments, as illustrated in Fig. 2. It consists of two tightly coupled components: an Active Spatial Brain and a Generalizable Action Cerebellum. Operating as a high-level policy, the Brain actively explores the spatial environment, decomposes the goal into sub-tasks, and invokes specific action agents within the Cerebellum to execute low-level robot actions to complete them. Once action agents are invoked, feedback is routed back to the Brain. Together, these two components work in a perception-planning-control closed-loop until the goal is successfully completed.

3.2 Active Spatial Brain

To enable the humanoid robot to autonomously perceive complex 3D environments, understand natural language instructions and make manipulation planing, we propose the Active Spatial Brain empowered by the mainstream VLM.

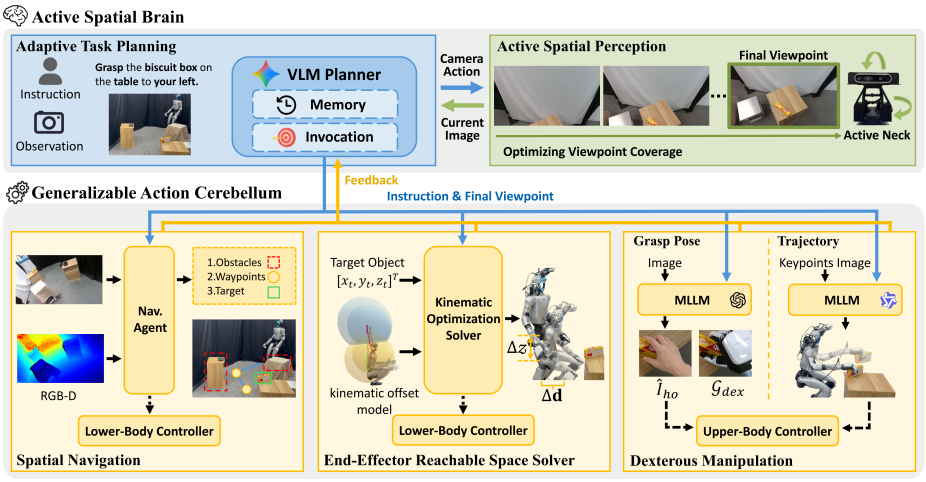


Fig. 2: The overview of our humanoid whole-body manipulation framework. Our framework consists of two components: an Active Spatial Brain for active spatial perception, understanding and planing; and a Generalizable Action Cerebellum for executable action generation.

The Brain integrates three modules: Active Spatial Perception for active surrounding perception, Memory Bank archiving perceived observations to support spatial awareness consistency, and Adaptive Task Planning leveraging them to generate spatially adaptive task plans.

Active Spatial Perception In complex real-world scenes, objects are frequently occluded or arranged in intricate spatial configurations, making fixed-perspective perception inadequate for accurate localization and manipulation. To address this challenge, we propose an active spatial perception agent to dynamically adjust the camera viewpoint to improve spatial understanding in unstructured 3D environments. We fully leverage the spatial reasoning capability of VLM to make viewpoint decisions directly from visual observations and task instructions. Specifically, we prompt the VLM with a pre-defined robot action space and its corresponding effects, where the space consists of a 2-DoF camera neck motion and a 4-DoF humanoid base movement including translations along the x, y, and z axes and z-axis rotation (Fig. 3). With this capability, the agent can actively explore complex environments to locate task-relevant targets.

Memory Bank In loco-manipulation, drastic shifts in the robot's base position can cause inherent viewpoint inconsistencies that mislead the VLM. To address this, we implement a hybrid recording movement-conditioned strategy. Specifically, the bank archives only the images captured by the perception module since the last base movement, pairing each with its corresponding camera pose.

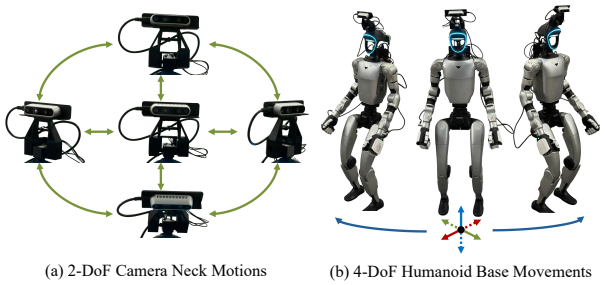


Fig. 3: The illustration of the degrees of freedom in the active camera, consisting of two parts: 2-DoF camera neck motions and 4-DoF camera base changes induced by humanoid body movements.

Coupled with constrained visuals, it maintains a textual log recording all commands dispatched from the Brain to the Cerebellum, alongside their respective feedback. By synthesizing these selectively archived images and the global text-based execution history, the framework enables the VLM to accurately infer the task stage and maintain consistent spatial awareness.

Adaptive Task Planning Building upon the active spatial perception module and the memory bank, the adaptive task planning agent integrates spatial reasoning, sub-task decomposition, and dynamic replanning. Given a user instruction, the agent first decomposes the long-horizon goal into a sequence of sub-tasks, followed by an iterative closed-loop execution. At each step, the planner evaluates the sequence against the memory bank to determine the current task stage and triggers one of two behaviors: (i) active viewpoint adjustment to achieve more comprehensive spatial perception, or (ii) invoking downstream agents in the Cerebellum to progress the overall plan.

However, the initial plan may be partially infeasible due to insufficient observations, and environmental disturbances can result in agent execution failures. To mitigate these issues, the planner dynamically replans its sub-tasks based on the memory bank at each step. For instance, upon detecting a failed grasp, it autonomously re-invokes the manipulation agent to execute a second attempt (Fig. 4). This unified replanning process achieves robust closed-loop control in complex, unstructured 3D environments.

3.3 Generalizable Action Cerebellum

To translate high-level commands from the Spatial Brain into executable and physically feasible humanoid action sequences in a generalizable manner, we introduce the Generalizable Action Cerebellum, which decomposes whole-body control into lower-body and upper-body control, executed by multiple specialized action agents. The lower-body locomotion module handles obstacle-aware

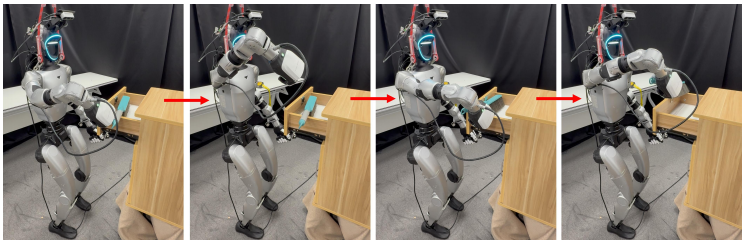


Fig. 4: The planner adjusts the plan via execution history and visual validation. Left to right: the robot misses the target, readjusts its pose, and successfully grasps it.

navigation and base adjustments to ensure end-effector reachability. The upper-body dexterous manipulation module generates physically feasible upper-body motions for manipulating the objects. By coordinating specialized action agents for both components without requiring task-specific demonstrations, the system achieves robust and generalizable whole-body humanoid manipulation.

3.3.1 Humanoid Lower-Body Locomotion

To achieve reliable and precise robot locomotion for subsequent manipulation, the lower-body locomotion module is built around two agents that can be invoked by the Brain based on perceived spatial relationships. In terms of their roles, a navigation agent manages long-range obstacle-aware coarse approach, while a reachable space solving agent provides short-range positioning to secure the object within the manipulation workspace. Beneath them, an RL policy is trained to convert the base movements from the agents into executable joint actions.

Spatial Navigation For obstacle-aware navigation, this agent requests an RGB-D observation which manifests the target along with the ground, and the 2D image coordinate of the target $\mathbf{u}_{\text{target}} = (u_{tar}, v_{tar})$. We first lift the depth map into a 3D point set in the camera frame. The resulting point cloud is then projected onto the ground plane to construct a discretized grid map with cell size d.

A grid cell is marked as occupied if there exists a point within the cell whose height exceeds a threshold h:

$$\mathcal{G}(i,j) = \begin{cases} 1, & \exists\, \mathbf{x} \in \mathcal{C}_{ij} \text{ s.t. } z(\mathbf{x}) > h, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The target coordinate $\mathbf{u}_{\text{target}}$ undergoes the same lifting and projection process to determine the corresponding goal cell in the grid map.

After constructing the grid map, we apply the A* algorithm to compute a collision-free path from the robot origin to the goal cell. The resulting path is

then simplified using the Ramer–Douglas–Peucker (RDP) algorithm, producing a waypoint sequence $\{\mathbf{p}_i\}_{i=1}^{N}$ that serves as the final trajectory. The robot tracks $\mathbf{p}_i$ using closed-loop feedback, computing $v_{x,t}$ and $\omega_{y,t}$ from the relative pose to the next waypoint while maintaining a nominal base height h_t.

End-Effector Reachable Space Solver To ensure the EE's reachability at close range, the robot should make precise adjustments to its base position. We therefore introduce an optimization based solver that computes the required base adjustment.

Given the target object position in the original robot's base frame $\mathbf{p}_t = [x_t, y_t, z_t]^T$, let $\mathbf{p}_{\text{obj}}(\mathbf{p}_t, \Delta\mathbf{d}, \Delta z)$ denote the target object position expressed in the adjusted base frame, where $\Delta\mathbf{d} \in \mathbb{R}^2$ and Δz represent the horizontal and vertical base adjustments. The solver minimizes the adjustment magnitude while bringing the object closer to the shoulder workspace:

$$\min_{\Delta\mathbf{d},\, \Delta z} \; \lambda_1 \|\Delta\mathbf{d}\|_2^2 + \lambda_2 (\Delta z)^2 + \lambda_3 \left\|\mathbf{p}_{\text{obj}}(\mathbf{p}_t, \Delta\mathbf{d}, \Delta z) - \mathbf{p}_{\text{shoulder}}(\Delta z)\right\|_2^2 . \tag{2}$$

To ensure reachability, we impose a constraint on the shoulder-to-object distance:

$$\left\|\mathbf{p}_{\text{obj}}(\mathbf{p}_t, \Delta\mathbf{d}, \Delta z) - \mathbf{p}_{\text{shoulder}}(\Delta z)\right\|_2 \leq \eta L, \tag{3}$$

where L is the arm length and $\eta = 0.8$ defines a safety margin.

Changing the base height alters the relative pose between the base and the shoulder due to the humanoid kinematic structure. To account for this effect, we estimate a kinematic offset model $\Phi(z)$ using MuJoCo [46] simulation:

$$\Phi(z) = (\delta x(z),\, \delta z(z)), \tag{4}$$

which predicts the horizontal and vertical offsets of the shoulder relative to the base at height z.

The shoulder position in the base frame is therefore

$$\mathbf{p}_{\text{shoulder}}(\Delta z) = [\delta x(z_0 + \Delta z),\; 0,\; \delta z(z_0 + \Delta z)]^T \tag{5}$$

where z_0 is the nominal base height. Solving this optimization yields the refined base displacement and height that place the object within a safe and reachable region for subsequent manipulation. We then compute the $v_{x,t}$, $\omega_{y,t}$, h_t corresponding to Δd and Δz for the RL policy.

Reinforcement Learning-Based Humanoid Locomotion We finally adopt a policy π_{loco} following HOMIE [1] to translate the base commands from agents above to actual lower-body joints actions. Given the proprioception state $\mathbf{s}_t$ and base command $\mathbf{c}_t = [v_{x,t}, \omega_{y,t}, h_t]$, the policy outputs $\mathbf{a}_{\text{lower},t} = \pi_{\text{loco}}(\mathbf{s}_t, \mathbf{c}_t)$ to achieve stable lower-body locomotion.

3.3.2 Humanoid Upper-Body Dexterous Manipulation

Humanoid upper-body dexterous manipulation requires precise control of both arm and dexterous hand joints, posing challenges due to the high degrees of freedom and limited generalization across tasks and objects. To address these challenges, we decompose manipulation into grasp generation and post-grasp trajectory generation implemented by two agents empowered by MLLM. One synthesizes object-specific grasp poses, while the other one produces manipulation trajectories after grasp acquisition.

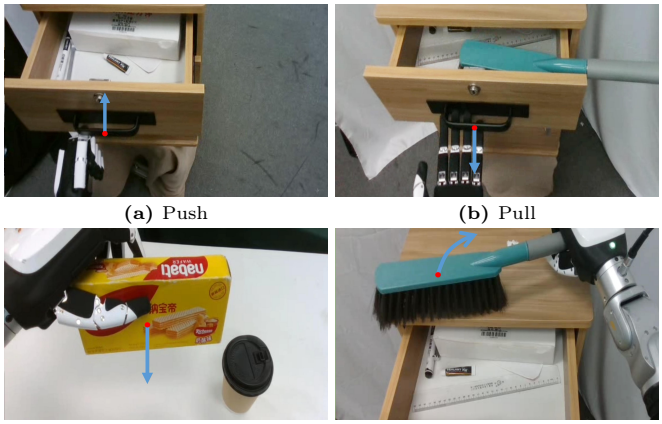
Grasp Pose Generation Inspired by the paradigm of leveraging foundation models to generate human grasp images and transferring them into executable robot actions [52], the grasp pose generation agent leverages the multimodal foundation generative models to synthesize hand–object interaction images $\hat{I}_{ho}$ conditioned on the visual observation I and task instruction $\mathcal{L}$. By harnessing the foundation model's ability to learn generalizable grasp patterns from large-scale data, these images encode transferable human grasp strategies that serve as a strong prior for subsequent robot grasp generation, even though they cannot be executed directly due to embodiment and physical constraints.

To bridge this gap, the agent performs a human-image-to-robot grasp transfer via hand–object reconstruction and dexterous retargeting. In the hand–object reconstruction stage, the agent first invokes the pretrained Hyper3D [20] model to generate a 3D object mesh $\mathcal{M}_o$ from the observation I. Given $\mathcal{M}_o$ and $\hat{I}_{ho}$, the agent then applies EasyHOI [34] to estimate the human hand pose $\mathcal{G}_{mano}$. The reconstructed human grasp is then retargeted to the dexterous hand by aligning the fingertip positions $p_k^{dex,ft}$ with the corresponding human fingertip positions $p_k^{mano,ft}$, and the optimization objective is formulated as:

$$\min_{\mathcal{G}_{dex}} \sum_k \left\| p_k^{dex,ft} - p_k^{mano,ft} \right\|_2^2. \tag{6}$$

The optimized $p_k^{dex,ft}$ are then retargeted to dexterous hand actions $\mathcal{G}_{dex}$ as the hand pose output.

Post-grasp Trajectory Generation This agent identifies the post-grasp trajectory through parameterized action primitives (Fig. 5). Given a visual observation and language instruction, the large model predicts semantic 2D keypoints and selects a corresponding action primitive. After lifting these keypoints to 3D coordinates, we derive the target 6-DoF EE poses by applying the primitive's specific kinematic rules. All primitives begin with an alignment phase, moving the EE to a target keypoint or preparatory pose. Subsequently, *Pushing* (Fig. 5a) and *Pulling* (Fig. 5b) apply linear translation to manipulate articulated objects like drawers. *Placing* (Fig. 5c) preserves the initial hand-object spatial offset at the new location, while *Rotating* (Fig. 5d) executes a pivot-based local rotation. Finally, an IK solver converts these sequential poses into continuous joint commands for whole-body execution.



(a) Push **(b)** Pull

(c) Place **(d)** Rotate

Fig. 5: Implementation of fundamental manipulation primitives. Red dots denote target spatial keypoints, and blue arrows indicate trajectory directions.

4 Experiments

4.1 Spatial Intelligence Benchmark for Manipulation Tasks

Spatial intelligence constitutes one of the cores of our framework. To investigate the spatial intelligence capabilities of mainstream VLMs, we systematically evaluate them through several tasks. These tasks assess spatial relationship understanding, 3D scene modeling, and active perception, respectively.

Task 1 Map a shuffled set of nine adjacent images back to their original 3×3 spatial grid, where proficiency is quantified by structural matching accuracy.

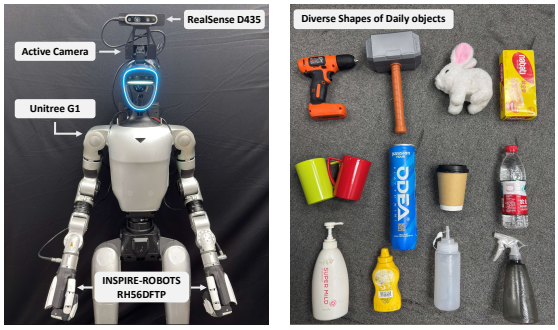
Task 2 The VLM drives the active camera to locate an initially out-of-view or occluded target, continuing until successful detection or running out budget.

Task 3 Predict a sequence of ground waypoints leading to the target within an obstacle-cluttered scene. Trajectory formed by waypoints is graded by obstacle clearance as Appropriate (moderate distance), Inappropriate (excessively close/far) or Fail (collision).

4.2 Real World Experiments Setting

Robot Platform We conduct real world experiments on a Unitree G1 humanoid equipped with two 6-DoF Inspire FTP hands and a 2-DoF active Realsense D435i camera (Fig. 6(a)). The set of objects used for manipulation is shown in Fig. 6(b).

Task Description To evaluate the proposed humanoid mobile manipulation framework under diverse conditions, we design five task settings. Each setting includes an *easy* and a *hard* mode.



(a) Hardware Setup (b) Objects

Fig. 6: Real-world experimental setup. (a) The robot hardware platform. (b) The objects to be grasped in our experiments.

1. **Different Heights.** The robot performs manipulation without base translation while adjusting its body height. The easy mode uses moderate heights, while the hard mode includes more extreme height variations.

2. **Different Positions.** The robot maintains a fixed height but navigates to the target without obstacles. The easy mode requires short-range motion, whereas the hard mode involves longer distances.

3. **Different Heights and Positions.** The robot must adjust both height and position. The easy mode corresponds to the intersection of the easy cases above, and the hard mode to the union of their hard cases.

4. **Obstacle Avoidance.** The robot avoids obstacles during navigation, with one obstacle in the easy mode and multiple obstacles in the hard mode.

5. **Occlusion Handling.** The robot removes occluding objects before manipulating the target. The easy mode uses a fixed occluder, while the hard mode includes diverse occluding objects.

For each setting, we conduct 10–30 trials and report the success rate, ensuring that the number of trials is identical across methods for fair comparison. A trial is considered successful only if the robot completes the manipulation task without any collision with obstacles.

4.3 Can our framework surpass data-driven methods?

Reproduction of Data-driven Method. We select TrajBooster [32] and Ψ_0 [49] as the learning-based baseline for comparison. We initialize the model with the officially released post-pre-trained weights and fine-tune it on our task settings. For each setting, we collect at least 20 trajectories under the easy configuration for post-training. The fine-tuned model is then evaluated on both the easy (in-domain) and hard (out-of-domain) configurations to assess its generalization performance.

Table 1: Comparison with data-driven (TrajBooster, Ψ_0) and VLM-driven (CaP) baselines on diverse tasks. Easy denotes in-domain tasks, while Hard denotes more complex out-of-distribution tasks.

Method	Task 1	Task 2	Task 3	Task 4	Task 5
Easy Setting					
TrajBooster	70.0	**75.0**	55.0	60.0	40.0
Ψ_0	80.0	**75.0**	70.0	70.0	60.0
CaP	75.0	70.0	70.0	**80.0**	65.0
Ours	**85.0**	**75.0**	**80.0**	**80.0**	**70.0**
Hard Setting					
TrajBooster	20.0	40.0	10.0	0	20.0
Ψ_0	35.0	45.0	25.0	0	30.0
CaP	50.0	55.0	35.0	30.0	35.0
Ours	**60.0**	**70.0**	**55.0**	**60.0**	**60.0**

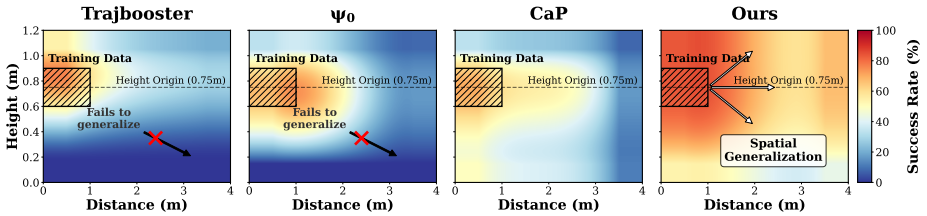


Fig. 7: Spatial reachability heatmaps across methods. The heatmap visualizes the reachable regions under different target locations, showing that our framework maintains more robust reachability than the data-driven baseline.

Comparison Results. As shown in Table 1, our framework consistently outperforms both data-driven baselines (TrajBooster, Ψ_0) and the VLM-driven baseline (CaP [30]) on spatial humanoid manipulation tasks. To analyze performance under different difficulty levels, we divide the tasks into easy and hard settings. The easy setting involves simple spatial relationships and limited generalization demands, whereas the hard setting requires stronger spatial reasoning and broader generalization in complex environments. We observe that data-driven methods remain competitive on the easy tasks, but their performance drops significantly on the hard tasks. This is likely because such methods depend heavily on the quality and diversity of training data, making them less robust to out-of-distribution scenarios and complex spatial relationships not covered during training. Notably, the performance gap between the easy and hard settings is far more pronounced for the data-driven baselines than for our framework, indicating that their learned policies are tightly coupled to the training distribution. Additionally, existing VLM methods (e.g., CaP) generalize well but underperform on humanoids without specific designs. In contrast, our framework maintains strong performance in both settings, demonstrating better generalization through the multi-agent design.

4.4 Can our framework achieve generalization?

We further analyze the generalization capability of our framework from two complementary perspectives: spatial reachability and object categories. Fig. 7 visualizes the reachable regions across different target locations, showing that our framework maintains a broader and more reliable reachable region than the data-driven baseline. Table 2 further demonstrates strong generalization across both seen and unseen object categories. In contrast, data-driven methods perform reasonably well when the test scenarios are similar to the training domain, but their performance degrades significantly when encountering unseen spatial layouts or novel objects. These results highlight the advantage of our approach in handling both reachability variations in space and category variations in real-world environments.

Table 2: Generalization across object categories. * indicates objects unseen in the fine-tuning dataset of TrajBooster.

Objects	Ours	Traj.
Tools	**77.8**	50.0
Tools*	**52.0**	50.0
Items	**60.0**	**60.0**
Items*	**69.6**	50.0

4.5 Is each component of our framework effective?

We conduct an ablation study to evaluate the key modules: Active Perception (AP) and End-Effector Reachable Space Solver (EES). In detail, Robot's viewpoint is fixed without AP, and it maintains a predefined distance and relative height difference with the target object without EES. Shown in Table 3, Task 3 performance is consistently low without AP, confirming that AP significantly expands the robot's observation space to maintain visual context during long-distance base movements. Success rates all drop without EES in the experiment, especially in Task 4 where sensor errors accumulated during non-linear movement, showing that EES further improves manipulation by refining object alignment and positioning. Results demonstrate that our design effectively improves humanoid ability in whole-body manipulation tasks.

Table 3: Ablation study of our framework on two representative tasks.

AP	EES	Diff. Heights and Positions	Obstacle Avoidance
×	×	18.8	0
✓	×	55.6	16.7
×	✓	24.9	33.4
✓	✓	**67.5**	**70.0**

Table 4: Spatial intelligence benchmark of VLMs. Time: average seconds per task.

Models	Correctness	Time
gpt5.1	94.12	94.59
gemini-3-flash	95.21	**54.93**
gemini-3.1-pro	**98.84**	94.50
qwen3.5-plus	93.29	155.63

(a) Map images back to grid

Models	Success	Observe Times	Time
gpt5.1	82.35	2.53	45.75
gemini-3-flash	88.23	**2.00**	**27.07**
gemini-3.1-pro	**94.11**	**2.00**	44.73
qwen3.5-plus	70.00	4.80	32.15

(b) Active perception

Models	Appropriate	Inappropriate	Fail	Time
gpt5.1	30.35	25.00	44.64	118.38
gemini-3-flash	57.89	26.31	**15.78**	33.76
gemini-3.1-pro	**64.91**	**14.03**	21.05	35.12
qwen3.5-plus	17.54	15.78	66.67	**33.03**

(c) Waypoints generation

4.6 How is the spatial intelligence of VLMs?

Table 4 highlights a distinct spatial intelligence gap in current VLMs. While they excel at coarse, relational spatial reasoning (Tasks 1 and 2), they struggle with fine-grained metric perception from 2D observations (Task 3). This validates our framework: by decoupling VLM-based high-level planning from fine-grained action generation, we leverage their perceptual strengths while ensuring precise, generalizable locomotion and manipulation.

5 Conclusion

In this work, we study spatial-aware humanoid whole-body manipulation in complex 3D environments, which requires strong spatial understanding and robust action generalization. To address these challenges, we propose a generalizable humanoid loco-manipulation framework composed of two components: an Active Spatial Brain for active spatial perception, task planning, and subtask decomposition, and a Generalizable Action Cerebellum for executable action generation without task-specific real-robot data. Experiments on diverse spatial manipulation tasks demonstrate strong performance in both spatial perception and real-robot task execution across varied environments. Our results highlight the potential of integrating large-model-driven active spatial perception and decision making for generalizable humanoid manipulation.

Acknowledgements

This work was supported partially by NSFC(92470202), Guangdong NSF Project (No. 2023B1515040025), Guangdong Key Research and Development Program (No.2024B0101040004, No. 2025B0909020002).

References

1. Ben, Q., Jia, F., Zeng, J., Dong, J., Lin, D., Pang, J.: Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. arXiv preprint arXiv:2502.13013 (2025)
2. Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 (2025)
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (2020)
4. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv:2503.06669 (2025)
5. Contributors, I.M.: Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. arXiv preprint arXiv:2510.13778 (2025)
6. Dai, Y., Lee, J., Zhang, Y., Ma, Z., Yang, J., Zadeh, A., Li, C., Fazeli, N., Chai, J.: Aimbot: A simple auxiliary visual cue to enhance spatial awareness of visuomotor policies. Conference on Robot Learning (2025)
7. Ding, P., Ma, J., Tong, X., Zou, B., Luo, X., Fan, Y., Wang, T., Lu, H., Mo, P., Liu, J., et al.: Humanoid-vla: Towards universal humanoid control with visual integration. arXiv preprint arXiv:2502.14795 (2025)
8. Dong, Y., Jiang, X., Qian, J., Wang, T., Zhang, K., Jin, Z., Li, G.: A survey on code generation with llm-based agents. arXiv preprint arXiv:2508.00083 (2025)
9. Fang, H.S., Wang, C., Gou, M., Lu, C.: Graspnet-1billion: A large-scale benchmark for general object grasping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11444–11453 (2020)
10. Fang, K., Liu, F., Abbeel, P., Levine, S.: Moka: Open-world robotic manipulation through mark-based visual prompting. Robotics: Science and Systems (2024)
11. Fu, Z., Zhao, T.Z., Finn, C.: Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. arXiv preprint arXiv:2401.02117 (2024)
12. He, T., Luo, Z., He, X., Xiao, W., Zhang, C., Zhang, W., Kitani, K.M., Liu, C., Shi, G.: Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In: Conference on Robot Learning (2024)
13. He, T., Luo, Z., Xiao, W., Zhang, C., Kitani, K., Liu, C., Shi, G.: Learning human-to-humanoid real-time whole-body teleoperation. In: IEEE International Conference on Intelligent Robots and Systems (2024)
14. He, T., Wang, Z., Xue, H., Ben, Q., Luo, Z., Xiao, W., Yuan, Y., Da, X., Castañeda, F., Sastry, S., et al.: Viral: Visual sim-to-real at scale for humanoid loco-manipulation. arXiv preprint arXiv:2511.15200 (2025)
15. Huang, H., Lin, F., Hu, Y., Wang, S., Gao, Y.: Copa: General robotic manipulation through spatial constraints of parts with foundation models. In: IEEE International Conference on Intelligent Robots and Systems (2024)
16. Huang, W.J., Zhang, Y.Y., Wei, Y.L., Xia, Z.W., Tan, J., Li, Y.M., Zhao, Z., Zheng, W.S.: Beyond mimicry: Learning whole-body human-humanoid interaction from human-human demonstrations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30740–30749 (2026)

17. Huang, W., Wang, C., Li, Y., Zhang, R., Fei-Fei, L.: Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In: Conference on Robot Learning (2024)
18. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models. In: Conference on Robot Learning (2023)
19. Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., Sermanet, P., Jackson, T., Brown, N., Luu, L., Levine, S., Hausman, K., Ichter, B.: Inner monologue: Embodied reasoning through planning with language models. In: Conference on Robot Learning (2022)
20. Hyper3D: Hyper3d: Ai-powered 3d model generator (2024), `https://hyper3d.ai/`, accessed: 2026-02-28
21. Ichter, B., Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., Kalashnikov, D., Levine, S., Lu, Y., Parada, C., Rao, K., Sermanet, P., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Yan, M., Brown, N., Ahn, M., Cortes, O., Sievers, N., Tan, C., Xu, S., Reyes, D., Rettinghouse, J., Quiambao, J., Pastor, P., Luu, L., Lee, K., Kuang, Y., Jesmonth, S., Joshi, N.J., Jeffrey, K., Ruano, R.J., Hsu, J., Gopalakrishnan, K., David, B., Zeng, A., Fu, C.K.: Do as I can, not as I say: Grounding language in robotic affordances. In: Conference on Robot Learning (2022)
22. Jeon, S., Jung, M., Choi, S., Kim, B., Hwangbo, J.: Learning whole-body manipulation for quadrupedal robot. IEEE Robotics and Automation Letters (2023)
23. Jiang, H., Chen, J., Bu, Q., Chen, L., Shi, M., Zhang, Y., Li, D., Suo, C., Wang, C., Peng, Z., et al.: Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control. arXiv preprint arXiv:2512.11047 (2025)
24. Jiang, J.J., Wu, X.M., He, Y.X., Zeng, L.A., Wei, Y.L., Zhang, D., Zheng, W.S.: Rethinking bimanual robotic manipulation: Learning with decoupled interaction framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12427–12437 (2025)
25. Jiang, J.J., Yang, Y., Wei, L., Luo, Y., Wu, X.M., Chen, X., Fan, B., Zhang, D., Zheng, W.S.: Task editing for generalizable 3d visuomotor policy learning. arXiv preprint arXiv:2606.07012 (2026)
26. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Advances in Neural Information Processing Systems (2022)
27. Li, J., Cheng, X., Huang, T., Yang, S., Qiu, R.Z., Wang, X.: Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control. arXiv preprint arXiv:2505.03738 (2025)
28. Li, J., Zhu, Y., Xie, Y., Jiang, Z., Seo, M., Pavlakos, G., Zhu, Y.: Okami: Teaching humanoid robots manipulation skills through single video imitation. In: Conference on Robot Learning (2024)
29. Li, Y., Lin, Y., Cui, J., Liu, T., Liang, W., Zhu, Y., Huang, S.: Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In: Conference on Robot Learning (2025)
30. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as policies: Language model programs for embodied control. In: IEEE International Conference on Robotics and Automation (2023)
31. Lin, M., Wei, Y.L., Chen, J., Lin, Y., Chen, S., Lyu, J., Chen, J., Tang, Y., Wang, H., Zheng, W.S.: Bidexgrasp: Coordinated bimanual dexterous grasps across object geometries and sizes. arXiv preprint arXiv:2604.06589 (2026)

32. Liu, J., Ding, P., Zhou, Q., Wu, Y., Huang, D., Peng, Z., Xiao, W., Zhang, W., Yang, L., Lu, C., et al.: Trajbooster: Boosting humanoid whole-body manipulation via trajectory-centric learning. arXiv preprint arXiv:2509.11839 (2025)
33. Liu, Y., Liu, Y., Meng, Y., Zhang, J., Zhou, Y., Li, Y., Jiang, J., Ji, K., Ge, S., Wang, Z., et al.: Spatial policy: Guiding visuomotor robotic manipulation with spatial-aware modeling and reasoning. arXiv preprint arXiv:2508.15874 (2025)
34. Liu, Y., Long, X., Yang, Z., Liu, Y., Habermann, M., Theobalt, C., Ma, Y., Wang, W.: Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7037–7047 (2025)
35. Liu, Z., Gu, Y., Wang, Y., Xue, X., Fu, Y.: Activevla: Injecting active perception into vision-language-action models for precise 3d robotic manipulation. arXiv preprint arXiv:2601.08325 (2026)
36. Madaan, A., Zhou, S., Alon, U., Yang, Y., Neubig, G.: Language models of code are few-shot commonsense learners. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2022)
37. Nai, R., Zheng, B., Zhao, J., Zhu, H., Dai, S., Chen, Z., Hu, Y., Hu, Y., Zhang, T., Wen, C., et al.: Humanoid manipulation interface: Humanoid whole-body manipulation from robot-free demonstrations. arXiv preprint arXiv:2602.06643 (2026)
38. Niu, Y., Zhang, Y., Yu, M., Lin, C., Li, C., Wang, Y., Yang, Y., Yu, W., Zhang, T., Li, Z., et al.: Human2locoman: Learning versatile quadrupedal manipulation with human pretraining. arXiv preprint arXiv:2506.16475 (2025)
39. Pan, M., Zhang, J., Wu, T., Zhao, Y., Gao, W., Dong, H.: Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
40. Qi, Z., Zhang, W., Ding, Y., Dong, R., Yu, X., Li, J., Xu, L., Li, B., He, X., Fan, G., et al.: Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. In: Advances in Neural Information Processing Systems (2025)
41. Qiu, R.Z., Yang, S., Cheng, X., Chawla, C., Li, J., He, T., Yan, G., Yoon, D.J., Hoque, R., Paulsen, L., et al.: Humanoid policy˜ human policy. In: Conference on Robot Learning (2025)
42. Shi, H., Xie, B., Liu, Y., Yue, Y., Wang, T., Fan, H., Zhang, X., Huang, G.: Spatialactor: Exploring disentangled spatial representations for robust robotic manipulation. arXiv preprint arXiv:2511.09555 (2025)
43. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A.: Progprompt: Generating situated robot task plans using large language models. In: IEEE International Conference on Robotics and Automation (2023)
44. Su, C., Shang, W., Qian, C., Zhang, F., Cong, S.: Resemact: Advancing fine-grained robotic manipulation via semantic structuring and affordance refinement. arXiv preprint arXiv:2507.18262 (2025)
45. Sun, W., Feng, L., Cao, B., Liu, Y., Jin, Y., Xie, Z.: Ulc: A unified and fine-grained controller for humanoid loco-manipulation. arXiv preprint arXiv:2507.06905 (2025)
46. Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 5026–5033. IEEE (2012)
47. Wang, A.L., Chen, N., Lin, K.Y., Li, Y.M., Zheng, W.S.: Task-oriented 6-dof grasp pose detection in clutters. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 5692–5698. IEEE (2025)

48. Wang, J., Rajabov, J., Xu, C., Zheng, Y., Wang, H.: Quadwbg: Generalizable quadrupedal whole-body grasping. In: IEEE International Conference on Robotics and Automation (2025)
49. Wei, S., Jing, H., Li, B., Zhao, Z., Mao, J., Ni, Z., He, S., Liu, J., Liu, X., Kang, K., Zang, S., Yuan, W., Pavone, M., Huang, D., Wang, Y.: ψ_0: An open foundation model towards universal humanoid loco-manipulation. In: Robotics: Science and Systems (2026)
50. Wei, Y.L., Jiang, J.J., Xing, C., Tan, X.T., Wu, X.M., Li, H., Cutkosky, M., Zheng, W.S.: Grasp as you say: Language-guided dexterous grasp generation. Advances in Neural Information Processing Systems **37**, 46881–46907 (2024)
51. Wei, Y.L., Lin, M., Lin, Y., Jiang, J.J., Wu, X.M., Zeng, L.A., Zheng, W.S.: Affordexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11818–11828 (2025)
52. Wei, Y.L., Luo, Z., Lin, Y., Lin, M., Liang, Z., Chen, S., Zheng, W.S.: Omnidexgrasp: Generalizable dexterous grasping via foundation model and force feedback. In: IEEE International Conference on Robotics and Automation (ICRA) (2026)
53. Wu, X.M., Cai, J.F., Jiang, J.J., Zheng, D., Wei, Y.L., Zheng, W.S.: An economic framework for 6-dof grasp detection. In: European Conference on Computer Vision. pp. 357–375. Springer (2024)
54. Wu, Z., Zhou, Y., Xu, X., Wang, Z., Yan, H.: Momanipvla: Transferring vision-language-action models for general mobile manipulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
55. Xu, G.H., Wei, Y.L., Zheng, D., Wu, X.M., Zheng, W.S.: Dexterous grasp transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17933–17942 (2024)
56. Xue, H., Huang, X., Niu, D., Liao, Q., Kragerud, T., Gravdahl, J.T., Peng, X.B., Shi, G., Darrell, T., Sreenath, K., et al.: Leverb: Humanoid whole-body control with latent vision-language instruction. arXiv preprint arXiv:2506.13751 (2025)
57. Xue, Z., Deng, S., Chen, Z., Wang, Y., Yuan, Z., Xu, H.: Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning. arXiv preprint arXiv:2502.16932 (2025)
58. Yang, L., Huang, X., Wu, Z., Kanazawa, A., Abbeel, P., Sferrazza, C., Liu, C.K., Duan, R., Shi, G.: Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. arXiv preprint arXiv:2509.26633 (2025)
59. Yang, R., Yu, Q., Wu, Y., Yan, R., Li, B., Cheng, A.C., Zou, X., Fang, Y., Cheng, X., Qiu, R.Z., et al.: Egovla: Learning vision-language-action models from egocentric human videos. arXiv preprint arXiv:2507.12440 (2025)
60. Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A.S., Khanna, M., Gervet, T., Yang, T.Y., Jain, V., Clegg, A., Turner, J.M., et al.: Homerobot: Open-vocabulary mobile manipulation. In: Annual Conference on Robot Learning (2023)
61. Yuan, H., Bai, Y., Fu, Y., Zhou, B., Feng, Y., Xu, X., Zhan, Y., Karlsson, B.F., Lu, Z.: Being-0: A humanoid robotic agent with vision-language models and modular skills. arXiv preprint arXiv:2503.12533 (2025)
62. Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., Fox, D.: Robopoint: A vision-language model for spatial affordance prediction for robotics. arXiv preprint arXiv:2406.10721 (2024)
63. Ze, Y., Chen, Z., Araujo, J.P., Cao, Z.a., Peng, X.B., Wu, J., Liu, K.: Twist: Teleoperated whole-body imitation system. In: Conference on Robot Learning (2025)

64. Ze, Y., Chen, Z., Wang, W., Chen, T., He, X., Yuan, Y., Peng, X.B., Wu, J.: Generalizable humanoid manipulation with 3d diffusion policies. In: IEEE International Conference on Intelligent Robots and Systems (2025)
65. Zhang, Y., Yuan, Y., Gurunath, P., Gupta, I., Omidshafiei, S., Agha-mohammadi, A.a., Vazquez-Chanlatte, M., Pedersen, L., He, T., Shi, G.: Falcon: Learning force-adaptive humanoid loco-manipulation. arXiv preprint arXiv:2505.06776 (2025)
66. Zhang, Z., Chen, C., Xue, H., Wang, J., Liang, S., Liu, Y., Zhang, Z., Wang, H., Yi, L.: Unleashing humanoid reaching potential via real-world-ready skill space. IEEE Robotics and Automation Letters (2025)
67. Zhao, S., Ze, Y., Wang, Y., Liu, C.K., Abbeel, P., Shi, G., Duan, R.: Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. arXiv preprint arXiv:2510.05070 (2025)
68. Zhou, J., Ma, T., Lin, K.Y., Wang, Z., Qiu, R., Liang, J.: Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation. In: Proceedings of the computer vision and pattern recognition conference. pp. 22551–22561 (2025)
69. Zhou, J., Ye, K., Liu, J., Ma, T., Wang, Z., Qiu, R., Lin, K.Y., Zhao, Z., Liang, J.: Exploring the limits of vision-language-action manipulation in cross-task generalization. Advances in Neural Information Processing Systems **38**, 139899–139927 (2026)
70. Zhu, T., Cai, G., Zhaohui, Y., Ren, G., Xie, H., Wang, Z., Wu, J., Wang, J., Yang, X., Mu, Y., et al.: Clot: Closed-loop global motion tracking for whole-body humanoid teleoperation. arXiv preprint arXiv:2602.15060 (2026)