Learning Humanoid Loco-Manipulation for Transporting Unwieldy Construction Objects

Zaolin Pan[1,3], Yantao Yu[1,2], and Zhengbo Zou[3]

[1] Dept. of Civil and Environmental Engineering, HKUST, Hong Kong, China
[2] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China
[3] Dept. of Civil Engineering and Engineering Mechanics, Columbia Univ., New York, NY, USA
zpanaq@connect.ust.hk, ceyantao@ust.hk, zhengbo.zou@columbia.edu

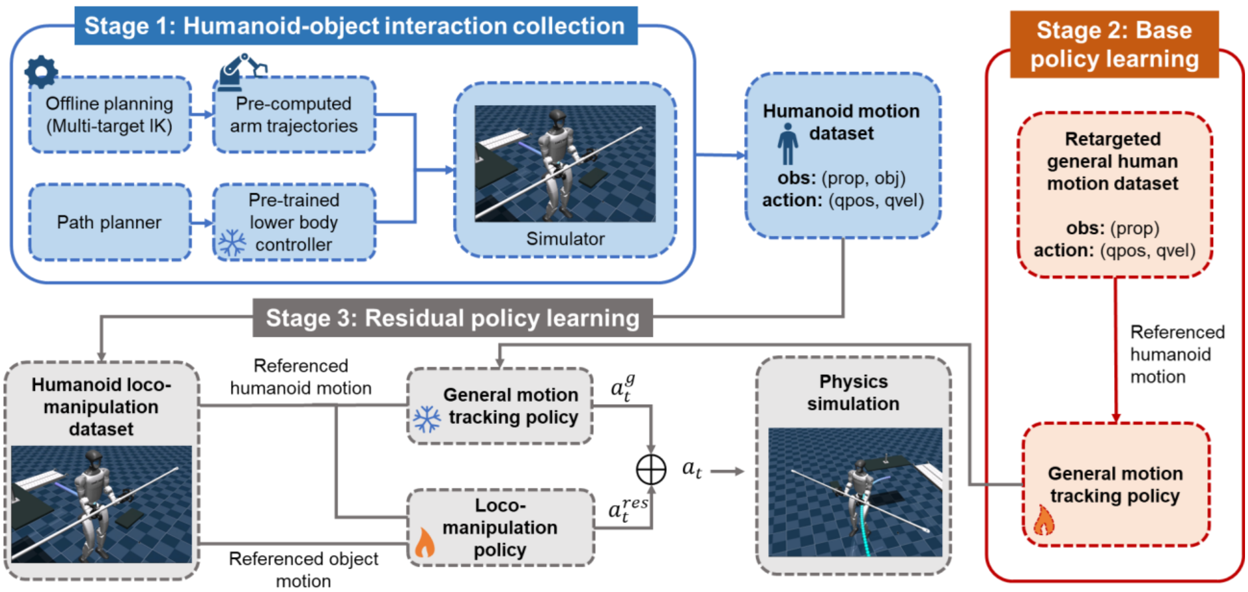

Figure 1 Proposed three-stage residual learning-based framework.

Abstract –

Enabling robust loco-manipulation for transporting unwieldy objects is a critical prerequisite for deploying humanoid robots in construction. However, this task is complicated by severe physical constraints, including irregular object geometries, limited actuator torque, and high moments of inertia caused by eccentric centers of mass. These factors induce complex robot-object dynamics that threaten locomotion stability—a challenge further exacerbated by the scarcity of domain-specific interaction data. Unlike previous approaches that rely on either large-scale motion capture data or torso-supported strategies ("hugging motion") to carry large objects, our method introduces (1) a scalable pipeline for synthesizing high-fidelity demonstrations in simulation and (2) a residual policy that refines locomotion priors for dexterous end-effector-based transportation. Experimental results demonstrate that our residual policy successfully transports unwieldy objects (up to 5 kg) while maintaining balance, outperforming both learning-from-scratch and fine-tuning baselines. This study serves as an initial exploration into equipping humanoids with the complex loco-manipulation skills required for construction sites.

Keywords –

Humanoid robots; Loco-manipulation; Construction robotics; Reinforcement learning; Whole-body control

1 Introduction

Humanoid robots hold great potential for deployment on construction sites, leveraging their flexibility and human-like morphology to navigate workplaces and perform diverse tasks. Demonstrations such as Boston

Dynamics' Atlas [1] and NASA's Valkyrie [2] have validated the utility of humanoids in construction through tasks like tool retrieval and scaffold frame assembly. However, achieving robust humanoid loco-manipulation in unstructured construction environments remains a formidable challenge. Key difficulties, such as robots' high-dimensional action space, the necessity to manipulate arbitrarily shaped objects, and the extended planning horizons required for construction tasks, render the learning of reliable whole-body loco-manipulation policies exceptionally difficult.

In this paper, we focus on a critical aspect of humanoid loco-manipulation: unwieldy object transportation. An unwieldy object is a physical load characterized by irregular geometry, large dimensions, and significant mass related to robot payload, such that its transportation introduces dynamic robot-object motion coupling that destabilizes the robot's balance. Given that construction sites are replete with unwieldy objects, such as scaffold frames, cross braces, and walk boards for scaffold assembly, this capability is a fundamental prerequisite for advancing robotic construction and providing effective support to human partners.

However, critical gaps remain in both the construction and robotics domains. In the construction context, previous research on collaborative robots has mainly relied on static robotic arms [3] or wheeled platforms [4], which bypass the complexities of whole-body balancing. Conversely, while recent robotics research has demonstrated humanoid loco-manipulation [5,6], these studies predominantly focus on compact everyday objects, such as scissors, hammers, or cups, whose small size and negligible mass have minimal impact on the robot's dynamics. Although some studies [7,8] have investigated larger objects like boxes or chairs, they typically rely on torso-supported strategies, which stabilize loads against the body via "hugging" rather than utilizing dexterous end-effector manipulation. Such strategies, however, are ill-suited for construction objects of high aspect ratio, such as tubes or scaffold frames. Hugging these objects introduces severe kinematic constraints, as the load often intersects with the robot's leg workspace during locomotion. Furthermore, the sparse geometry of these objects provides insufficient contact surface for stable friction, while the extended length creates large lever arms that magnify external disturbances, rendering torso-based friction insufficient for stabilization.

Beyond these physical dynamics, acquiring the necessary training data presents a separate bottleneck. We lack scalable demonstrations for construction objects because traditional Motion Capture (MoCap) methods—while accurate—are expensive and difficult to deploy for the large-scale interactions required in this domain.

To address these problems, we propose a residual learning framework, as illustrated in Fig. 1. First, we introduce a scalable simulation-based method for synthesizing physically plausible interaction data. By modeling the object with negligible mass, we can employ a pretrained locomotion controller and bimanual Inverse Kinematics (IK) to generate precise kinematic demonstrations for diverse construction objects, while eliminating dynamic disturbances that would otherwise compromise the robot's balance.

Second, to facilitate effective policy learning for unwieldy object transportation, we introduce a residual learning paradigm. This approach is motivated by the insight that fundamental locomotion dynamics are transferable across tasks, whereas unwieldy object interaction requires task-specific modulation. By leveraging a pre-trained base policy as a robust motion prior, our framework enables the residual policy to generate corrective actions that compensate for the additional load imposed by the object. To mitigate training instability of heavy objects, we implement a mass curriculum that incrementally increases the object's weight, preventing early failure.

Experimental results in a scaffold frame transportation task demonstrate that the residual policy can generalize to heavy objects (up to 5 kg) while maintaining robust balance and significantly outperform baselines trained from scratch or via standard fine-tuning. In summary, the primary contribution of this paper is a residual learning framework designed specifically for unwieldy object transportation. By effectively bridging the gap between robust locomotion and complex object manipulation, this work establishes a foundation for the deployment of humanoids in construction environments.

2 Literature review

2.1 Humanoid-object interaction collection

To equip humanoids with human-like loco-manipulation skills, obtaining high-quality human demonstrations is a critical requirement. High-fidelity demonstrations can accelerate the robot learning process by reducing the sample complexity. Existing data collection pipelines typically fall into three categories: (1) physical MoCap, (2) real-robot teleoperation, and (3) simulation-based methods (VR teleoperation and synthetic generation). However, each method faces distinct limitations when scaled to the specific domain of unwieldy object transportation.

Physical MoCap provides kinematic references of the highest fidelity, capturing accurate motion trajectories for both the human and the object [9,10]. However, scalability is a major bottleneck. The process is cost-prohibitive, requiring not only expensive capture infrastructure but also the fabrication of physical props for every unwieldy object variant, alongside the logistical costs of recruiting human subjects. Furthermore, standard MoCap is inherently limited in resolving contact dynamics, necessitating the use of physics-based simulation or haptic instrumentation to infer accurate interaction forces.

Real-robot teleoperation (via VR controllers or vision-based retargeting) offers an alternative that ensures physically plausible robot states. However, most teleoperation frameworks [11–13] focus on kinematic retargeting (i.e., controlling robot movement) rather than dynamic interaction, which could lead to instability or even failure when interacting with unwieldy objects. Crucially, unlike simulation, real-world teleoperation does not inherently provide ground-truth object states, necessitating complex external tracking setups to acquire the object trajectories needed for policy learning.

Simulation-based methods, specifically VR teleoperation [14] and synthetic data generation [15], offer lower costs and better scalability. VR teleoperation allows users to control a simulated humanoid to interact with diverse 3D object models without physical props. While this provides object ground truth, the lack of haptic feedback often results in "jittery" or noisy motions that are difficult to transfer to real hardware. Conversely, synthetic methods generate demonstrations using motion priors [16] or targeted object trajectories [15]. While highly scalable, these methods often suffer from physical implausibility, such as contact violations (e.g., mesh penetration) and a lack of fine-grained finger details.

Synthesizing these insights reveals two essential requirements for an effective data collection pipeline in the context of unwieldy object transportation: (1) *Scalability*, to readily accommodate the diverse objects, terrains, and robot models found in construction; and (2) *Physical fidelity*, ensuring the capture of physically plausible, synchronized trajectories for both the agent and the object to facilitate policy convergence.

2.2 Humanoid loco-manipulation

To function as effective collaborators in construction environments, humanoid robots must master the skill of loco-manipulation—specifically, the transportation of unwieldy objects using end-effectors. However, existing literature has predominantly focused on the loco-manipulation of compact household objects. Methods like Luo et al. [5] and Tessler et al. [6] (MaskedManipulator) focus on small objects manageable by a single hand or bimanually that do not account for significant external loads. However, unwieldy objects introduce distinct dynamic challenges, requiring the robot to maintain whole-body balance against large inertial forces caused by an eccentric center of mass and dynamic robot-object motion coupling.

While some studies address large object transportation, they often simplify the problem by allowing the robot to "hug" the object against its torso rather than using dexterous end-effectors [7,8]. Hugging reduces the moment arm and motion coupling effects by fixing the object close to the robot's center of mass. However, unlike compact household items, construction materials (e.g., scaffold frames and walk boards) are typically large and irregularly shaped. These geometric constraints prevent simple "hugging" strategies, necessitating precise grasping and stabilization via the robot's hands while moving, i.e., an end-effector-based approach. While there is research that does consider end-effector interactions [17], they frequently rely on simulated humanoid characters (i.e., SMPL-X robots) rather than real robots, e.g., the Unitree G1. These models often neglect real-world constraints such as setting large joint impedances, friction, or torque limits. This discrepancy can create a large "sim-to-real" gap.

To summarize, realizing effective humanoid transportation in construction necessitates that robots leverage dexterous end-effectors to manipulate unwieldy construction objects. Crucially, this requires managing complex robot-object interaction dynamics while strictly adhering to the robot's physical hardware constraints.

2.3 Research gaps

In summary, we identified two research gaps. First, current data collection pipelines lack a cost-efficient and scalable solution for collecting physically plausible humanoid-object interaction data, particularly for the unwieldy construction object transportation. Second, there is a notable absence of research addressing unwieldy object loco-manipulation that simultaneously accounts for end-effector interactions, complex object dynamics, and rigorous real-robot physical constraints.

3 Methodology

To bridge the first gap, we developed a scalable simulation-based method that integrates a pre-trained locomotion policy with bimanual IK to coordinate the humanoid's lower and upper body. By modeling the object with negligible mass in simulation, we mitigate dynamic disturbances, enabling the efficient generation of humanoid-object kinematic demonstrations without stability failures. To address the second gap, we propose a residual learning-based policy. This architecture leverages robust pretrained motion priors from the base policy while efficiently adapting to complex object dynamics and end-effector interactions using the residual policy.

3.1 Problem formulation

We cast the whole-body loco-manipulation task as a goal-conditioned Reinforcement Learning (RL) problem, formalized as a Markov Decision Process (MDP): $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$. At each time step t, the state $s_t \in \mathcal{S}$ encapsulates the robot's proprioceptive data s_r, the object state s_o, and their respective goal configurations g_r and g_o, such that $s_t = \{s_r, s_o, g_r, g_o\}$. The action $a_t \in \mathcal{A}$ denotes target joint angles, which are converted into actuation torques via a PD controller. The immediate reward is defined as $r_t = R(s_t, a_t)$, and we aim to learn a policy that maximizes the expected cumulative discounted return defined by $\mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} r_t]$.

We define an object ($\mathcal{O}$) as unwieldy if its mass distribution significantly alters the robot-object system's centroidal dynamics, such that the Combined Capture Point (CCP) exits the static support polygon $\mathcal{S}_{support}$, i.e., $\mathbf{x}_{CCP}(\mathcal{O}) \notin \mathcal{S}_{support}$, necessitating a whole-body response to prevent failure.

3.2 Stage 1: Generating humanoid-object interaction data

To efficiently collect reference trajectories for humanoid loco-manipulation, we decouple the control strategy into upper and lower body components. For the lower body, we deploy an existing pre-trained locomotion policy [11], which enables locomotion and body pose adjustment capabilities (e.g., crouching, bending). For the upper body, we leverage multi-target IK [18] to plan dual-arm movements, while hand actuation is managed via hardcoded open-close commands. Crucially, by modeling the object with a negligible mass (e.g., 0.01 g), we minimize dynamic disturbances to the pre-trained lower-body controller. This allows us to robustly generate kinematic demonstrations for diverse construction objects without stability failures. Furthermore, by utilizing a physics-based simulator, we ensure that the synchronized humanoid-object states remain physically plausible and fully documented.

3.3 Stage 2: General motion tracking policy

In this stage, our objective is to train a general motion tracking policy π_g (the base policy) that takes humanoid proprioception s_r and a reference motion g_r as input, and outputs actions a_g to track the reference trajectory. While our framework is compatible with any general motion tracking policy, we present a specific implementation here for demonstration.

We utilize the retargeted LAFAN1 dataset from Unitree [19], which contains motion data from 5 subjects across 77 sequences, totaling 496,672 frames (~4.6 hours) at 30 fps. To train the base policy, we adopt the single-stage RL framework from BeyondMimic [20], which provides a one-recipe-fits-all framework. The proprioceptive observation is defined as $\{\theta_t, v_t, \omega_t, q_t - q_0, \dot{q}_t - \dot{q}_0, a_{t-1}\}_t$, including the root orientation θ_t, root linear velocity v_t, root angular velocity ω_t, related joint positions $q_t - q_0 \in \mathbb{R}^{29}$, related joint velocities $\dot{q}_t - \dot{q}_0$, and the previous action a_{t-1}. The reference motion input is given by $\{\hat{p}_t - p_t, \hat{\theta}_t \ominus \theta_t, \hat{q}_t, \hat{\dot{q}}_t\}_t$, where $\hat{p}_t - p_t$ is the related reference root translation, $\hat{\theta}_t \ominus \theta_t$ is the related reference root orientation, $\hat{q}_t$ is the reference joint position, and $\hat{\dot{q}}_t$ is the reference joint velocity.

Following BeyondMimic [20], the reward function (r) comprises two main components: task rewards and regularization terms. The task reward (r_{task}) is the sum of tracking errors for position (p), orientation (θ), linear velocity (v), and angular velocity (ω). Errors ($\bar{e}_i$) are computed as Mean Squared Error (MSE) and mapped through a Gaussian-shaped exponential: $r(\bar{e}_i, \sigma_i) = \exp(-\bar{e}_i/\sigma_i^2) \in [0,1]$, where σ_s is determined empirically. The total task reward is: $r_{\text{task}} = \sum_{i \in \{p,\theta,v,\omega\}} r(\bar{e}_i, \sigma_i)$.

Three penalties are applied to ensure physical consistency and safety: (1) joint limit (r_{limit}), which keeps joints within soft limits to prevent hardware damage; (2) action rate (r_{smooth}) that promotes smooth transitions to avoid jitter; and (3) contact penalty (r_{contact}) which penalizes undesired contacts, excluding the links of ankles and end effectors. The final reward subtracts the weighted penalties from the task reward: $r = r_{\text{task}} - \lambda_l r_{\text{limit}} - \lambda_s r_{\text{smooth}} - \lambda_c r_{\text{contact}}$, where $\lambda > 0$ denote the weight for each term.

To improve generalization and robustness, the training randomizes ground friction and restitution coefficients, default joint positions to simulate offset errors, torso center of mass position, and applies random velocity perturbations.

3.4 Stage 3: Residual policy

Building on the pre-trained base policy π_g, a residual policy π_{res} is introduced to refine coarse actions and complete the desired loco-manipulation task. This approach adopts single-stage RL with PPO, where the residual policy takes the state tuple $\langle s_r, s_o, g_r, g_o \rangle$ as input and outputs a residual action $a_{res} \in \mathbb{R}^{29}$. Specifically, the object state is represented as $s_o = [\theta_t^o, v_t^o, \omega_t^o]$, and the reference object trajectory as $g_o = [\hat{p}_t^o, \hat{\theta}_t^o, \hat{v}_t^o, \hat{\omega}_t^o]$. The final executed action is the summation of the residual and base actions, $a = a_{res} + a_g$. While this framework performs well with light objects, it often becomes unstable when handling larger masses. To address this instability, an object mass curriculum is implemented, starting with a small object mass and gradually increasing it during training to

Table 1 Quantitative comparison of three loco-manipulation policies of 1 kg object mass.

Method	*Success rate*	*Avg. length*	*Anchor pos.*	*Joint pos.*	*Object rot.*	*Iteration*
Learn from scratch	0%	8.20	0.69	1.42	1.48	11360
Fine-tuning	0%	5.98	0.71	1.22	1.53	2000
Residual learning	70%	25.84	0.81	1.42	1.23	6250

Table 2 Quantitative comparison of different object weights using the proposed method.

Method	*Success rate*	*Avg. length*	*Anchor pos.*	*Joint pos.*	*Object rot.*
1 KG	70%	25.84	0.81	1.42	1.23
3 KG	40%	16.90	0.84	1.63	1.20
5 KG	10%	7.27	0.81	1.64	1.15

prevent failure.

For training the residual policy, the system reuses the motion reward and domain randomization from the base policy but introduces an additional object tracking reward r_t^o, to encourage task completion. This term measures the pose differences between the simulated and reference objects using the formula $r_t^o = \exp(-\lambda_p \parallel p_t^o - \hat{p}_t^o \parallel_2) + \lambda_w \cdot \exp(-\lambda_\theta \parallel \theta_t^o \ominus \hat{\theta}_t^o \parallel_2)$.

Additionally, early termination conditions are expanded; the episode ends not only if the robot makes unintended ground contact or deviates from its reference, but also if the object's translation or orientation deviates significantly from its reference threshold.

4 Experiments

Our experiments are designed to: (1) validate the effectiveness of the proposed residual policy; and (2) assess its robustness to unwieldy objects of varying masses.

4.1 Experimental Setup

Robot. We perform simulation experiments on a Unitree G1 humanoid robot (29-DoF) with two Inspire hands (24-DoF). For simplicity, we assume that the robot has already grasped the object, and the joint positions of both hands are set to close by default. This helps reduce the action space for speeding up the learning process.

Task Description. We evaluate our approach within a scaffolding assembly scenario, specifically focusing on the transportation of vertical scaffold frames. We selected this task because these frames are representative of unwieldy construction objects, characterized by their large dimensions, substantial weight, and eccentric centers of mass. The objective is to maintain whole-body balance during transportation while accurately tracking a long-horizon reference motion trajectory (2 minutes).

Evaluation metrics. We assess policy performance using two primary criteria: success rate (a trial is deemed successful if the object is transported for over 30 seconds without being dropped) and average episode length (s). Additionally, consistent with prior works that report both robot and object tracking errors [5,8], we quantify the anchor (torso link) position error (m), joint position tracking error (rad), and object pose errors (rad).

Baselines. To validate our method's reliability and effectiveness, we consider two representative baselines: (1) *Learn from scratch.* A single-stage RL policy is trained from scratch to track both human motion and object trajectories. (2) *Finetuning base policy.* The general motion tracking (base) policy is finetuned to track human motion and object trajectories. However, because the base policy's neural network was pre-trained exclusively on proprioceptive and human motion data, its input dimensions are fixed. Consequently, direct fine-tuning cannot incorporate novel object state observations (such as object pose or velocity) without destructively modifying the network's input layer. Thus, this baseline is constrained to its original observation space, forcing the policy to infer object dynamics indirectly.

For fairness, we use the same reward terms as the proposed method for these two baselines without task-specific tuning. All policies are trained in IsaacLab to leverage its massive parallelism for accelerated data collection.

4.2 Simulation Results

Based on the experimental results (Tables 1 and 2 and Fig. 2), several key conclusions can be drawn:

First, fine-tuning the base policy alone proves insufficient for loco-manipulation tasks, although it establishes a strong prior for joint tracking. As shown in Table 1, this method yields a 0% success rate and the shortest average episode length (~6 s). This failure stems from the lack of explicit object state observations, causing the policy to prioritize internal joint tracking while ignoring the scaffold frame's dynamics. This is evidenced in Fig. 2, which shows that the robot fails to stabilize the object, resulting in high tracking errors and significant oscillation (fluctuations in roll, pitch, and yaw). However, due to pre-training on a general motion dataset, this method maintains low joint position and global anchor errors, confirming its utility as a robust kinematic prior.

Second, residual learning significantly enhances both training efficiency and task effectiveness. Compared to the learning-from-scratch baseline, our residual approach achieves a substantially higher success rate (70%) and more than triples the average episode length (26 s vs. 8 s). It also exhibits superior stability, characterized by lower object orientation tracking errors and smoother trajectories, as shown in Fig. 2. Furthermore, it demonstrates better sample efficiency, achieving convergence in approximately half the training iterations required by the baseline.

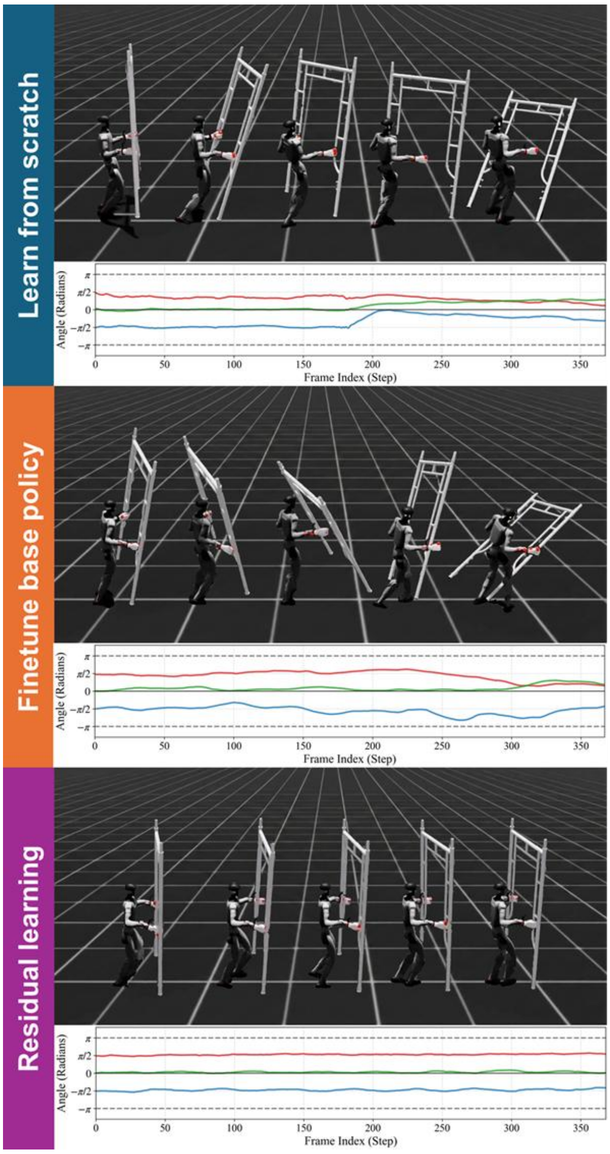

Figure 2 Comparison between three policies. The corresponding curves quantify the object's roll (red), pitch (green), and yaw (blue) angles relative to the robot's base orientation.

Third, policy performance degrades as object mass increases. As detailed in Table 2, the success rate and average episode length drop dramatically when generalizing to objects heavier than the 1 kg training standard. This deterioration can be attributed to the increased inertial forces generated by the heavier load, coupled with the vertical bouncing and lateral swaying inherent to humanoid walking. These dynamic forces exceed the robot's wrist torque limits, making it difficult to maintain a stable grasp and resulting in object slippage. This limitation motivates future research into mass-adaptive control strategies.

5 Discussion

In this study, we addressed the critical challenge of humanoid loco-manipulation for unwieldy construction objects by proposing a novel three-stage residual learning framework. Our experimental results demonstrate that leveraging pre-trained locomotion priors alongside task-specific residual adjustments effectively mitigates the severe dynamic coupling introduced by large, eccentric loads. While these findings validate our core hypothesis in simulation, translating this capability into practical, real-world applications requires a careful examination of its broader implications and operational boundaries. The following sections contextualize our framework within existing construction workflows, analyze the anticipated sim-to-real gap, and outline the current limitations and future directions of this research.

Integration with existing workflows. Beyond the proof-of-concept experiment, the proposed framework is designed to integrate into broader construction workflows, serving as a critical primitive for autonomous material logistics and human-robot collaborative assembly. In a real-world construction scenario, high-level task planners or human supervisors would provide target waypoints to the robot. The residual learning framework then acts as a robust low-level controller, enabling the humanoid to autonomously navigate the site and deliver diverse unwieldy materials directly to the point of installation. By delegating the physically demanding and dynamically destabilizing task of material transport to the humanoid, this workflow allows skilled human workers to focus exclusively on complex, high-dexterity assembly tasks, thereby improving overall site productivity and safety.

Sim-to-real gaps. Successfully deploying this framework on physical hardware, such as the Unitree G1, requires bridging a substantial sim-to-real gap. A primary challenge involves hardware constraints under dynamic stress; the physical limits of wrist actuators and the complex friction dynamics of the robotic hands are difficult to perfectly replicate in simulation, particularly when subjected to the impact forces of walking.

Furthermore, acquiring the real-time object state presents a significant perception problem. While vision-based pose estimation can be utilized, it is highly susceptible to latency and visual occlusions from the robot's own body. To overcome this, future hardware deployments will heavily prioritize proprioceptive feedback. Integrating force/torque sensors at the wrists or utilizing high-frequency joint effort estimations presents a robust alternative or addition to visual object pose tracking. This tactile and force feedback is immune to occlusion and directly measures the dynamic coupling forces, providing the residual policy with the low-latency signals necessary to prevent slippage and maintain whole-body balance in real-world construction scenarios.

Limitations. However, this approach is not without limitations. The current method relies on the assumption that the residual policy can fully compensate for the lack of dynamic information in the demonstration data. While effective for moderate weights, this paradigm may struggle as object mass approaches the robot's hardware limits, where kinematic adaptation alone is insufficient. Additionally, our current evaluation is limited to simulation on flat terrain and a single type of object. Real-world construction environments present uneven topography and clutter that require more robust perception-locomotion coupling and testing on more diverse construction objects.

Future work. Future work will address these challenges by incorporating terrain perception into the policy inputs, tackling the complex contact dynamics required to autonomously grasp and pick up these materials, and exploring mass-adaptive control strategies and multi-robot collaboration to distribute the load of heavier construction elements. We also aim to validate the proposed framework on physical hardware to assess its robustness against real-world sensing noise and actuation delays.

6 Conclusion

In this paper, we addressed the challenging problem of humanoid loco-manipulation for unwieldy object transportation in construction environments. We identified that the primary barriers to this task are the complex physical dynamics—specifically geometric constraints, eccentric static loads, and inertial coupling—and the scarcity of domain-specific interaction data. To overcome these challenges, we proposed a scalable residual learning framework. By leveraging synthetic data generation and residual policy, our approach enables robust transport of irregular objects without requiring expensive motion capture. Experimental results demonstrate that our policy significantly outperforms baseline methods, successfully transporting scaffold frames up to 5 kg while maintaining whole-body balance. This work provides a foundational step toward deploying humanoid robots as effective collaborative partners in unstructured construction sites.

References

[1] Boston Dynamics. Atlas gets a grip, 2023. On-line: https://www.youtube.com/watch?v=BTMKuvO0VNg, Accessed: 12/31/2025.

[2] NASA. NASA's Valkyrie robot can erect scaffolding, 2019. https://www.youtube.com/watch?v=j1CgTzQSYyo Accessed: 12/31/2025.

[3] L. Atanasova, D. Mitterberger, T. Sandy, F. Gramazio, M. Kohler, K. Dorfler, Prototype as artefact - Design tool for open-ended collaborative assembly processes, *Proc. Annu. Conf. Assoc. Comput. Aided Des. Archit.: Distrib. Prox.*, ACADIA, ACADIA, Virtual, Online, 2020: pp. 350–359.

[4] Y. Liu, M. Habibnezhad, H. Jebelli, Brain-computer interface for hands-free teleoperation of construction robots, *Automation in Construction* 123:103523, 2021.

[5] Z. Luo, J. Cao, S. Christen, A. Winkler, K. Kitani, W. Xu, Grasping diverse objects with simulated humanoids. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pages: 2161–2184, 2024.

[6] C. Tessler, Y. Jiang, E. Coumans, Z. Luo, G. Chechik, X.B. Peng, MaskedManipulator: Versatile whole-body control for loco-manipulation. *Arxiv*, 2025.

[7] L. Pan, Z. Yang, Z. Dou, W. Wang, B. Huang, B. Dai, T. Komura, J. Wang, TokenHSI: Unified synthesis of physical human-scene interactions through task tokenization, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages:5379–5391, 2025.

[8] S. Zhao, Y. Ze, Y. Wang, C.K. Liu, P. Abbeel, G. Shi, R. Duan, ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *Arxiv*, 2025.

[9] N. Mahmood, N. Ghorbani, N.F. Troje, G. Pons-Moll, M.J. Black, AMASS: Archive of motion capture as surface shapes. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages: 5442-5451, 2019.

[10] O. Taheri, N. Ghorbani, M.J. Black, D. Tzionas, GRAB: A dataset of whole-body human grasping of objects. *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science*, 12349:581–600, Springer, Cham, 2021.

[11] Z. Zhang, C. Chen, H. Xue, J. Wang, S. Liang, Y. Liu, Z. Zhang, H. Wang, L. Yi, Unleashing humanoid reaching potential via real-world-ready skill space. *Arxiv*, 2025.

[12] Y. Ze, Z. Chen, J.P. Araújo, Z. Cao, X.B. Peng, J. Wu, C.K. Liu, TWIST: Teleoperated whole-body imitation system. *Arxiv*, 2025.

[13] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, G. Shi, OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning, *Proceedings of The 8th Conference on Robot Learning, PMLR*, 270:1516–1540, 2025.

[14] M. Seo, S. Han, K. Sim, S.H. Bang, C. Gonzalez, L. Sentis, Y. Zhu, Deep imitation learning for humanoid loco-manipulation through human teleoperation, *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8, Austin, TX, USA, 2023.

[15] J. Li, J. Wu, C.K. Liu, Object motion guided human motion synthesis, *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

[16] D. Kalaria, S.S. Harithas, P. Katara, S. Kwak, S. Bhagat, S. Sastry, S. Sridhar, S. Vemprala, A. Kapoor, J.C.-K. Huang, DreamControl: Human-inspired whole-body humanoid control for scene interaction via guided diffusion. *Arxiv*, 2025.

[17] S. Xu, H.Y. Ling, Y.-X. Wang, L.-Y. Gui, InterMimic: towards universal whole-body control for physics-based human-object interactions, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12266-12277, 2025.

[18] C.M. Kim, B. Yi, H. Choi, Y. Ma, K. Goldberg, A. Kanazawa, PyRoki: A modular toolkit for robot kinematic optimization. *Arxiv*, 2025.

[19] Lafan1 retargeting dataset (unitree g1/h1/h1-2, (2025). On-line: https://huggingface.co/datasets/lvhaidong/LAFAN1_Retargeting_Dataset. Accessed: 01/04/2019.

[20] Q. Liao, T.E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, C.K. Liu, BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion. *Arxiv*, 2025.