CLONE: Closed-Loop Whole-Body Humanoid Teleoperation for Long-Horizon Tasks

Yixuan Li [*,1,2] lyx@bit.edu.cn
Yutang Lin [*,3,4,5,6] yutang.lin@stu.pku.edu.cn
Jieming Cui [2,3,4,6] cuijieming@stu.pku.edu.cn
Tengyu Liu [2,7] liutengyu@bigai.ai
Wei Liang✉ [,1] liangwei@bit.edu.cn
Yixin Zhu✉ [,3,4,6,8] yixin.zhu@pku.edu.cn
Siyuan Huang✉ [,2,7] syhuang@bigai.ai

[*] Equal contributors [1] School of Computer Science and Technology, Beijing Institute of Technology
[2] State Key Laboratory of General Artificial Intelligence, BIGAI
[3] School of Psychological and Cognitive Sciences, Peking University
[4] Institute for Artificial Intelligence, Peking University [5] Yuanpei College, Peking University
[6] Beijing Key Laboratory of Behavior and Mental Health, Peking University
[7] Joint Laboratory of Embodied AI and Humanoid Robots, BIGAI & UniTree Robotics
[8] Embodied Intelligence Lab, PKU-Wuhan Institute for Artificial Intelligence

https://humanoid-clone.github.io/

Figure 1: **CLONE** employs an MoE-based policy with **closed-loop** error correction for humanoid teleoperation, enabling precise **whole-body coordination** and **long-horizon** task execution.

Abstract: Humanoid teleoperation plays a vital role in demonstrating and collecting data for complex humanoid-scene interactions. However, current teleoperation systems face critical limitations: they decouple upper- and lower-body control to maintain stability, restricting natural coordination, and operate open-loop without real-time position feedback, leading to accumulated drift. The fundamental challenge is achieving precise, coordinated whole-body teleoperation over extended durations while maintaining accurate global positioning. Here we show that an MoE-based teleoperation system, **CLONE**, with closed-loop error correction enables unprecedented whole-body teleoperation fidelity, maintaining minimal positional drift over long-range trajectories using only head and hand tracking from an MR headset. Unlike previous methods that either sacrifice coordination for stability or suffer from unbounded drift, **CLONE** learns diverse motion skills while preventing tracking error accumulation through real-time feedback, enabling complex coordinated movements such as "picking up objects from the ground." These results establish a new milestone for whole-body humanoid teleoperation for long-horizon humanoid-scene interaction tasks.

Keywords: Humanoid; Whole-body teleoperation; Humanoid-scene interaction
9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

1 Introduction

The ability to seamlessly coordinate whole-body movements while navigating complex environments represents one of humanity's most remarkable capabilities [1, 2]. From squatting to retrieve objects from the ground to walking across rooms while carrying items, humans effortlessly integrate locomotion and manipulation in ways that remain challenging for robots [3–7]. Humanoid robots (humanoids henceforth), with their human-like morphology, offer the promise of replicating these capabilities—potentially enabling applications from household assistance to operations in hazardous environments where human-like dexterity and mobility are essential [8–11].

However, realizing this potential requires solving a fundamental challenge: enabling intuitive and precise teleoperation that maintains coordination across the entire body over extended periods. Long-horizon tasks, such as navigating to distant locations while manipulating objects, demand not only moment-to-moment stability but also sustained accuracy in both movement execution and global positioning. Current teleoperation approaches fall short of these requirements, creating a significant capability gap between human operators and humanoids.

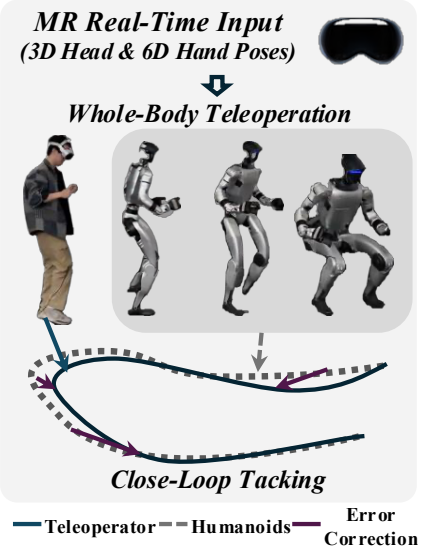


Figure 2: **Whole-body humanoid teleoperation from minimal input.** Our approach enables intuitive control of a humanoid robot using only head and hand poses from mixed reality input, generating coordinated whole-body motions including natural locomotion. Through closed-loop tracking, the system maintains accurate correspondence between operator and robot over extended operation periods, enabling complex long-horizon tasks that require sustained precision.

Recent advances in humanoid teleoperation and loco-manipulation [12–19] have made notable progress. Nevertheless, existing methods struggle with precise teleoperation over extended durations and fall short of the whole-body coordination necessary for humanoid-scene interaction. Two fundamental challenges persist in bridging this capability gap.

The first challenge centers on achieving **coordinated whole-body coordination**. Many systems decouple upper- and lower-body control for stability [17, 20], sacrificing the natural synergies required for fluid motion. While this separation provides safety, it fundamentally limits integrated actions such as reaching while walking or adjusting posture during manipulation. Alternative approaches that rely on motion capture data [12, 14, 18, 21–26] often emphasize stability at the cost of expressiveness, yielding conservative motions constrained by training data distributions. Moreover, these methods consistently overlook key factors like hand orientation that are critical for dexterous tasks, further restricting humanoids' potential for sophisticated whole-body movements.

The second challenge involves **accumulated positional drift** over time due to the absence of real-time feedback about the robot's actual position in the environment. Unlike wheeled robots with straightforward odometry, humanoids exhibit complex foot-ground interactions and non-holonomic dynamics that complicate accurate state estimation. Without closed-loop correction, small pose errors compound with each step, progressively degrading the operator's spatial awareness and control authority, eventually leading to complete task failure. This drift becomes particularly acute during manipulation tasks that require precise positioning relative to environmental objects.

To tackle the above challenges in humanoid long-horizon tasks requiring whole-body coordination and accurate positioning, we present **CLONE**. As illustrated in Fig. 2, **CLONE** is a **closed-loop** whole-body teleoperation system combining learning-based coordination and real-time feedback correction. Our system employs an Mixture-of-Experts (MoE) architecture that learns to coordinate diverse motion skills while a LiDAR-based error correction mechanism prevents the accumulation of positional drift. Critically, **CLONE** requires only head and hand tracking from a single commercial Mixed Reality (MR) headset, making it practical for real-world deployment while achieving unprecedented fidelity in long-horizon tasks.

Our approach integrates three key components: (i) We develop an MoE framework that enables unified learning of diverse motion skills while maintaining natural upper- and lower-body coordination. (ii) We implement closed-loop error correction using LiDAR odometry [27] and Apple Vision Pro (AVP) tracking to provide continuous global pose feedback and prevent drift accumulation. (iii) We curate a comprehensive dataset, **CLONED**, that augments AMASS [28] with additional motion-captured sequences and an online hand orientation generation method, ensuring robust generalization to complex manipulation scenarios involving coordinated whole-body movements.

Our experiments demonstrate that **CLONE** enables capabilities previously unattainable with existing systems: whole-body coordination over long trajectories with minimal positional drift, complex coordinated movements like object retrieval from ground level, and robust performance across diverse operator configurations and environmental conditions. Using only minimal input from a commercial MR headset, **CLONE** achieves improved tracking precision over existing open-loop approaches, opening new possibilities for practical humanoid applications in unstructured environments.

Our contributions are four-fold: (i) the first MoE-based framework for coordinated whole-body teleoperation that maintains natural movement synergies; (ii) a closed-loop system that solves the fundamental position drift problem in long-horizon tasks through real-time pose correction; (iii) a comprehensive dataset, **CLONED**, enabling robust learning of dexterous whole-body motions with proper hand orientation coverage; and (iv) extensive validation demonstrating substantial improvements in real-world humanoid-scene interaction capabilities across diverse scenarios.

2 Related Work

Whole-Body Humanoid Teleoperation Humanoid teleoperation enables robots to replicate human movements for complex tasks using motion capture systems [23, 16, 29], haptic devices [30–32], or virtual reality interfaces [14, 18, 33–35]. A key challenge is developing whole-body control[1] policies that balance robot stability with motion tracking fidelity. Current methods struggle to reproduce the full diversity and fluidity of human motions [40], primarily due to monolithic MLP-based architectures that inadequately handle conflicting objectives across different motion types (*e.g.*, walking *vs*. crouching) [41–43]. Although mixture-based models have shown promise in other domains [44–47], their application to humanoid teleoperation remains underexplored. To address these limitations, we leverage an MoE framework for adaptive learning and unified representation of diverse motion patterns within a single policy.

Long-Horizon Loco-Manipulation Long-horizon task execution [48] has been extensively studied for fixed-base arms [49–52], mobile manipulators [53–57], and aerial manipulators [58], typically in structured settings. In contrast, humanoid teleoperation remains limited to short-horizon motion replication [18, 22, 23], operating open-loop due to difficulties in real-time global state estimation for bipedal systems. Although recent advances in odometry have improved state tracking for legged robots [59–61], their application to long-horizon humanoid control remains largely unexplored. To bridge this gap, we integrate LiDAR odometry into our teleoperation framework to enable closed-loop error correction and significantly reduce accumulated drift.

Datasets for Training Humanoids Large-scale motion capture (MoCap) datasets [28, 62] have been instrumental in training humanoid control policies [14, 18, 22, 63, 64]. Even after augmenting the datasets with generative models [23], these datasets were still confined primarily to animation and graphics [65, 24] rather than robotics applications. While they contain semantically distinct actions (*e.g.*, waving, hugging, drinking), they underrepresent the kinematic configurations and dynamic transitions required for robust, generalizable controller training in real-world scenarios. To

[1] Whole-Body Control (WBC) in robotics literature [36–38] was traditionally formulated as an optimization problem [3, 39], coordinating multiple competing tasks (such as balance and reaching) through hierarchical control objectives. Learning-based approaches [12, 18, 23] have recently extended this concept by formulating whole-body control as a reinforcement learning problem. We refer to our approach as a whole-body control policy, as it similarly coordinates all degrees of freedom of the humanoid in a unified manner.

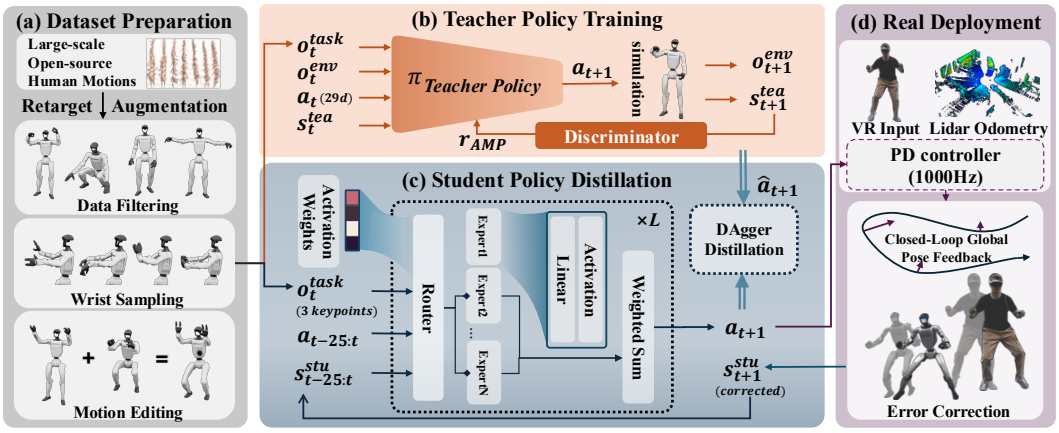


Figure 3: **The CLONE framework.** (a) **CLONED** curates and augments retargeted AMASS [28] data through motion editing to introduce diverse humanoid motions and detailed hand movements. (b) A teacher policy is trained using privileged information, including full robot state and environmental context. (c) An MoE network serves as the student policy, distilled from the teacher to operate with real-world observations only. (d) For real-world deployment, we integrate LiDAR odometry to obtain real-time humanoid states, enabling closed-loop error correction during teleoperation.

address these limitations, we introduce **CLONED** by augmenting AMASS [28] through motion editing and collecting additional human MoCap data, specifically tailored for humanoid controllers. This expansion increases coverage of motions and transitions relevant to humanoid control tasks.

3 The CLONE Framework

Our teleoperation framework captures a minimal set of control signals from the teleoperator, consisting solely of the 6D poses (position and orientation) of both wrists and the 3D position of the head, tracked using an AVP headset. These three points (see also Fig. 3) serve as the complete control interface, providing an intuitive yet powerful means of directing the humanoid's whole-body motion while maintaining a simple setup that requires no additional hardware or complex calibration.

CLONE addresses two fundamental challenges through complementary components. First, we develop a teacher-student policy learning approach that transforms these sparse control signals into coordinated whole-body movements (Sec. 3.1). Second, we implement a closed-loop error correction mechanism that maintains positional accuracy during extended operation (Sec. 3.2). The system is supported by carefully designed reward structures and randomization techniques (Sec. 3.3) and trained on a newly curated dataset, **CLONED**, that ensures robust generalization (Sec. 3.4). Additional details of the implementation are provided in Appendix C.

3.1 Policy Learning

We employ a teacher-student training strategy for the teleoperation policy, following the overall framework of OmniH2O [18] (see Appendix A.1 for the problem formulation). This approach first trains a teacher policy with privileged information, then distills this knowledge into a student policy that operates using only real-world observations.

Teacher Policy Training The teacher policy π_{tea} is implemented as an Multi-Layer Perceptrons (MLP) that leverages comprehensive state information unavailable on real robots. At each timestep t, it processes observations $\mathbf{o}_t^{\text{tea}} = [\mathbf{s}_t^{\text{tea}}, \mathbf{o}_t^{\text{task}}, \mathbf{a}_t, \mathbf{o}_t^{\text{env}}]$ and outputs target joint positions $\mathbf{a}_{t+1} \in \mathbb{R}^{29}$ for PD control. The privileged states $\mathbf{s}_t^{\text{tea}} = [\mathbf{p}_t, \theta_t, \mathbf{v}_t, \omega_t]$ include joint angular positions $\mathbf{p}_t$ and the 6D poses, linear velocities, and angular velocities $\theta_t, \mathbf{v}_t, \omega_t$ of all robot links. Task observations $\mathbf{o}_t^{\text{task}} = [\hat{\mathbf{p}}_{t+1} - \mathbf{p}_t, \hat{\theta}_{t+1} - \theta_t, \hat{\mathbf{v}}_{t+1} - \mathbf{v}_t, \hat{\omega}_{t+1} - \omega_t, \hat{\mathbf{p}}_{t+1}, \hat{\theta}_{t+1}]$ capture both reference motion (denoted by $\hat{\cdot}$) and tracking errors between reference and current states. Environmental observations $\mathbf{o}_t^{\text{env}}$ provide context including ground friction coefficient and robot mass distribution.

Student Policy Distillation The student policy must operate without privileged information, following $a_{t+1} = \pi_{\text{stu}}(s^{\text{stu}}_{t-25:t}, a_{t-25:t}, o^{\text{task}}_t)$. The robot state sequence $s^{stu}_{t-25:t}$ contains joint positions q, joint velocities $\dot{q}$, root angular velocity ω^{root}, and root gravity vector g obtained from on-device IMU over the past 25 frames. Task observations o^{task}_t consist of $\hat{p}_{t+1} - p_t$, $\hat{p}_{t+1}$, $\hat{\dot{p}}_{t+1}$, h_t, and $\hat{h}_{t+1}$, where p_t represents the 3D positions of head and two wrists obtained from LiDAR odometry and forward kinematics, $\hat{p}_{t+1}$ and $\hat{\dot{p}}_{t+1}$ are target positions and velocities from reference motion, and h_t, $\hat{h}_{t+1}$ represent current and target wrist orientations.

The key challenge lies in handling diverse motion patterns within a single policy. Walking requires different control strategies than crouching or reaching, yet traditional monolithic architectures struggle with these conflicting objectives. We address this through an MoE architecture, as shown in Fig. 3, which allows specialized processing for different motion types.

The MoE design consists of L layers, each comprising N experts that function as independent feed-forward sub-layers with distinct parameters. At each layer, a router dynamically selects which experts are activated based on the input, generating weight distributions over all experts. The layer output combines the top-k experts with highest routing weights: $f = \sum_i^k w_i \cdot E_i(\cdot)$, where w_i is the routing weight for the i-th selected expert and $E_i(\cdot)$ is the output of the i-th expert. This design enables different experts to focus on distinct motion patterns. To prevent model collapse to only a few experts, we introduce a balancing loss that encourages uniform expert selection:

$$\mathcal{L}_{balance} = \sum_{l=1}^{L} \sum_{e=1}^{N} [\max(p_e - \frac{1+\epsilon}{N}, 0) + \min(\frac{1-\epsilon}{N} - p_e, 0)], \tag{1}$$

where $p_e = \mathbb{E}[w_e]$ represents the expected activation probability of expert e, and ϵ is a slack constant that allows slight deviations from perfect uniformity.

3.2 Closed-Loop Error Correction

Traditional humanoid teleoperation systems operate in an open-loop configuration, where small errors in position tracking accumulate over time, leading to significant drift during extended operations. This fundamental limitation becomes particularly problematic during long-horizon tasks that require sustained positional accuracy. To address this challenge, we implement a closed-loop error correction mechanism that continuously monitors and compensates for positional discrepancies between the teleoperator and the humanoid.

Our approach utilizes LiDAR odometry to maintain accurate global position estimates for both the humanoid and the teleoperator. We employ FAST-LIO2 [27], an algorithm that tightly couples IMU and LiDAR data through an iterated Kalman filter to provide robust real-time state estimation even during dynamic movements (see more details in Appendix A.2). This choice ensures reliable tracking performance across diverse motion patterns, from walking to complex manipulation tasks.

The system tracks global positions for both agents: the humanoid's position $p \in \mathbb{R}^3$ is computed from onboard sensors, while the teleoperator's position $\hat{p} \in \mathbb{R}^3$ is similarly tracked through a MR hardware equipped with a comparable odometry pipeline. The student teleoperation policy directly consumes the difference between p and $\hat{p}$, enabling it to generate actions that systematically reduce positional drift and maintain accurate correspondence between the operator and the humanoid.

3.3 Reward Design and Domain Randomization

We build upon the reward terms and domain randomizations from OmniH2O [18] as the foundation of our approach, with specific enhancements to address the challenges of real-world teleoperation. Detailed reward functions and domain randomization settings are provided in Appendix B.

To enhance robustness against LiDAR odometry errors, we introduce a velocity-dependent Stochastic Differential Equation (SDE) noise model during training that reflects real-world error characteristics. For the head position $\vec{p}_{\text{head}}$, we define the randomized position $\vec{P}_{\text{head}}$ as:

$$\mathrm{d}\vec{P}_{\text{head}} = \dot{\vec{p}}_{\text{head}}\mathrm{d}t + (\frac{\| \dot{\vec{p}}_{\text{head}} \|}{c_{\text{vel}}} + c_{\min})\mathrm{d}\vec{W}, \tag{2}$$

where $\vec{W}$ is a standard Wiener process, and c_{vel} and $c_{\min}$ are constants that scale the noise proportionally to movement speed and establish a minimum randomization level. This formulation mirrors real-world dynamics, where faster movements tend to produce greater odometry errors. We use forward kinematics to compute other body positions based on the randomized head position, while periodically resetting and constraining the maximum deviation to avoid unrealistic drift.

Since **CLONE** provides only upper-body references (head and wrists), we must generate appropriate lower-body behaviors without explicit guidance. To tackle this challenge, we employ an Adversarial Motion Priors (AMP) reward [66] to regularize lower-body movements and encourage natural, stable behavior. Through this combination of specialized domain randomization and reward design, **CLONE** learns to generate robust lower-body behaviors while maintaining precise upper-body control aligned with operator commands.

3.4 The CLONED Dataset

The training dataset **CLONED** comprises three complementary components to support robust whole-body teleoperation. These include: (i) an augmented AMASS [28] subset of 149 curated sequences featuring diverse pairings of upper- and lower-body movements, enhanced via motion editing to increase compositional diversity and policy generalization(see Appendix C.1); (ii) 14 custom sequences captured with an IMU-based Xsens MoCap system to fill coverage gaps, emphasizing continuous transitions and diverse upper-body poses critical for manipulation; and (iii) systematic hand orientation augmentation through procedurally generated 6D wrist targets, smoothed via Spherical Linear Interpolation (SLERP) to ensure coherent and natural hand motions for teleoperation.

4 Real-World Experiments

We evaluated **CLONE** on a physical Unitree G1 humanoid through comprehensive experiments demonstrating exceptional whole-body motion fidelity and precise position tracking. Our experiments focus on two key capabilities: (i) global position tracking accuracy during extended teleoperation to validate our closed-loop error correction mechanism, and (ii) whole-body motion tracking fidelity across diverse skills to demonstrate coordination capabilities. These experiments collectively validate both the technical performance and practical applicability of our approach to real-world humanoid teleoperation.

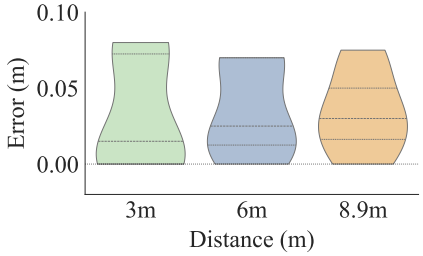
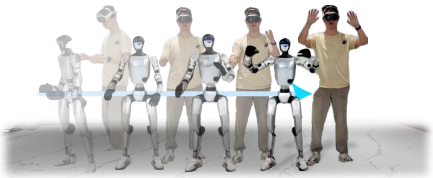


Figure 4: **Global position tracking accuracy in real-world experiments. CLONE** achieves mean tracking errors of 5.1cm across distances up to 8.9m, demonstrating effective closed-loop error correction in extended teleoperation.

Global Position Tracking To evaluate global positioning accuracy over extended distances, we designed a controlled path-following experiment. (i) **Straight-path tracking.** We fixed the initial positions of both the operator and the robot. The operator then walked along straight paths to target positions at 3m, 6m, and 8.9m while teleoperating the robot. We measured the discrepancy between the robot's final and expected positions as the tracking error and repeated each condition ten times. (ii) **Curved-path tracking.** We teleoperated the humanoid along a 10m trajectory with two 90° turns, representative of typical household paths, and repeated the trial six times to measure translational and rotational tracking errors.

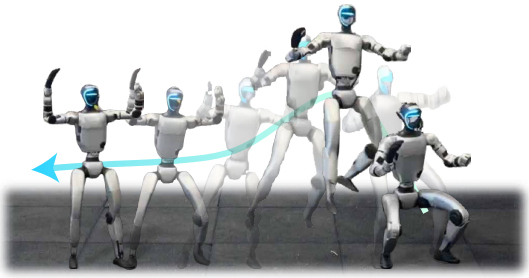
CLONE achieved a mean tracking error of 5.1 cm in straight-path tracking, with a maximum deviation of 12.0 cm at 8.9m (see Fig. 4). These results show that **CLONE**'s closed-loop error correction effectively mitigates accumulated errors during extended teleoperation.

Statistical analysis confirms consistent tracking performance across all tested distances. Independent samples t-tests comparing distance groups yielded: 3m versus 6m ($t = 0.165, p = 0.871$) and 6m versus 8.9m ($t = 0.048, p = 0.963$). With all p-values > 0.05, **CLONE** shows no significant performance degradation across the tested range, demonstrating that our approach maintains high accuracy over extended distances without drift accumulation.



(a) Waving

(b) Squatting



(c) Jumping

(d) Squatting

Figure 5: **Whole-body motion tracking on Unitree G1.** **CLONE** successfully tracks diverse skills including (a) waving, (b)(d) squatting, and (c)jumping, showcasing comprehensive whole-body coordination capabilities.

In curved-path tracking, the mean error is 20 cm (maximum 27 cm), and the mean rotational drift was 2° between the operator's and humanoid's orientations. These results highlight **CLONE** 's ability to sustain high precision even over long and complex trajectories.

Whole-Body Motion Tracking As shown in Fig. 5, **CLONE** successfully enables real-time teleoperation across a diverse range of whole-body skills. The robot accurately tracks complex motions including waving, squatting, standing up from squatted positions, and jumping. These results demonstrate high whole-body motion fidelity for real-time humanoid teleoperation, particularly for dynamic skills like jumping requiring precise coordination of balance control and force application.

Long-Horizon Mixed Navigation To validate system performance in complex scenarios, we conducted extended teleoperation sessions incorporating multiple movement types. As visualized in Fig. 6, we recorded a continuous teleoperation sequence where the operator traversed a complex path spanning over 15m, incorporating diverse locomotion patterns including forward walking, turning, side-stepping, and returning to the original position.

The robot consistently tracked the operator's movements throughout this extended sequence and returned to its starting position with minimal drift. This result validates **CLONE**'s robustness for extended teleoperation sessions that combine locomotion and whole-body motion control—a critical capability for practical humanoid applications requiring sustained coordination over long horizons.

(a) Sequence 1

(b) Sequence 2

Figure 6: **Long-horizon teleoperation.** The humanoid accurately tracks both the operator's local pose and global translation throughout a complex navigation sequence, demonstrating robust performance.

5 Simulations

We present comprehensive evaluations of **CLONE** in simulation across four key settings: reference motion tracking, diverse stance tracking, ablation studies (Appendix D.2), and expert activation analysis (Appendix D.4). These experiments are designed to: (i) quantify motion tracking accuracy in controlled simulation environments, and (ii) assess robustness across diverse stance configurations. The evaluation metrics are detailed in Appendix D.1.

Motion Tracking We compared **CLONE** against two ablated baselines: **CLONE** † and **CLONE** *. **CLONE** † employs an MLP as the student policy, resembling the OmniH2O baseline trained on our data and task. **CLONE** * represents our **CLONE** trained on OmniH2O data instead of **CLONED**. Quantitative results in Tab. 1 reveal that both the MoE architecture and **CLONED** contribute significantly to accurate reference motion tracking. Qualitative comparisons are provided in Appendix D.3.

Table 1: **Motion tracking evaluation on CLONED dataset.** Comparison of **CLONE** against ablations: **CLONE** † uses an MLP instead of MoE architecture, while **CLONE** * trains on OmniH2O data instead of **CLONED**.

Method	SR ↑	E_{mpkpe} ↓	$E_{\text{r-mpkpe}}$ ↓	E_{vel} ↓	$E_{\text{hand-rot}}$ ↓
CLONE †	100%	113.97	35.55	245.11	4.73
CLONE *	100%	102.20	41.07	309.65	4.61
CLONE	100%	**87.84**	**33.30**	**227.17**	**3.61**

Tracking Diverse Stances To assess **CLONE**'s robustness across varying postures, we evaluated performance in tracking motions with head heights from 1.2m (standing) to 0.6m (deep squatting) in 0.1m decrements. We generated these challenging motions by systematically editing sequences from the **CLONED** dataset, creating unseen poses that test teleoperation system limits.

Fig. 7 reveals an interesting performance trade-off: while **CLONE** underperforms baselines in absolute position accuracy (MPKPE), it consistently outperforms them in local motion metrics (R-MPKPE, velocity error, and hand orientation). This pattern indicates that **CLONE** prioritizes faithful reproduction of reference motions—particularly for challenging postures—sometimes at the expense of global positioning accuracy. All methods exhibit increased tracking errors at lower heights, confirming the inherent difficulty of teleoperating robots in squatting postures.

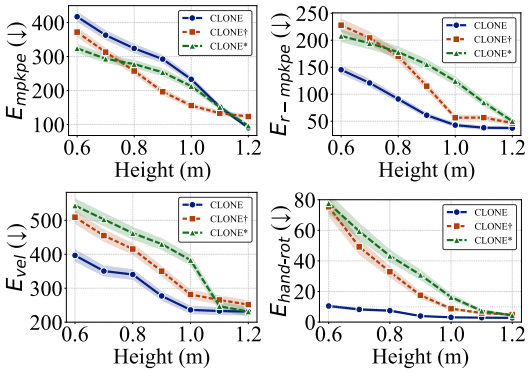


Figure 7: **Motion tracking performance across stance heights.** Comparison between **CLONE** (blue solid), **CLONE** * (green dashed), and **CLONE** † (red dashed) across different postures from standing to deep squatting. Lower values indicate better performance.

6 Conclusion

We present **CLONE**, a closed-loop MoE-based teleoperation system that enables comprehensive humanoid control while addressing accumulated tracking errors in long-horizon tasks. Our approach integrates three key innovations: (i) an MoE architecture that coordinates diverse motion skills while maintaining natural upper- and lower-body coordination, (ii) LiDAR-based closed-loop error correction that prevents positional drift accumulation through real-time feedback, (iii) and the **CLONED** dataset that augments AMASS with hand orientations and additional motion-captured sequences for robust whole-body coordination training. Experimental validation demonstrates exceptional performance: 5.1 cm mean global position tracking error over 8.9 m trajectories, accurate whole-body coordination across diverse skills including complex coordinated movements like object retrieval from ground level, and robust long-horizon teleoperation requiring only head and hand tracking from a single commercial MR headset.

7 Limitations

While **CLONE** demonstrates significant capabilities, it still has important limitations that warrant further investigation. The minimal input configuration, though enabling intuitive control with head and hand tracking, inherently limits fine-grained stability control in certain scenarios. Additionally, the system exhibits reduced performance during highly dynamic movements like jumping, stemming from training data constraints and balance control challenges.

Future work should explore additional sensing modalities to enhance stability while preserving interface simplicity, and expand motion datasets with specialized reward functions for dynamic behaviors. **CLONE** establishes new benchmarks for practical humanoid teleoperation, achieving unprecedented fidelity in complex tasks while maintaining essential simplicity.

Acknowledgments

The authors gratefully acknowledge Unitree Robotics for their support with hardware. This work is supported in part by the National Science and Technology Major Project (2022ZD0114900), the National Natural Science Foundation of China (62376009), the Beijing Nova Program, the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

[1] B. Henze, M. A. Roa, and C. Ott. Passivity-based whole-body balancing for torque-controlled humanoid robots in multi-contact scenarios. *International Journal of Robotics Research (IJRR)*, 35(12):1522–1543, 2016.

[2] P. M. Wensing and D. E. Orin. Improved computation of the humanoid centroidal dynamics and application for whole-body control. *International Journal of Humanoid Robotics*, 13(01): 1550039, 2016.

[3] L. Sentis and O. Khatib. Synthesis of whole-body behaviors through hierarchical control of behavioral primitives. *International Journal of Humanoid Robotics*, 2(04):505–518, 2005.

[4] O. Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal on Robotics and Automation*, 3(1):43–53, 2003.

[5] T. Fukuda, P. Dario, and G.-Z. Yang. Humanoid robotics—history, current state of the art, and challenges. *Science Robotics*, 2(13):eaar4043, 2017.

[6] A. Hereid, C. M. Hubicki, E. A. Cousineau, and A. D. Ames. Dynamic humanoid locomotion: A scalable formulation for hzd gait optimization. *IEEE Transactions on Robotics (T-RO)*, 34 (2):370–387, 2018.

[7] J. Cui, T. Liu, Z. Meng, J. Yu, R. Song, W. Zhang, Y. Zhu, and S. Huang. Grove: A generalized reward for learning open-vocabulary physical skill. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[8] Y. Tong, H. Liu, and Z. Zhang. Advancements in humanoid robots: A comprehensive review and future prospects. *IEEE/CAA Journal of Automatica Sinica*, 11(2):301–328, 2024.

[9] Vecna Robotics. Battlefield Extraction-Assist Robot (BEAR). https://robotsguide.com/robots/bear, 2006.

[10] U.S. Naval Research Laboratory. Autonomous Shipboard Humanoid (ASH). https://www.navy.mil/Resources/Fact-Files/Display-FactFiles/Article/2160601/shipboard-autonomous-firefighting-robot-saffir/, 2012.

[11] AGIBOT Robotics. AGIBOT A2 Humanoid Robot. https://www.agibot.com/products/A2, 2024.

[12] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning (CoRL)*, 2024.

[13] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu. Generalizable humanoid manipulation with improved 3d diffusion policies. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[14] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, C. Liu, G. Shi, X. Wang, L. Fan, and Y. Zhu. Hover: Versatile neural whole-body controller for humanoid robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[15] J. Cui, T. Liu, N. Liu, Y. Yang, Y. Zhu, and S. Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] S. Dafarra, K. Darvish, R. Grieco, G. Milani, U. Pattacini, L. Rapetti, G. Romualdi, M. Salvi, A. Scalzo, I. Sorrentino, et al. icub3 avatar system: Enabling remote fully immersive embodiment of humanoid robots. *Science Robotics*, 9(86):eadh3834, 2024.

[17] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. In *Robotics: Science and Systems (RSS)*, 2025.

[18] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning (CoRL)*, 2024.

[19] T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, J. Humplik, M. Wulfmeier, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner, et al. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *Science Robotics*, 9(89):eadi8022, 2024.

[20] A. Matsiko. Humanoid robot learning of complex behaviors with llms. *Science Robotics*, 10 (98):eadv4627, 2025.

[21] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[22] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang. Expressive whole-body control for humanoid robots. In *Robotics: Science and Systems (RSS)*, 2024.

[23] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang. Exbody2: Advanced expressive humanoid whole-body control, 2024.

[24] N. Jiang, Z. Zhang, H. Li, X. Ma, Z. Wang, Y. Chen, T. Liu, Y. Zhu, and S. Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[25] N. Jiang, Z. He, H. Li, Y. Chen, S. Huang, and Y. Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia Conference Papers*, 2024.

[26] N. Jiang, H. Li, Z. Yuan, Z. He, Y. Chen, T. Liu, Y. Zhu, and S. Huang. Dynamic motion blending for versatile motion editing. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[27] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics (T-RO)*, 38(4):2053–2073, 2022.

[28] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.

[29] K. Darvish, Y. Tirupachuri, G. Romualdi, L. Rapetti, D. Ferigo, F. J. A. Chavez, and D. Pucci. Whole-body geometric retargeting for humanoid robots. In *International Conference on Humanoid Robots (Humanoids)*, 2019.

[30] A. Brygo, I. Sarakoglou, N. Garcia-Hernandez, and N. Tsagarakis. Humanoid robot teleoperation with vibrotactile based balancing feedback. In *Haptics: Neuroscience, Devices, Modeling, and Applications*, 2014.

[31] L. Peternel and J. Babič. Learning of compliant human–robot interaction using full-body haptic interface. *Advanced Robotics*, 27(13):1003–1012, 2013.

[32] J. Ramos and S. Kim. Dynamic locomotion synchronization of bipedal robot and human operator via bilateral feedback teleoperation. *Science Robotics*, 4(35):eaav4282, 2019.

[33] J. Chagas Vaz, D. Wallace, and P. Y. Oh. Humanoid loco-manipulation of pushed carts utilizing virtual reality teleoperation. In *International Mechanical Engineering Congress and Exposition*, 2021.

[34] L. Penco, N. Scianca, V. Modugno, L. Lanari, G. Oriolo, and S. Ivaldi. A multimode teleoperation framework for humanoid loco-manipulation: An application for the icub robot. *IEEE Robotics & Automation Magazine*, 26(4):73–82, 2019.

[35] S. Tachi, Y. Inoue, and F. Kato. Telesar vi: Telexistence surrogate anthropomorphic robot vi. *International Journal of Humanoid Robotics*, 17(05):2050019, 2020.

[36] Y. Nakamura, H. Hanafusa, and T. Yoshikawa. Task-priority based redundancy control of robot manipulators. *International Journal of Robotics Research (IJRR)*, 6(2):3–15, 1987.

[37] O. Khatib, L. Sentis, J. Park, and J. Warren. Whole-body dynamic behavior and control of human-like robots. *International Journal of Humanoid Robotics*, 1(01):29–43, 2004.

[38] A. Dietrich, C. Ott, and A. Albu-Schäffer. An overview of null space projections for redundant, torque-controlled robots. *International Journal of Robotics Research (IJRR)*, 34(11):1385–1400, 2015.

[39] F. L. Moro and L. Sentis. Whole-body control of humanoid robots. *Humanoid robotics: a reference*, pages 1161–1183, 2019.

[40] M. Moniruzzaman, A. Rassau, D. Chai, and S. M. S. Islam. Teleoperation methods and enhancement techniques for mobile robots: A comprehensive survey. *Robotics and Autonomous Systems*, 150:103973, 2022.

[41] R. Huang, S. Zhu, Y. Du, and H. Zhao. Moe-loco: Mixture of experts for multitask locomotion. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[42] S. Zhou, W. Zhang, J. Jiang, W. Zhong, J. GU, and W. Zhu. On the convergence of stochastic multi-objective gradient manipulation and beyond. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[43] K. Darvish, L. Penco, J. Ramos, R. Cisneros, J. Pratt, E. Yoshida, S. Ivaldi, and D. Pucci. Teleoperation of humanoid robots: A survey. *IEEE Transactions on Robotics (T-RO)*, 39(3): 1706–1727, 2023.

[44] C. Yang, K. Yuan, Q. Zhu, W. Yu, and Z. Li. Multi-expert learning of adaptive legged locomotion. *Science Robotics*, 5(49):eabb2174, 2020.

[45] Z. Xie, S. Starke, H. Y. Ling, and M. van de Panne. Learning soccer juggling skills with layer-wise mixture-of-experts. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2022.

[46] W. Song, H. Zhao, P. Ding, C. Cui, S. Lyu, Y. Fan, and D. Wang. Germ: A generalist robotic model with mixture-of-experts for quadruped robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[47] G. Cheng, L. Dong, W. Cai, and C. Sun. Multi-task reinforcement learning with attention-based mixture of experts. *IEEE Robotics and Automation Letters (RA-L)*, 8(6):3812–3819, 2023.

[48] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez. Integrated task and motion planning. *Annual Review of Control, Robotics, and Autonomous Systems*, 4(1):265–293, 2021.

[49] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning (CoRL)*, 2023.

[50] Z. Lin, Y. Chen, and Z. Liu. Hierarchical human-to-robot imitation learning for long-horizon tasks via cross-domain skill alignment. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[51] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. *Conference on Robot Learning (CoRL)*, 2023.

[52] Z. Zhao, Y. Li, W. Li, Z. Qi, L. Ruan, Y. Zhu, and K. Althoefer. Tac-man: Tactile-informed prior-free manipulation of articulated objects. *IEEE Transactions on Robotics (T-RO)*, 41: 538–557, 2024.

[53] Z. Jiao, Z. Zeyu, W. Wang, D. Han, S.-C. Zhu, Y. Zhu, and H. Liu. Efficient task planning for mobile manipulation: a virtual kinematic chain perspective. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[54] Z. Jiao, Z. Zeyu, X. Jiang, D. Han, S.-C. Zhu, Y. Zhu, and H. Liu. Consolidating kinematic models to promote coordinated mobile manipulations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[55] Z. Jiao, Y. Niu, Z. Zhang, S.-C. Zhu, Y. Zhu, and H. Liu. Planning sequential tasks on contact graph. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[56] P. Zhi, Z. Zhang, M. Han, Z. Zhang, Z. Li, Z. Jiao, B. Jia, and S. Huang. Closed-loop open-vocabulary mobile manipulation with gpt-4v. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[57] P. Zhi, P. Li, J. Yin, B. Jia, and S. Huang. Learning unified force and position control for legged loco-manipulation. In *Conference on Robot Learning (CoRL)*, 2025.

[58] Y. Su, J. Li, Z. Jiao, M. Wang, C. Chu, H. Li, Y. Zhu, and H. Liu. Sequential manipulation planning for over-actuated unmanned aerial manipulators. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[59] D. Wisth, M. Camurri, and M. Fallon. Vilens: Visual, inertial, lidar, and leg odometry for all-terrain legged robots. *IEEE Transactions on Robotics (T-RO)*, 39(1):309–326, 2022.

[60] G. Ou, D. Li, and H. Li. Leg-kilo: Robust kinematic-inertial-lidar odometry for dynamic legged robots. *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[61] A. Allshire, H. Choi, J. Zhang, D. McAllister, A. Zhang, C. M. Kim, T. Darrell, P. Abbeel, J. Malik, and A. Kanazawa. Visual imitation enables contextual humanoid control. In *Conference on Robot Learning (CoRL)*, 2025.

[62] F. G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[63] L. Ma, Z. Meng, T. Liu, Y. Li, R. Song, W. Zhang, and S. Huang. Styleloco: Generative adversarial distillation for natural humanoid robot locomotion, 2025.

[64] H. Geng, F. Wang, S. Wei, Y. Li, B. Wang, B. An, C. T. Cheng, H. Lou, P. Li, Y.-J. Wang, et al. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. In *Robotics: Science and Systems (RSS)*, 2025.

[65] Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[66] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)*, 40(4): 1–20, 2021.