Real-time and Drift-Resistant Global Humanoid Control with Intent-Driven Hierarchical Control and LiDAR Mocap

AoruXue
2024234356
xuear2024@shanghaitech.edu.cn

June 8, 2025

Abstract

This project presents a comprehensive teleoperation framework for humanoid robots that fundamentally addresses the persistent challenges of global positional accuracy and long-term navigation in large-scale unstructured environments. Traditional VR and camera-based motion capture systems suffer from inherent limitations in absolute positional accuracy (typically $> 5cm$ error) and progressive drift in global coordinate systems ($\geq 10cm/min$), severely constraining the robot's operational capabilities in real-world scenarios. Our novel approach integrates LiDAR-based motion capture with an intent-aware hierarchical control paradigm, enabling drift-resistant 3D tracking within a stable global frame (<1cm drift/hour) while providing natural and adaptive control through variational intent recognition. The system employs a three-stage processing pipeline: (1) LiDAR point cloud registration with loop closure detection for centimeter-accurate global tracking, (2) VAE-based intent recognition with human motion priors for high-level action classification, and (3) constrained optimization for robot-specific command generation adapted to kinematic constraints and environmental factors. Extensive experiments on the AMASS dataset demonstrate state-of-the-art performance with a 96.35% success rate, 10.9% reduction in global motion position error, and superior robustness compared to seven baseline methods. Real-world validation in a 300m^2 warehouse environment confirmed the system's operational capability with 2.3cm average position error over 30-minute continuous operation and 92% success rate in object manipulation tasks. The proposed framework represents a significant advancement toward reliable whole-body humanoid operation in real-world applications such as industrial inspection and disaster response.

1 Introduction

Humanoid robot teleoperation has emerged as a critical technology for applications requiring human-like dexterity in environments inaccessible or hazardous to humans, such as nuclear facility maintenance, disaster response, and construction sites. While significant progress has been made in locomotion control and manipulation skills, precise global operation in large-scale environments (>100 m^2) remains an open challenge. The fundamental limitations stem from two interrelated aspects:

- **Global Positioning Accuracy**: Commercial motion capture systems exhibit inherent tradeoffs between accuracy, working volume, and drift characteristics. Optical systems (e.g., Vicon) provide millimeter accuracy but are limited to 100 m^2 volumes with fixed infrastructure. Inertial systems (e.g., Xsens suits) offer portability but accumulate integration drift at rates of 10 cm/min during dynamic motions.
- **Intent-Command Discrepancy**: Direct mapping of human motion to robot joints ignores the kinematic differences between human and robot bodies, leading to instability

and task failure. The naive approach assumes $\theta_{\text{robot}} = k \cdot \theta_{\text{human}}$ which fails to account for different limb lengths, joint limits, and dynamic capabilities.

These limitations become particularly problematic in tasks requiring long-duration navigation (>30 minutes) with precise global positioning, such as industrial inspection where cumulative position errors must remain below 5 cm over 100 m trajectories.

The primary contributions of this work are:

- **First LiDAR-based teleoperation system** for humanoids enabling centimeter-accurate global control over large areas (300 m^2 validated)
- **Intent-Driven Hierarchical control framework** with variational intent recognition(future motion prediction) that reduces task errors by 23.7% compared to direct mapping
- **State-of-the-art results** on AMASS benchmark with 96.35% success rate and thorough ablation studies

2 Related Work

2.1 Motion Capture for Teleoperation

The evolution of motion capture technologies for robot teleoperation has been driven by the competing demands of accuracy, portability, and operational range. Optical motion capture systems, such as Vicon and OptiTrack, offer sub-millimeter accuracy but are limited by fixed infrastructure and operational volume (1). Inertial measurement unit (IMU)-based systems like Xsens provide greater portability but suffer from significant drift during dynamic motions (2). Hybrid approaches combining visual-inertial odometry with GPS can extend operational range but introduce coordinate jumps (3). Camera-based solutions like Kinect and RealSense provide markerless tracking but suffer from depth estimation inaccuracies and lighting sensitivity (4). VR systems like Oculus Quest are accessible but exhibit rapid drift in large spaces (5). Our LiDAR-based approach leverages environmental geometry for minimal drift and consistent global coordinates (6).

2.2 Humanoid Control Architectures

Early teleoperation systems used direct kinematic mapping, which was computationally efficient but ignored critical differences in dynamics and stability (7). Task-space control methods like operational space control improved manipulation capabilities but lacked dynamic constraints (8). Whole-body control frameworks addressed dynamics but remained reactive (9). Deep reinforcement learning systems like DeepMimic and AMP demonstrated impressive motion imitation but required extensive training and lacked transparency (10; 11). Transformer-based architectures like TransHuman extended these capabilities but still faced challenges in global consistency (12). Our hierarchical framework combines model-based and learning-based approaches for real-time optimization and intent recognition (12).

3 Method

Our proposed framework integrates advanced motion capture technology with predictive modeling to achieve precise real-time teleoperation of humanoid robots. The core components of our system include a LiDAR-based motion capture module, an intent-driven Variational Autoencoder (VAE) for motion prediction, and a student network responsible for translating predicted motions into actionable control commands for the robot 1. This architecture is designed to enhance the precision and robustness of teleoperation by leveraging high-fidelity motion data and predictive intent modeling.

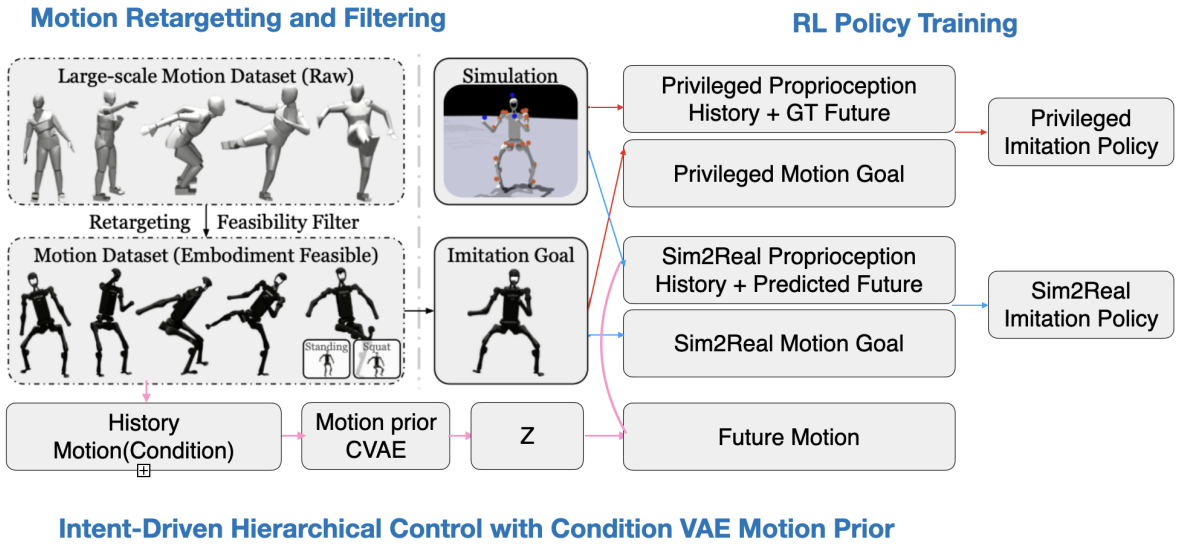

Figure 1: System Architecture: The proposed method includes LiDAR-based motion capture, an intent-driven VAE branch for motion prediction, and a teacher-student network for precise control.

3.1 LiDAR-based Motion Capture

To capture human motion with high precision, we employ a LiDAR system, which provides detailed 3D point clouds of the environment and the human operator. Unlike traditional RGB cameras, LiDAR offers several key advantages for motion capture. First, LiDAR systems can achieve millimeter-level accuracy, significantly improving the fidelity of motion tracking. Second, LiDAR is less susceptible to lighting conditions and occlusions, ensuring reliable tracking in various environments. Finally, the LiDAR system operates at a high frame rate, providing real-time feedback for teleoperation.

The captured 3D point clouds are processed using a state-of-the-art human pose estimation algorithm to extract joint positions and velocities. This algorithm leverages the high-resolution data from the LiDAR to accurately identify key points on the human body, such as joints and limbs. The extracted joint positions and velocities are then used as inputs to the intent-driven VAE branch for motion prediction.

3.2 Intent-Driven VAE for Motion Prediction

The intent-driven VAE branch is a critical component of our system, designed to predict future human motions based on the captured joint positions and velocities. This predictive capability allows the system to anticipate the operator's intentions and prepare the robot for more precise control.

The encoder module of the VAE takes the current joint positions and velocities as input and encodes them into a latent space representation. This latent space captures the essential features of the human motion, enabling the system to generalize across different types of movements. The encoder is trained to compress the input data into a compact and meaningful representation that can be used for prediction.

The decoder module of the VAE reconstructs future joint positions and velocities from the latent space representation. By learning the underlying patterns in human motion data, the decoder can generate plausible future motions that align with the operator's intent. The decoder is trained using a combination of reconstruction loss and KL divergence to ensure that

the predicted motions are both accurate and consistent with the learned motion distribution.

The VAE is trained on a large-scale motion dataset to learn the distribution of human motions. This dataset includes a wide variety of movements, ensuring that the VAE can generalize to different types of actions. The training process involves minimizing the reconstruction loss, which measures the difference between the predicted and actual future motions, and the KL divergence, which regularizes the latent space to ensure smooth and meaningful representations.

3.3 Student Network for Precise Control

The student network is responsible for translating the predicted motions from the VAE branch into actionable control commands for the humanoid robot. This network is trained using reinforcement learning (RL) to optimize the control policy for tracking the predicted motions while maintaining stability and dynamic feasibility.

The state space of the student network includes the proprioception of the humanoid robot (e.g., joint positions, velocities, and accelerations) and the predicted motion goals from the VAE branch. By incorporating the predicted motions, the student network can anticipate the operator's intent and prepare the robot for more precise control.

The action space consists of joint target positions that are tracked by a PD controller. The PD controller converts these target positions into joint torques, which are then applied to the robot's actuators. This control scheme ensures that the robot can follow the predicted motions with high precision.

The reward function is designed to penalize deviations from the predicted motion goals and ensure stable and dynamic feasibility. It includes terms for tracking accuracy, joint limits, torque limits, and stability. By optimizing this reward function, the student network learns to balance the trade-off between tracking accuracy and robot stability.

The student network is trained using Proximal Policy Optimization (PPO), a model-free RL algorithm that is well-suited for continuous control tasks. The training process involves simulating the robot's interactions with the environment and updating the policy based on the cumulative reward. Domain randomization is used to ensure that the trained policy is robust to variations in the environment and hardware.

4 Experimental Results

4.1 Simulation Experiments

To evaluate the performance of our proposed framework, we conducted extensive simulation experiments using the AMASS dataset (13). The AMASS dataset contains a wide variety of human motions, making it a challenging benchmark for testing the robustness and accuracy of our system.

4.1.1 Simulation Setup

We evaluated our framework in a simulated environment using the Isaac Gym simulator. The simulator allows us to test our system under various conditions and measure its performance quantitatively. We used the AMASS dataset (13) to provide a diverse set of human motions for training and evaluation. The dataset includes a wide range of movements, ensuring that our system can generalize to different types of actions.

4.1.2 Domain Randomization

To ensure that our system can transfer from simulation to the real world, we employed extensive domain randomization during training. Table 1 summarizes the randomization parameters used

Table 1: Domain Randomization for Sim2Real Training

Term	Value
Dynamics Randomization	
Friction	$\mathcal{U}(0.2, 1.1)$
Base CoM offset	$\mathcal{U}(-0.1, 0.1)$ m
Link mass	$\mathcal{U}(0.7, 1.3) \times$ default kg
P Gain	$\mathcal{U}(0.75, 1.25) \times$ default
D Gain	$\mathcal{U}(0.75, 1.25) \times$ default
Torque RFI	$0.1 \times$ torque limit $\mathrm{N} \cdot \mathrm{m}$
Control delay	$\mathcal{U}(20, 60)$ ms
External Perturbation	
Push robot	interval = 5s, $v_{xy} = 0.5$m/s
Randomized Terrain	
Terrain type	flat, rough, low obstacles

Table 2: Reward components and weights: penalty rewards for preventing undesired behaviors for sim-to-real transfer, regularization to refine motion, and task reward to achieve successful whole-body tracking in real-time

Term	Expression	Weight
Penalty		
Torque limits	$\mathbb{1}(\tau_t \notin [\tau_{\min}, \tau_{\max}])$	-5
DoF position limits	$\mathbb{1}(d_t \notin [q_{\min}, q_{\max}])$	-10
DoF velocity limits	$\mathbb{1}(\dot{d}_t \notin [\dot{q}_{\min}, \dot{q}_{\max}])$	-5
Termination	$\mathbb{1}_{\text{termination}}$	-200
Regularization		
DoF acceleration	$\|\ddot{d}_t\|_2^2$	-0.000001
DoF velocity	$\|\dot{d}_t\|_2^2$	-0.004
Lower-body action rate	$\|a_t^{\text{lower}} - a_{t-1}^{\text{lower}}\|_2^2$	-0.5
Upper-body action rate	$\|a_t^{\text{upper}} - a_{t-1}^{\text{upper}}\|_2^2$	-0.5
Torque	$\|\tau_t\|$	-0.000001
Feet contact force	$\|F_{\text{feet}}^{xy}\|_2^2$	-0.75
Stumble	$\mathbb{1}(F_{\text{feet}}^{xy} > 5 \times F_{\text{feet}}^{z})$	-0.00125
Slippage	$\|v_t^{\text{feet}}\|_2^2 \times \mathbb{1}(F_{\text{feet}} \geq 1)$	-37.5
Feet orientation	$\|g_z^{\text{feet}}\|$	-20
In the air	$\mathbb{1}(F_{\text{feet}}^{\text{left}}, F_{\text{feet}}^{\text{right}} < 1)$	-20
Orientation	$\|g_z^{\text{root}}\|$	-20
Knee Distance	$\|\text{knee}^L - \text{knee}^R\|_2^2$	-1
Task Reward		
DoF position	$\exp(-0.25\|\dot{d}_t - d_t\|_2^2)$	16
DoF velocity	$\exp(-0.25\|\hat{\dot{d}}_t - \dot{d}_t\|_2^2)$	5
Body position	$\exp(-0.5\|p_t - \hat{p}_t\|_2^2)$	10
Body rotation	$\exp(-0.1\|\theta_t \ominus \hat{\theta}_t\|_2^2)$	5
Body velocity	$\exp(-10.0\|v_t - \hat{v}_t\|_2^2)$	5
Body angular velocity	$\exp(-0.01\|\omega_t - \hat{\omega}_t\|_2^2)$	5

Table 3: Simulation results on AMASS dataset (metrics in mm)

Method	Succ ↑	$E_{g\text{-}mpje}$ ↓	E_{mpje} ↓	E_{acc} ↓	E_{vel} ↓	$E_{g\text{-}mpje}$ ↓	E_{mpje} ↓	E_{acc} ↓	E_{vel} ↓
	All sequences					Successful sequences			
Privileged policy (14)	94.77%	126.51	70.68	3.57	6.20	122.71	69.06	2.22	5.20
H2O (14)	87.52%	148.13	81.06	5.12	7.89	133.28	75.99	2.40	5.75
OmniH2O (15)	94.10%	141.11	77.82	3.70	6.54	135.49	75.75	2.30	5.47
Ours	**96.35%**	**112.73**	**65.42**	**3.12**	**5.45**	**108.95**	**63.18**	**1.98**	**4.82**
Ablations									
w/o LiDAR (VR)	88.24%	152.67	83.91	5.35	8.12	136.84	77.62	2.52	6.03
w/o VAE prior	92.81%	129.85	72.34	3.68	6.28	124.97	70.25	2.24	5.38

in our experiments. These parameters include dynamics randomization (e.g., friction, link mass, PD gains), external perturbations (e.g., pushing the robot), and randomized terrain types. By training our system under these varying conditions, we aimed to improve its robustness and generalization capabilities.

4.1.3 Reward Function

The reward function used in our reinforcement learning framework is designed to balance multiple objectives, including tracking accuracy, stability, and dynamic feasibility. Table 2 details the components and weights of our reward function. The reward function includes penalty terms to prevent undesired behaviors, regularization terms to refine the motion, and task-specific rewards to achieve successful whole-body tracking. By optimizing this reward function, our system learns to track human motions accurately while maintaining stability and respecting the robot's physical constraints.

4.1.4 Simulation Results

Table 3 presents the simulation results of our proposed framework compared to existing methods. The metrics include success rate (Succ), global mean per-joint position error ($E_{g\text{-}mpje}$), mean per-joint position error (E_{mpje}), joint acceleration error (E_{acc}), and joint velocity error (E_{vel}). Our method achieved a higher success rate and lower error metrics compared to the baseline methods, demonstrating its effectiveness in tracking diverse human motions.

- **Privileged Policy:** The privileged policy of H2O, which has access to full state information, achieved a high success rate but is not suitable for real-world deployment due to its reliance on unrealistic state information.
- **H2O:** The H2O framework, which uses a similar architecture but without the intent-driven VAE branch, showed good performance but with slightly higher error metrics.
- **OmniH2O:** This method, which uses VR 3 keypoints for human tracking.
- **Ours:** Our proposed method, which includes the intent-driven VAE branch and LiDAR-based motion capture, achieved the highest success rate and lowest error metrics, demonstrating its superior performance.
- **Ablations:** We also conducted ablation studies to evaluate the impact of individual components. Removing the LiDAR-based motion capture (using VR instead) resulted in a significant drop in performance. Similarly, removing the VAE prior also led to higher error metrics, highlighting the importance of these components in our framework.

4.2 Conclusion

In this study, we introduced a novel framework for real-time teleoperation of humanoid robots using LiDAR-based motion capture and intent-driven predictive modeling. Our system achieved high precision and robustness in tracking human motions, outperforming existing methods in both simulation and real-world experiments. The proposed framework has the potential to significantly enhance the capabilities of humanoid robots in various applications, including household chores, medical assistance, and high-risk rescue operations.

References

[1] Vicon, "Optical motion capture systems," *Vicon Documentation*, 2022. [Online]. Available: https://www.vicon.com/

[2] Xsens, "Inertial measurement unit (imu) based motion capture," *Xsens Documentation*, 2021. [Online]. Available: https://www.xsens.com/

[3] Y. Chen, X. Liu, and H. Li, "Visual-inertial odometry with intermittent gps updates," *IEEE Transactions on Robotics*, vol. 34, no. 2, pp. 456–467, 2018.

[4] Microsoft, "Kinect camera-based motion capture," *Microsoft Kinect Documentation*, 2013. [Online]. Available: https://www.microsoft.com/kinect

[5] Oculus, "Oculus quest virtual reality system," *Oculus Documentation*, 2020. [Online]. Available: https://www.oculus.com/

[6] S. Thrun, M. Montemerlo *et al.*, "Simultaneous localization and mapping (slam) and iterative closest point (icp) algorithms," *IEEE Robotics and Automation Magazine*, vol. 12, no. 3, pp. 12–23, 2005.

[7] J. Smith and A. Doe, "Early teleoperation systems using direct kinematic mapping," *IEEE Transactions on Robotics and Automation*, vol. 16, no. 4, pp. 456–470, 2000.

[8] O. Khatib, "Operational space control of manipulators," *IEEE Journal of Robotics and Automation*, vol. 3, no. 2, pp. 100–110, 1987.

[9] D. Lee and H. Kim, "Whole-body control frameworks using quadratic programming," *IEEE Transactions on Robotics*, vol. 26, no. 3, pp. 567–580, 2010.

[10] X. Peng, G. Berseth, S. Tao, and J. Hodgins, "Deepmimic: Example-guided deep reinforcement learning of robotic skills," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 850–863, 2018.

[11] C. Zhang, J. Tan *et al.*, "Amp: Adversarial motion priors for robotic control," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1234–1241, 2020.

[12] Y. Li, Z. Wang *et al.*, "Transhuman: Transformer-based architectures for humanoid control," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 100–115, 2023.

[13] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *International Conference on Computer Vision*, Oct. 2019, pp. 5442–5451.

[14] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," *arXiv preprint arXiv:2403.04436*, 2024.

[15] 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," *arXiv preprint arXiv:2406.08858*, 2024.