2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

TWIST2

Scalable, Portable, and Holistic Humanoid Data Collection System

Yanjie Ze[12] Siheng Zhao[13] Weizhuo Wang[12]
Angjoo Kanazawa[14†] Rocky Duan[1†] Pieter Abbeel[14†] Guanya Shi[15†] Jiajun Wu[2†] C. Karen Liu[12†]
[1]Amazon FAR [2]Stanford University [3]USC [4]UC Berkeley [5]CMU [†]Equal Advising

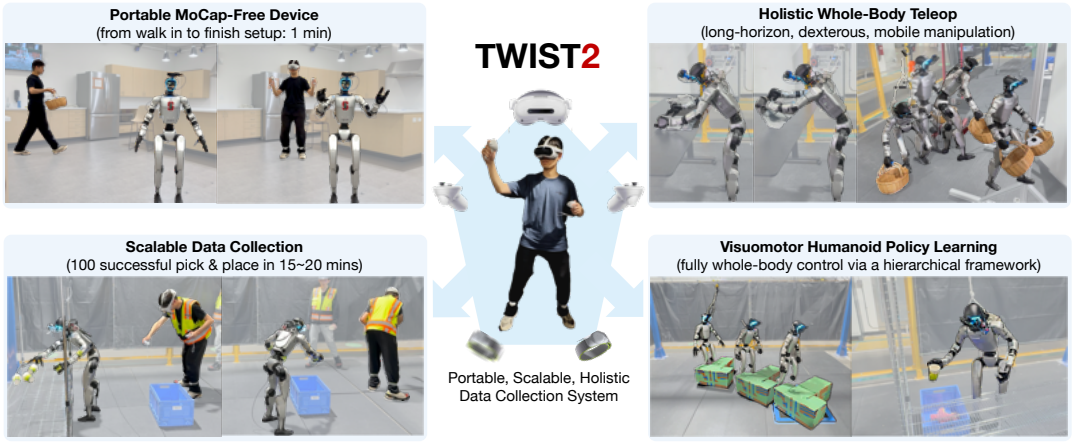

Fig. 1: We introduce TWIST**2**, a holistic humanoid data collection system designed with scalability and portability. TWIST**2** enables scalable data collection, fast setup, and enjoyable user experience compared to MoCap solutions such as TWIST [1], while maintaining the full whole-body control. We build a 2-DoF Neck (TWIST**2** Neck) to enable egocentric teleoperation, which costs $250. With TWIST**2**, our robots are able to perform long-horizon, dexterous, mobile whole-body manipulation and legged manipulation. All tasks are achieved with streamed robot egocentric vision, full whole-body control, and a single operator. We further train visuomotor policies upon data collected via TWIST**2**. Our entire system is open-sourced at https://yanjieze.com/TWIST2 and ensure full reproducibility.

***Abstract*— Large-scale data has driven breakthroughs in robotics, from language models to vision-language-action models in bimanual manipulation. However, humanoid robotics lacks equally effective data collection frameworks. Existing humanoid teleoperation systems either use decoupled control or depend on expensive motion capture setups. We introduce TWIST2, a portable, mocap-free humanoid teleoperation and data collection system that preserves full whole-body control while advancing scalability. Our system leverages PICO4U VR for obtaining real-time whole-body human motions, with a custom 2-DoF robot neck (cost around $250) for egocentric vision, enabling holistic human-to-humanoid control. We demonstrate long-horizon dexterous and mobile humanoid skills and we can collect 100 demonstrations in 15 minutes with an almost 100% success rate. Building on this pipeline, we propose a hierarchical visuomotor policy framework that autonomously controls the full humanoid body based on egocentric vision. Our visuomotor policy successfully demonstrates whole-body dexterous manipulation and dynamic kicking tasks. The entire system is fully reproducible and open-sourced at https://yanjieze.com/TWIST2. Our collected dataset is also open-sourced at https://twist-data.github.io.**

I. Introduction

The transformative power of large-scale data has fundamentally reshaped machine learning, driving breakthrough achievements from large language models like GPT-4 [5] to the recent success of vision-language-action (VLA) models in robotics. In the realm of bimanual manipulation, models such as π_0 [6] and $\pi_{0.5}$ [7] have demonstrated unprecedented capabilities, directly enabled by the robust and scalable data collection infrastructure [8]–[10]. However, this data-driven revolution has yet to reach humanoid robots, where the absence of equally effective data collection frameworks continues to limit progress toward human-level versatile manipulation and locomotion.

As summarized in Table I, existing humanoid teleoperation systems fall into three broad categories: a) *Decoupled control* of lower and upper body (e.g., MobileTV [11], HOMIE [2]); b) *Partial whole-body control* that coordinates selected body segments such as arms and torso while legs track base velocity commands (e.g., AMO [3], CLONE [4]); c) *Full whole-body control* that directly tracks human body pose across all joints including arms, torso, and legs in a unified manner (e.g., HumanPlus [12], TWIST [1]). Among these, VR-based solutions such as AMO and CLONE offer practicality but are limited to mobile skills with simple locomotion, falling short of capturing dynamic whole-body coordination skills that humans naturally exhibit. In contrast, full whole-body control holds the greatest promise for unleashing the versatility of

[1]Work done during the internship of Yanjie Ze, Siheng Zhao, and Weizhuo Wang at Amazon Frontier AI & Robotics (FAR).

TABLE I: **Comparison of recent humanoid data collection systems**. We compare existing humanoid teleoperation systems across key dimensions essential for effective data collection. TWIST**2** is the first system to combine full whole-body control with portability, achieving comprehensive capabilities including egocentric teleoperation, accurate tracking, and single-operator efficiency. Unlike previous works that either sacrifice portability for full whole-body control (TWIST) or sacrifice full whole-body control for portability (AMO, CLONE), our system achieves all critical requirements for scalable humanoid data collection.

Humanoid Data Collection System	Category	Portability & Scalability				Holistic Control			
		Source	Portable	No Calibration	Single Operator	Whole-Body Tracking	Egocentric Teleop	Foot Control	Wrist Control
HOMIE [2]	Decoupled	Exoskeleton	✗	✓	✓	✗	✗	✗	✓
AMO [3]	Partial	VR	✓	✓	✗	✗	✓	✗	✓
CLONE [4]	Partial	VR	✓	✓	✓	✗	✗	✗	✓
TWIST [1]	Full	MoCap	✗	✗	✓	✓	✗	✓	✗
TWIST2 (ours)	Full	VR	✓	✓	✓	✓	✓	✓	✓

humanoid robots, as evidenced by TWIST [1]. However, such systems typically depend on expensive, non-portable motion capture setups, restricting deployment to lab environments.

In this work, we introduce TWIST**2**, a humanoid teleoperation and data collection system that preserves the power of full whole-body control while advancing portability and scalability. Our design leverages PICO4U [13], a lightweight VR device that provides whole-body motion streaming using a head goggle, handheld controllers, and two motion trackers on the ankles, without requiring expensive motion capture systems. Recognizing that egocentric vision is crucial for human-like task execution, we design a low-cost and non-invasive neck that seamlessly integrates with Unitree G1 and our VR teleoperation ecosystem. With these portable components, we build a comprehensive retargeting pipeline from full human body poses of PICO to corresponding humanoid motor joint positions. To execute the retargeted motions on the robot, we train a robust motion tracking controller using reinforcement learning and large-scale simulation interaction on carefully curated motion data.

These elements together enable efficient, long-horizon, in-the-wild teleoperation and data collection without reliance on motion capture systems, and only requiring a single operator. We showcase that 1) we can teleoperate robots to perform very long-horizon and fine-grained whole-body dexterous skills such as folding towels and mobile skills such as transporting objects through the door, and 2) we can collect human demonstrations efficiently, *e.g.*, collecting around 100 successful demonstrations in 20 minutes without failure. We also find that egocentric active stereo vision is essential for the long-horizon mobile and dexterous teleoperation.

Building on this scalable data collection pipeline, we further propose a hierarchical visuomotor policy learning framework consisting of two components. The first component is the same motion tracking controller used during teleoperation, which serves as a low-level controller. The second component is a Diffusion Policy that directly predicts whole-body joint positions based on visual observations that feeds into the low-level controller. To our knowledge, this is the first policy learning framework that enables vision-based autonomous control of the full humanoid body, moving beyond simplified commands such as root velocity. Importantly, this capability is made possible by our data collection system, which provides the high-quality demonstrations needed for training.

We showcase a few representative results where our humanoid robot autonomously performs a) consecutive whole-body dexterous pick & place and b) continuous kicking of a T-shaped box to target regions (Kick-T), illustrating the potential of this new framework.

To summarize, our main contributions are:

1) A portable, mocap-free humanoid teleoperation and data collection system with full whole-body control, enhanced with an attachable neck for egocentric active vision.
2) A hierarchical whole-body visuomotor policy learning framework that achieves full whole-body control.
3) Demonstration of long-horizon teleoperation skills such as towel folding/unfolding and object transporting through the door, effective data collection, and new autonomous humanoid skills including whole-body dexterous pick & place and Kick-T.

Our system, data, and model are fully open-sourced at https://yanjieze.com/TWIST2 to ensure full reproducibility.

II. Related Work

A. Whole-Body Humanoid Teleoperation

Teleoperation is crucial for enabling humanoid robots to interact with complex real-world environments and perform sophisticated loco-manipulation tasks. Unlike wheel-based robots or tabletop arms, the anthropomorphic nature of humanoids makes whole-body control the most natural and effective teleoperation approach [1], [3], [4], [12], [14]–[16]. As shown in Table I, we categorize recent works into three categories: a) decoupled control, b) partial whole-body control, and c) full whole-body control. Full whole-body control, as demonstrated by TWIST [1], shows promising results in coordinated whole-body dexterity, which is the primary focus of this work. As detailed in Table I, we identify several critical aspects in scalable & holistic teleoperation and data collection that remain lacking in previous works for real-world deployment. which we address comprehensively in this work.

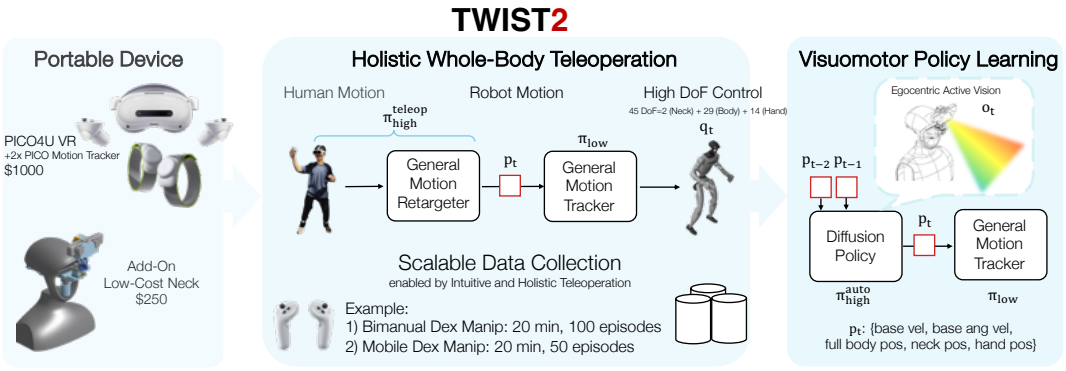

Fig. 2: System overview of TWIST**2**. We build a holistic humanoid teleoperation system with portable devices and egocentric active vision, enabling scalable imitation data collection. With data collected, we build a hierarchical visuomotor policy learning framework that directly predicts whole-body joint positions.

B. Visual Humanoid Control

Previous works on visual humanoid control predominantly rely on LiDAR for perceptive locomotion [17]–[19], typically employing task-specific sim-to-real reinforcement learning (RL) approaches. Recent works like HEAD [20] propose keypoint-based hierarchical frameworks with humanoid egocentric vision, while limiting in simple navigation tasks. VideoMimic [18] introduced a real2sim2real pipeline that enables real robots to perform environment interactions such as sitting, though their interactions remain limited to static settings like the ground or stone chairs. Some works such as PDC [21] are conducted only in simulation and face significant sim-to-real transfer challenges. In contrast, our work focuses on developing general visuomotor humanoid policies that can interact with complex environments and perform long-horizon whole-body loco-manipulation and legged manipulation tasks—capabilities not demonstrated in previous works.

III. Our System

We introduce TWIST**2**, a scalable, portable, and holistic humanoid teleoperation and data collection system (see Figure 1 for capabilities). As illustrated in Figure 2, our system consists of four main components: a humanoid robot equipped with active vision (Section III-B), portable motion capture using VR devices (Section III-C), holistic human-to-robot motion retargeting (Section III-D), a general motion tracker for low-level control (Section III-E). These components work together to enable scalable data collection (Section III-F) and autonomous visuomotor policy execution (Section III-G).

A. Problem Formulation

We focus on enabling humanoid robots to perform diverse whole-body dexterous tasks with their own egocentric vision and proprioception within a single unified framework. To this end, we propose a two-level hierarchical control framework, consisting of a low-level controller π_{low} and a high-level controller π_{high}.

Low-level control. We formulate the low-level controller π_{low} as a *general motion tracking* problem, so that our low-level control is task-agnostic. At each timestep, the low-level controller receives a reference command vector composed of root translational velocity in the x and y axes, root z position, root roll/pitch angles, root yaw angular velocity, and whole-body joint positions:

$$\mathbf{p}_{\text{cmd}} = \left[\dot{x}_{\text{ref}},\ \dot{y}_{\text{ref}},\ z_{\text{ref}},\ \phi_{\text{ref}},\ \theta_{\text{ref}},\ \dot{\psi}_{\text{ref}},\ \mathbf{q}_{\text{ref}}\right]. \tag{1}$$

In addition, it has access to robot proprioception, including root orientation and angular velocity from IMU readings, as well as joint positions and velocities from encoders:

$$\mathbf{s} = \left[\omega,\ \dot{\omega},\ \mathbf{q},\ \dot{\mathbf{q}}\right]. \tag{2}$$

The controller outputs desired joint positions,

$$\mathbf{q}_{\text{tgt}} = \pi_{\text{low}}(\mathbf{s}, \mathbf{p}_{\text{cmd}}), \tag{3}$$

at 50Hz, which are then tracked by a PD controller to generate the final torque:

$$\tau = K_{\text{P}}\left(\mathbf{q}_{\text{tgt}} - \mathbf{q}\right) - K_{\text{D}}\,\dot{\mathbf{q}}. \tag{4}$$

High-level control. The high-level controller π_{high} focuses on generating task-specific motion commands $\mathbf{p}_{\text{cmd}}$ conditioned on egocentric vision. We have two variants in this work: (1) a teleoperation policy $\pi^{\text{teleop}}_{\text{high}}$, and (2) a visuomotor policy $\pi^{\text{auto}}_{\text{high}}$. Both map visual observations $\mathbf{o}$ and proprioceptive states $\mathbf{s}$ into commands:

$$\mathbf{p}_{\text{cmd}} = \pi_{\text{high}}(\mathbf{o}, \mathbf{s}). \tag{5}$$

In this work, we employ $\pi^{\text{teleop}}_{\text{high}}$, *i.e.*, the human teleoperator plus the motion retargeter, to collect observation–action pairs $(\mathbf{o}, \mathbf{s}, \mathbf{p}_{\text{cmd}})$, which are then used to train $\pi^{\text{auto}}_{\text{high}}$, *e.g.*, a Diffusion Policy.

Interface design. There are two key aspects of our command interface $\mathbf{p}_{\text{cmd}}$: (1) We use relative root translations/rotations rather than absolute poses, so that our system does not rely on accurate global state estimation [22], and remains stable during very long-horizon operation; (2) We include whole-body joint positions instead of simplifying lower-body control as root velocity only [3], [4], [11], which enables more precise control of lower-body movements and unlocks tasks such as legged manipulation and dancing.

B. Humanoid Robot with Active Vision

We use Unitree G1 with 29 DoF (3 DoF waist + two 6 DoF legs + two 7 DoF arms), equipped with two 7 DoF Dex31 hands. We find that neck DoFs are essential for effective and long-horizon teleoperation, so we build a portable robot neck with yaw and pitch DoFs.

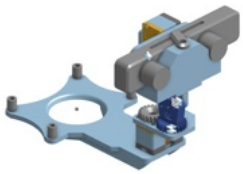

Fig. 3: TWIST**2** Neck. We design a simple yet effective 2-DoF neck that can be easily assembled for a non-expert user and can be attached/detached to/from a Unitree G1 without removing the original LiDAR.

Add-on low-cost neck (TWIST2 Neck). Unlike recent works [3], [23] that build built-in necks, we design an add-on neck module that can be seamlessly attached to the Unitree G1 without disassembling its original head (see Figure 3). Our design is inspired by ToddlerBot [24]. We use two Dynamixel XC330-T288 motors to control the yaw and pitch angles, connected via a U2D2 and powered by the onboard 12V/5A supply. All structural parts are 3D printed. The cost of the neck is $250. We use Zed Mini as our stereo camera attached to the neck (the ZED Mini stereo camera will cost extra $400). Since human roll DoF is rarely used in everyday interaction, we find that the two-DoF design already enables smooth and human-like neck motions (Figure **??**).

C. Portable MoCap-Free Whole-Body Human Data Source

To obtain real-time full human body poses in a portable manner, we utilize PICO 4U [13] combined with two PICO Motion Trackers [25] that are bound on the humans' calves to obtain global translations and rotations for each human body parts. Though PICO supports more than 2 motion trackers, we find the 2-tracker mode provides a more stable pose estimation. The cost for such a setup is around $1000. much cheaper and practical compared to an optical MoCap system. We use XRoboToolkit [26] for access to motion streaming from PICO. The motion can be streamed at 100Hz. Notably, PICO does not require heavy calibration compared to the MoCap system. As shown in Figure 1, it takes around only 1 minute to finish the setup of PICO.

Compared to HTC Vive Tracker [27] that is used in recent demos of Boston Dynamics [28], PICO's whole-body estimation does not require extra third-person view camera setup, thus more flexible.

D. Holistic Human-to-Humanoid Retargeting

In this section, we describe how human motion data is holistically leveraged to control the humanoid robot's body, hands, and neck.

Body retargeting. We adapt GMR [1], [29], a real-time motion retargeting method, to the PICO human motion format. The original GMR employs a two-stage optimization: (1) solving for link rotation consistency, and (2) refining global pose alignment. Since PICO motion capture often yields inaccurate global pose estimation, we modify the second optimization stage as follows: 1) for the lower body, optimize for position and rotation constraints; 2) for the upper body, only optimize for rotation constraints. This ensures 1) less feet sliding and 2) better upper-body teleportation experience.

We partition the retargeted links into lower-body $\mathcal{L}_{\text{low}}$ (e.g., pelvis, hips, knees, ankles, feet) and upper-body $\mathcal{L}_{\text{up}}$ (e.g., spine, shoulders, elbows, wrists, head). Let R_i^{human} and $R_i^{\text{robot}}(\mathbf{q})$ be the link orientations, and p_k^{human} and $p_k^{\text{robot}}(\mathbf{q})$ the link positions for a selected set of lower-body points $\mathcal{P}_{\text{low}}$ (typically feet/ankles and optionally pelvis). To reduce sensitivity to noisy global pose estimation (and to support user teleportation), we measure all human positions in a pelvis-centric frame. The stage-2 optimization is then formulated as:

$$\mathbf{q}^* = \arg\min_{\mathbf{q}} \sum_{i\in\mathcal{L}_{\text{up}}\cup\mathcal{L}_{\text{low}}} w_i^R \left\| R_i^{\text{human}} - R_i^{\text{robot}}(\mathbf{q}) \right\|_F^2 + \lambda_{\text{pos}} \sum_{k\in\mathcal{P}_{\text{low}}} w_k^p \left\| p_k^{\text{human, pelvis}} - p_k^{\text{robot}}(\mathbf{q}) \right\|_2^2. \tag{6}$$

Here w_i^R and w_k^p are per-link weights, λ_{pos} balances the rotation and position terms, and $p_k^{\text{human, pelvis}}$ denotes human keypoints expressed in the human pelvis frame. This formulation enforces accurate foot and ankle placement to mitigate foot sliding, while keeping the upper body free of positional terms so that global-pose jumps (e.g., teleportation) do not introduce artifacts—upper-body retargeting depends only on local rotations.

Hand retargeting. Directly mapping a human five-finger hand to the Unitree Dex31 hand is not intuitive for teleoperation, since the Dex31 only provides three fingers with limited degrees of freedom. In practice, the functionality of the Dex31 hand is much closer to a parallel-jaw gripper than to a dexterous multi-fingered hand. Therefore, we simplify hand retargeting by treating the Dex31 as a gripper and not using hand pose estimation but controlling it by pressing buttons with PICO handheld controllers. We define two canonical configurations: an *open pose* $\mathbf{q}_{\text{open}}$ and a *close pose* $\mathbf{q}_{\text{close}}$. A scalar grasp command $\alpha \in [0, 1]$ is computed from the human hand signals, where $\alpha = 0$ denotes fully open and $\alpha = 1$ denotes fully closed. The commanded Dex31 hand joint configuration is then interpolated as

$$\mathbf{q}_{\text{hand}} = (1-\alpha)\,\mathbf{q}_{\text{open}} + \alpha\,\mathbf{q}_{\text{close}}. \tag{7}$$

For tasks that require power grasp (*e.g.*, grasp a cup) and tasks that require fine-grained pinching (*e.g.*, folding cloths), we define two sets of $\mathbf{q}_{\text{open}}$ and $\mathbf{q}_{\text{close}}$.

Neck retargeting. Let $R_{\text{head}}, R_{\text{spine}} \in SO(3)$ be the global rotations of the human head and spine in the world frame, respectively. The relative rotation is

$$R_{\text{rel}} = R_{\text{spine}}^{\top} R_{\text{head}}. \tag{8}$$

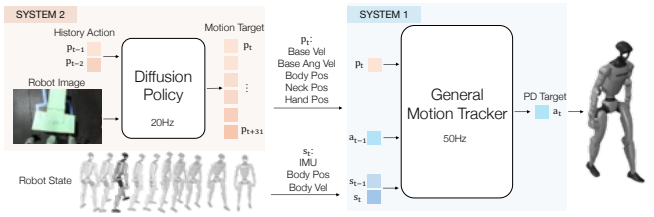

Fig. 4: Hierarchical whole-body visuomotor policy learning framework built upon data collected via TWIST**2**. Unlike previous works that focus on upper-body manipulation or lower-body locomotion separately, our visuomotor policy controls the entire body, enabling complex tasks such as Kick-T that require coordinated whole-body movements.

From $R_{\text{rel}} = [r_{ij}]$, the robot neck joint targets are defined as

$$q^{\text{yaw}}_{\text{neck}} = \psi = \arctan 2(r_{21}, r_{11}),\ q^{\text{pitch}}_{\text{neck}} = \theta = \arcsin(-r_{31}). \tag{9}$$

E. Training General Motion Trackers for Low-Level Control

To bring the retargeted kinematics motions onto a physical robot, we need a whole-body controller π_{low} that takes into reference motions and outputs the desired PD target. Different from previous works that adopt a complex teacher-student pipeline to train a reasonable whole-body controller [1], [14], [30], we design a simple one-stage training framework for general motion tracking.

More specifically, we first curate a humanoid motion dataset consisting of around 20k motion clips. The motion dataset includes data retargeted via GMR [1], [31] (7k clips) and the original motion dataset from TWIST [1] (13k clips). The motion data source includes AMASS [32], OMOMO [33], and our in-house MoCap data. This mixture of the dataset ensures our policy learns omnidirectional walking. Similarly as found in TWIST [1], we find that curating a small set of motions from the teleoperation device is essential to bridge the domain gap. We only collect 73 motions via PICO, as these motions already cover most daily movements like walking, crouching, and manipulation. We then generate reward supervision from the motion datasets. The rewards are defined as $r = r_{\text{track}} + r_{\text{reg}}$, where r_{track} is defined as:

$$r_{\text{track}} = e^{-\alpha \|\mathbf{p}_{\text{cmd}} - \mathbf{p}_{\text{cur}}\|} \tag{10}$$

where $\mathbf{p}_{\text{cur}}$ denotes the actual state the robot achieved. r_{reg} consists of the regularization terms, such as the penalty on the action change.

The actor π_{low} is trained via PPO and mainly consists of two parts: the convolutional history encoder and the MLP backbone. We find that compressing history robot proprioceptions and history reference motions into a compact latent vector boosts learning efficiency.

F. Scalable Humanoid Data Collection

We now describe our humanoid teleoperation and data collection system built with the aforementioned modules.

Egocentric whole-body teleoperation. During teleoperation, we obtain real-time streamed human motions from PICO (Section III-C) and map human motions into robot motion commands $\mathbf{p}_{\text{cmd}}$, and then send $\mathbf{p}_{\text{cmd}}$ to π_{low} (Section III-E) through Redis [34]. Additionally, our teleoperation system is equipped with stereoscopic vision via the custom shader implemented in [26] that adjusts the interpupillary distance and sets the focal point at approximately 3.3 feet, providing teleoperators with depth perception. The stereo images are streamed from ZED Mini to PICO via GStreamer in the h265 format and to the data collection process via ZMQ in the JPEG format.

Single operator. A practical teleoperation/data collection system should only require a single operator. Recent whole-body humanoid teleoperation systems focus on showing their capabilities [1], [3], [4], [11], but most of them do not explicitly show how the teleoperation sessions start, pause, and terminate. AMO [3] and MobileTV [11] require two operators: one for the upper body and one for the lower body. TWIST [1] and CLONE [4] require only one operator for teleoperating the robot, but need another one to control the start/end of the entire process. We program the PICO's handheld controllers to allow the demonstrator to safely and smoothly operate the entire system without the need for any assistance. The handheld controllers play the role of the control center, as shown in Figure 5.

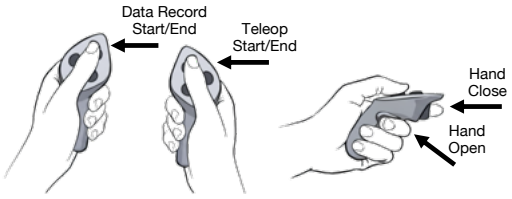

Fig. 5: Illustrations on using the PICO joystick controller as the control center to make TWIST**2** a single-operator system.

Safe control. Humanoid robots are brittle; and this problem becomes more critical when designing a system that can fully control the robot. In TWIST**2**, we use motion interpolation for smooth state transition. For example, our system supports *pause* via the origin joystick from PICO; and when *pause* mode ends, we interpolate from the last robot pose to current target pose, to avoid sudden jump. This guarantees our system can operate in a quite long time safely and stop anytime when human operators are tired.

System delay. All modules in our system stream at a speed above 50Hz, ensuring the overall delay to be lower than 0.1s, significantly improved upon prior work [1] (0.5s delay).

Data filtering. During data collection, we consecutively record episodes. To process these trajectories, we developed a demonstration post-processing GUI that segments long sequences into multiple episodes, each corresponding to a completed task. We also reduce idle actions and remove failure episodes through filtering.

G. Whole-Body Visuomotor Policy Learning

Using the high-quality demonstration data collected through our teleoperation system, we develop a hierarchical visuomotor policy framework, as illustrated in Figure 4.

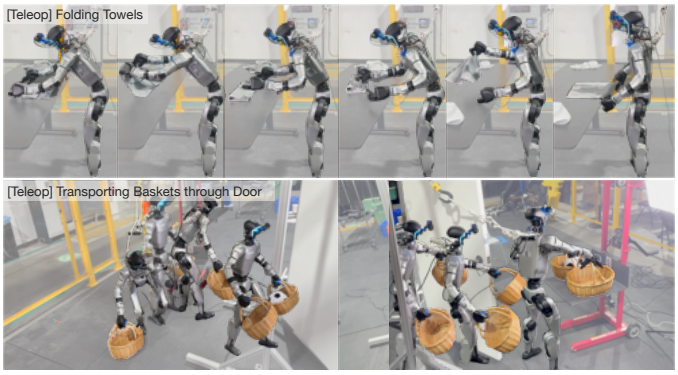

Fig. 6: Long-horizon humanoid teleoperation powered by TWIST**2**. All tasks are achieved with streamed robot egocentric vision, full whole-body control, and a single operator.

This section details the design and training of the high-level visuomotor policy $\pi^{\text{auto}}_{\text{high}}$.

Observation and action space. The visuomotor policy operates on visual observations and proprioceptive information to generate motion commands. Visual input consists of 360×640 RGB images captured by the ZED Mini camera, which are downsampled to 224×224 for computational efficiency. For robot proprioception, we use the historical command sequence $\mathbf{p}_{\text{cmd}}$ rather than raw robot states $\mathbf{s}$. This choice of proprioception serves two purposes: 1) it decouples the high-level policy from the low-level controller, enabling modular training and deployment, and 2) it mitigates error accumulation in this high-dimensional system by avoiding direct dependence on noisy raw robot states s. The action space consists of the same command vector $\mathbf{p}_{\text{cmd}}$ used during teleoperation, ensuring consistency between data collection and policy execution. All proprioceptive inputs are normalized to improve training stability.

Network architecture. We employ Diffusion Policy [35] as our policy learning framework, utilizing 1D convolutional blocks for temporal modeling of action sequences. The policy predicts 64 action chunks using sample-based prediction [16], [36], corresponding to 2 seconds of future motion commands at the policy execution frequency. For visual encoding, we use a ResNet-18 backbone pre-trained with R3M [37], which provides robust visual representations learned from diverse robotic datasets.

Data augmentation and regularization. To enhance the robustness and generalization of the learned policy, we apply both state-space and visual augmentations. We inject 10% Gaussian noise into the proprioceptive inputs, encouraging the policy to rely more heavily on visual observations rather than overfitting to precise state information. For visual augmentation, we employ a comprehensive set of techniques including random cropping, random rotation, and color jittering. These augmentations improve the policy's ability to generalize across different lighting conditions, camera viewpoints, and visual variations that may occur during deployment.

Deployment and inference. For efficient real-time execution, the trained Diffusion Policy is converted to ONNX format, achieving a 20Hz inference rate on a single NVIDIA RTX 4090. We execute 48 out of the predicted 64-step action chunks at 30Hz, maintaining consistency with the data collection frequency.

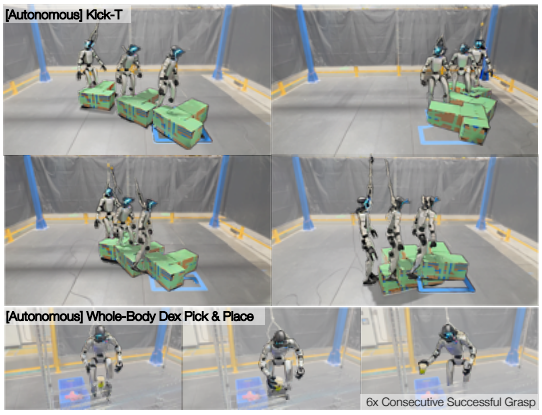

Fig. 7: Closed-loop whole-body visuomotor policy execution in the real world. TWIST**2** enables effective and holistic whole-body humanoid data collection, which further enables versatile autonomous whole-body humanoid loco-manipulation & legged manipulation skills.

IV. Experiment Results

In this section, we show that powered by TWIST**2**, we can 1) teleoperate Unitree G1 to perform long-horizon challenging whole-body dexterous tasks, 2) collect imitation learning data effectively, and 3) make Unitree G1 autonomously perform whole-body tasks via its egocentric vision.

A. Long-Horizon Teleoperation

TWIST**2** enables very long-horizon teleoperation. We showcase two representative tasks that cannot be achieved by previous systems (see Figure 6). We observe that 1) egocentric active perception and 2) smooth whole-body tracking instead of decoupled control are keys that enable such natural & smooth, long-horizon, whole-body, and mobile tasks.

Folding towels. The robot uses its egocentric vision to locate the towel, move the towel to its front, grasp it, and shakes it to spread. Then it will pinch the corner to fold the towel in half with two hands. It repeats the motion to fold into thirds (or quarters) to the target size, presses along the crease to set it, and neatly places the finished towel to its left-hand side. The entire process requires fine-grained control of the wrists and hands, active vision, and whole-body reaching. Our robot can continuously fold 3 towels that are randomly placed on the table for now; and this is only bottlenecked by the underlying motor robustness, such as motor overheating.

Transporting baskets through the door. The robot first adjusts its position via changing foot placements and bends down to pick up the baskets on its left side and on its right side, respectively. We casually put the basket so the teleoperator seeks the basket first via robot active perception. Then the robot moves close to the door, pushes the door open with the arm, walks across the door, and places the basket gently onto the shelf. Note that all the base movements of

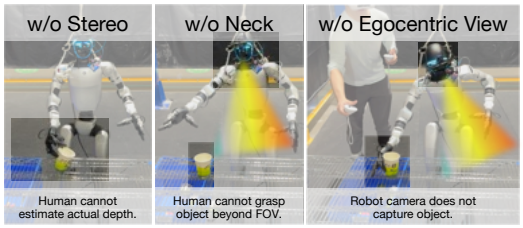

Fig. 8: Comparison of different teleoperation settings.

the robot are achieved via a single teleoperator by tracking the lower-body movements.

B. Efficient Data Collection

We show that 1) how effective TWIST**2** is in collecting imitation learning data and 2) how some key designs in our system improves data collection.

First, we show in Table II that within 20 minutes, the expert teleoperator can consecutively collect 1) around 100 successful bimanual pick&place or 2) around 50 successful mobile pick&place.

TABLE II: Scalable data collection. We show that we can easily collect several demonstrations via our system.

Task	Time	#Collected Episodes	Success Rate	Avg Time Per Episode
Bimanul Manip 1	18.5 min	98	100%	11 s
Mobile Manip	19.5 min	46	100%	25 s

Second, we conduct a user study to quantify the effectiveness of our data collection system. We evaluate two users: 1) an **expert** who has extensive experience using this system for data collection, and 2) a **novice** who is using the system for the first time during the test. Since the novice user gains proficiency through practice, we have them start with our complete system and then progressively remove features to isolate the impact of each component. As shown in Table III, TWIST**2** achieves the shortest completion times and highest success rates across all configurations.

As illustrated in Figure 8, we observe several key findings: 1) without stereo vision for teleoperation, users tend to grasp higher than the actual object location, significantly increasing grasp failure rates; 2) without the neck module, users cannot perceive objects beyond the fixed field of view, making teleoperation extremely challenging; 3) when using third-person view with VR pass-through (*i.e.*, *w/o Egocentric View*), the expert can collect data remarkably fast (10 episodes in 43 seconds), but this is only possible because the expert stands directly beside the robot, which is infeasible for long-horizon mobile manipulation tasks which require remote control via egocentric vision.

TABLE III: Data collection efficiency of TWIST**2** on different users and setups. The results show the necessity of using active egocentric stereo vision.

Collect 10 Demos	**Succes/Total Trials**			**Time Cost (s)**		
	Novice	Expert	Avg (Sum)	Novice	Expert	Avg (Mean)
TWIST**2**	10/12	10/11	**20/23**	75.6	59.9	**67.8**
w/o Stereo	10/12	10/15	20/27	90.6	105.9	98.3
w/o Neck	7/17	9/12	16/29	144.0	80.5	112.3
w/o Egocentric View	10/13	10/10	**20/23**	94.3	43.0	68.7

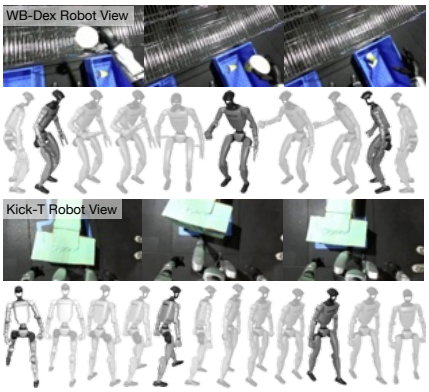

Fig. 9: Visualization of training demonstrations (egocentric robot view and whole-body joint positions) for WB-Dex and Kick-T tasks.

C. Whole-Body Policy Learning Results

We design two tasks to showcase autonomous results with our hierarchical visuomotor policy framework. We visualize the training data in Figure 9.

Whole-body dexterous pick & place (WB-Dex). In this task, the robot bends down to pick up a cup from the shelf using its dexterous hand and places it into a box on the ground. We train the policy with 170 human demonstrations and report the success and failure rates in Figure 10. We observe that the policy can reliably reach the cup in most cases. However, because the cup is very light, grasping it requires highly precise control; even a slight drift often results in grasp failure.

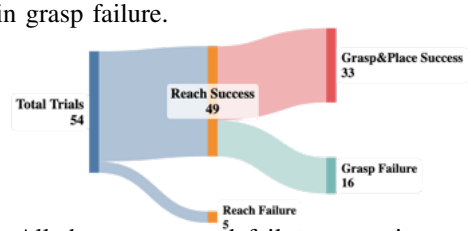

Fig. 10: All the success and failure cases in our WB-Dex task.

Kick T-shaped box to target (Kick-T). In this task, the robot uses its foot to kick a T-shaped green box toward a fixed T-shaped target position on the ground. The policy is trained with 50 demonstrations. In our data, the action pattern is consistent: the robot kicks with its left foot, and then takes a step forward with the right foot to maintain balance. This design ensures that the learned policy exhibits robust kicking behavior. We visualize policy rollouts in Figure 7. The policy successfully transports the T-shaped box to the target in 6 out of 7 trials. At present, the policy can only kick the box forward, without more flexible strategies such as walking around the box to adjust the kicking angle; we leave such capabilities to future work.

V. Conclusions and Limitations

We introduce TWIST**2**, a portable and holistic mocap-free data collection system for humanoid robots with full whole-body control. By combining lightweight VR devices with an attachable neck for egocentric vision, our framework

enables scalable data collection. On top of this, we designed a hierarchical visuomotor policy that allows a real humanoid robot to autonomously perform versatile whole-body skills including whole-body dexterous manipulation and Kick-T.

Limitations. 1) The general motion tracker struggles with highly dynamic movements such as sprinting due to challenges in tracking fast, complex motions. 2) PICO's whole-body pose estimation is less accurate than high-cost motion capture systems, particularly for elbows and knees where no trackers are placed, resulting in reduced motion quality.

ACKNOWLEDGMENTS

We want to thank Charlie Cheng, Shaofeng Yin, Yuanhang Zhang, Yunchu Zhang, and Raven Huang for their help in real-world experiments. We also thank Wenhao Wang, Ke Jing, Ning Yang, Liuchuan Yu, and Zhigen Zhao for the helpful discussion in the PICO usage. The human motion datasets used in this work, including AMASS [32] and OMOMO [33], are solely for research purposes.

REFERENCES

[1] Y. Ze, Z. Chen, J. P. Araújo, Z. ang Cao, X. B. Peng, J. Wu, and C. K. Liu, "Twist: Teleoperated whole-body imitation system," *arXiv preprint arXiv:2505.02833*, 2025.

[2] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, "Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit," *arXiv preprint arXiv:2502.13013*, 2025.

[3] J. Li, X. Cheng, T. Huang, S. Yang, R. Qiu, and X. Wang, "Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control," *Robotics: Science and Systems 2025*, 2025.

[4] Y. Li, Y. Lin, J. Cui, T. Liu, W. Liang, Y. Zhu, and S. Huang, "Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks," *arXiv preprint arXiv:2506.08931*, 2025.

[5] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[6] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[7] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025, published: April 22, 2025.

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

[9] J. Aldaco, T. Armstrong, R. Baruch, J. Bingham, S. Chan, K. Draper, D. Dwibedi, C. Finn, P. Florence, S. Goodrich, W. Gramlich, T. Hage, A. Herzog, J. Hoech, T. Nguyen, I. Storz, B. Tabanpour, L. Takayama, J. Tompson, A. Wahid, T. Wahrburg, S. Xu, S. Yaroshenko, K. Zakka, and T. Z. Zhao, "Aloha 2: An enhanced low-cost hardware for bimanual teleoperation," *arXiv preprint arXiv:2405.02292*, 2024.

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

[11] C. Lu, X. Cheng, J. Li, S. Yang, M. Ji, C. Yuan, G. Yang, S. Yi, and X. Wang, "Mobile-television: Predictive motion priors for humanoid whole-body control," *arXiv preprint arXiv:2412.07773*, 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] PICO Immersive Pte. Ltd., "PICO 4 Ultra: An All-New Mixed Reality Experience," https://www.picoxr.com/global/products/pico4-ultra, 2023.

[14] 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.

[15] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," *arXiv preprint arXiv:2407.01512*, 2024.

[16] 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," *arXiv preprint arXiv:2410.10803*, 2024.

[17] H. Wang, Z. Wang, J. Ren, Q. Ben, T. Huang, W. Zhang, and J. Pang, "Beamdojo: Learning agile humanoid locomotion on sparse footholds," in *Robotics: Science and Systems (RSS)*, 2025.

[18] 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," *arXiv preprint arXiv:2505.03729*, 2025.

[19] J. Long, J. Ren, M. Shi, Z. Wang, T. Huang, P. Luo, and J. Pang, "Learning humanoid locomotion with perceptive internal model," 2024. [Online]. Available: https://arxiv.org/abs/2411.14386

[20] S. Chen, Y. Ye, Z.-A. Cao, J. Lew, P. Xu, and C. K. Liu, "Hand-eye autonomous delivery: Learning humanoid navigation, locomotion and reaching," *arXiv preprint arXiv:2508.03068*, 2025.

[21] Z. Luo, C. Tessler, T. Lin, Y. Yuan, T. He, W. Xiao, Y. Guo, G. Chechik, K. Kitani, L. Fan *et al.*, "Emergent active perception and dexterity of simulated humanoids from visual reinforcement learning," *arXiv preprint arXiv:2505.12278*, 2025.

[22] T. E. Truong, Q. Liao, X. Huang, G. Tevet, C. K. Liu, and K. Sreenath, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," *arXiv preprint arXiv:2508.08241*, 2025.

[23] H. Xiong, X. Xu, J. Wu, Y. Hou, J. Bohg, and S. Song, "Vision in action: Learning active perception from human demonstrations," *arXiv preprint arXiv:2506.15666*, 2025.

[24] H. Shi, W. Wang, S. Song, and C. K. Liu, "Toddlerbot: Open-source ml-compatible humanoid platform for loco-manipulation," *arXiv preprint arXiv:2502.00893*, 2025.

[25] PICO Immersive Pte. Ltd., "PICO Motion Tracker," https://www.picoxr.com/global/products/pico-motion-tracker, 2023.

[26] Z. Zhao, L. Yu, K. Jing, and N. Yang, "Xrobotoolkit: A cross-platform framework for robot teleoperation," *arXiv preprint arXiv:2508.00097*, 2025.

[27] "Vive tracker (3.0)," https://www.vive.com/us/accessory/tracker3/, 2025.

[28] B. Dynamics and T. R. Team, "Large behavior models and atlas find new footing," https://bostondynamics.com/blog/large-behavior-models-atlas-find-new-footing/, Aug. 2025, accessed: 2025-09-08.

[29] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025.

[30] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang, "Gmt: General motion tracking for humanoid whole-body control," *arXiv:2506.14770*, 2025.

[31] Y. Ze, J. P. Araújo, J. Wu, and C. K. Liu, "Gmr: General motion retargeting," https://github.com/YanjieZe/GMR, 2025, gitHub repository.

[32] 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 the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[33] J. Li, J. Wu, and C. K. Liu, "Object motion guided human motion synthesis," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–11, 2023.

[34] Redis contributors, "redis/redis: The redis in-memory data structure store," 2025, accessed: 2025-09-09. [Online]. Available: https://github.com/redis/redis

[35] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[36] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.

[37] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.