

OmniClone: Engineering a Robust, All-Rounder Whole-Body Humanoid Teleoperation System

Yixuan Li^{*,1,2,7} Le Ma^{*,2,7} Yutang Lin^{*,3,2,6,7,8,10} Yushi Du^{4,2,7} Mengya Liu^{2,7} Kaizhe Hu⁵
Jieming Cui^{3,2,6,7,8,10} Yixin Zhu^{✉,6,3,7,8,10} Wei Liang^{✉,1,9} Baoxiong Jia^{✉,2,7} Siyuan Huang^{✉,2,7}

* Equal contribution ✉ Corresponding author Project Website: <https://omniclone.github.io/>

¹ School of Computer Science and Technology, Beijing Institute of Technology ² Beijing Institute for General Artificial Intelligence (BIGAI)

³ Institute for AI, Peking University ⁴ Department of Electrical and Electronic Engineering, The University of Hong Kong

⁵ Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University ⁶ School of Psychological and Cognitive Sciences, Peking University

⁷ State Key Lab of General AI ⁸ Beijing Key Laboratory of Behavior and Mental Health, Peking University

⁹ Yangtze Delta Region Academy of Beijing Institute of Technology ¹⁰ Embodied Intelligence Lab, PKU-Wuhan Institute for Artificial Intelligence

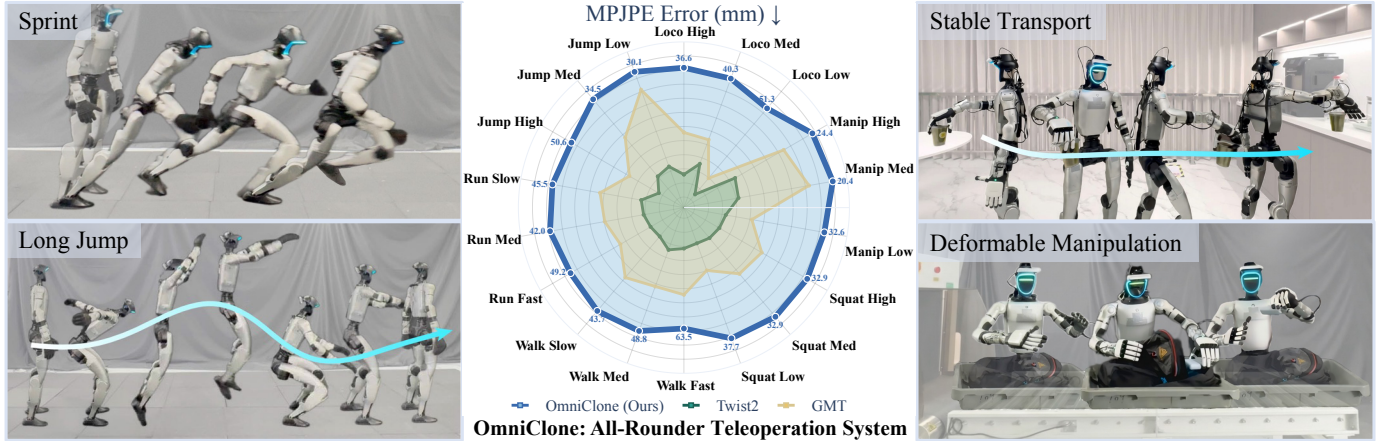


Fig. 1: **OmniClone** achieves well-balanced, high-fidelity whole-body tracking across all MPJPE dimensions on **OmniBench** while enabling versatile real-world deployment. Center: the radar map compares MPJPE (mm) across 18 stratified evaluation categories, showing that **OmniClone** consistently outperforms SOTA baselines GMT and Twist2. Surrounding panels illustrate real-world teleoperation with a unified policy, spanning dynamic whole-body motions (sprint, long jump) and stable, long-horizon dexterous loco-manipulation (transport, deformable object handling).

Abstract—Whole-body humanoid teleoperation enables humans to remotely control humanoid robots, serving as both a real-time operational tool and a scalable engine for collecting demonstrations for autonomous learning. Despite recent advances, existing systems are validated using aggregate metrics that conflate distinct motion regimes, masking critical failure modes. This lack of diagnostic granularity, compounded by tightly coupled and labor-intensive system configurations, hinders robust real-world deployment. A key open challenge is building a teleoperation system that is simultaneously robust, versatile, and affordable for practical use. Here we present **OmniClone**, a whole-body humanoid teleoperation system that achieves high-fidelity, multi-skill control on a single consumer GPU with modest data requirements. Central to our approach is **OmniBench**, a diagnostic benchmark that evaluates policies across stratified motion categories and difficulty levels on unseen motions, exposing the narrow specialization of prior systems. Guided by these diagnostics, we identify an optimized training data recipe and integrate system-level improvements—subject-agnostic re-targeting and robust communication—that collectively reduce Mean Per-Joint Position Error (MPJPE) by over 66% while requiring orders-of-magnitude fewer computational resources than comparable methods. Crucially, **OmniClone** is control-source-agnostic: a single unified policy supports real-time teleoperation, generated motion playback, and Vision-Language-Action (VLA) models, while generalizing across operators of vastly different

body proportions. By uniting diagnostic evaluation with practical engineering, **OmniClone** provides an accessible foundation for scalable humanoid teleoperation and autonomous learning.

I. INTRODUCTION

Humanoid robots are emerging as versatile, general-purpose platforms capable of executing complex tasks either autonomously [5, 13, 15, 18, 19, 22, 23, 39, 38] or via teleoperation [4, 12]. Whole-body teleoperation [2, 11, 17, 21, 40], in particular, serves as a foundational technology—enabling both real-time remote intervention and the collection of high-quality demonstrations for scalable imitation learning. Yet this broad operational scope demands the seamless integration of often conflicting capabilities: robustness, stability, dynamic agility, and precise control.

While reliable teleoperation frameworks exist for fixed-base manipulators [30, 32] and inherently stable quadrupeds [6, 8], achieving comparable maturity for whole-body humanoid teleoperation remains a formidable challenge. Recent methods [2, 3, 20, 21, 25, 41] have markedly expanded the capabilities of humanoid teleoperation, yet these systems still fall short of the robustness and task-agnostic versatility required for real-world deployment. We identify two root causes.

On the *evaluation* side, current methods tend to showcase isolated, high-complexity skills while reporting only coarse aggregate metrics that conflate distinct motion regimes. This masks critical nuances: a policy may maintain low tracking error during a standard upright pose yet exhibit significant instability during a deep squat. By averaging out such failure modes, existing evaluations obscure the narrow specialization of current models and the key bottlenecks that must be resolved for practical, generalizable deployment.

On the *system* side, the hardware and software configurations of current teleoperation systems are highly heterogeneous and tightly coupled with specific methods, hindering real-world reproducibility. Taking Motion Capture (MoCap) as an example, platforms like PICO VR headsets [29] and VICON systems [36] employ proprietary human-scale estimation algorithms that are often opaque to the user. The seemingly minor discrepancies in estimated scale lead to large performance gaps in practice. Compounded by factors such as control frequency and network latency, these systems necessitate labor-intensive calibration for every new operator and each distinct MoCap setting, severely impeding scalable deployment.

In this work, we adopt a system-level perspective to address these challenges, providing an effective and affordable solution. Rather than proposing yet another task or model in isolation, we begin by constructing **OmniBench**, the first comprehensive diagnostic evaluation benchmark for humanoid full-body teleoperation. **OmniBench** systematically assesses policy performance under varied workspaces and skill demands—from high-dynamic agile motions to stable, precise quasi-static manipulation—across multiple stratified difficulty levels. As shown in Fig. 1, this fine-grained evaluation uncovers persistent skill imbalances in existing systems, arising from both training data imbalance and model design. These limitations are further exacerbated in real-world deployment, where human-scale variabilities, retargeting errors, and data-stream disturbances caused by network fluctuation and latency jointly magnify performance gaps.

Guided by these diagnostics, we present **OmniClone**, a robust and affordable system for high-fidelity whole-body humanoid teleoperation across a comprehensive range of practical scenarios. **OmniClone** addresses the narrow coverage of existing models through a high-capacity transformer-based full-body tracking policy trained on a meticulously designed data recipe that balances diverse skills, from high agility to stable manipulation. Beyond policy learning, **OmniClone** incorporates system-level mechanisms to handle real-world disturbances, including subject-agnostic refined retargeting and robust data communication that mitigate network fluctuation and latency. Crucially, the entire system remains highly accessible, requiring only 30 hours of motion data and a single consumer GPU—orders of magnitude less than comparable methods. We validate **OmniClone** extensively on **OmniBench**, demonstrating significant improvements over state-of-the-art (SOTA) baselines with detailed ablations justifying each design choice. Moreover, **OmniClone** generalizes reliably across teleoperators ranging from 1.47 m to 1.94 m

in height, and serves as a control-source-agnostic platform compatible with real-time teleoperation, generated motion playback, and downstream VLA models. To demonstrate the latter, we train a VLA policy on **OmniClone**-collected data, achieving success rates of 85.71% and 80.00% on the ‘Pick-and-Place’ and ‘Squat to Pick-and-Place’ tasks, respectively.

Our contributions can be summarized as follows:

- We introduce **OmniBench**, the first comprehensive diagnostic benchmark for systematically evaluating humanoid full-body teleoperation across diverse motion categories and stratified difficulty levels, providing actionable insights for policy improvement.
- We propose **OmniClone**, a robust, affordable, and easily reproducible system for high-fidelity real-world humanoid teleoperation, integrating a skill-balanced policy with system-level mechanisms for handling deployment disturbances.
- We conduct extensive experiments demonstrating that **OmniClone** significantly outperforms SOTA baselines across all dimensions on **OmniBench** and in real-world testing, while further validating its utility as a data collection engine for learning autonomous VLA policies.

II. RELATED WORK

A. Dataset and Evaluation in Humanoid Teleoperation

While numerous recent works [19, 20, 21, 41, 37] develop custom datasets for humanoid teleoperation, rigorous analysis of how models perform across distinct functional capabilities remains largely absent. Standard evaluation protocols report overall Success Rate (SR) and tracking errors that conflate distinct movement patterns—from precise manipulation to dynamic locomotion—into single scalar values, obscuring critical performance discrepancies across dynamic regimes and masking specific failure modes. Consequently, most whole-body teleoperation efforts [3, 21, 41] evaluate performance exclusively on self-curated training distributions, where systems may exhibit superiority in isolated skills while suffering significant degradation in other skills. This practice further fails to capture true generalization, as evaluation and training distributions are identical.

To bridge this gap, we introduce **OmniBench**, a diagnostic evaluation suite that disentangles performance across multifaceted humanoid capabilities on unseen motions stratified by category and difficulty, thereby approximating real-world deployment conditions. **OmniBench** also provides actionable insights for dataset composition, enabling us to curate an optimized training recipe that maximizes tracking fidelity.

B. Whole-Body Humanoid Teleoperation

Whole-body teleoperation [7, 12] has established itself as a cornerstone for collecting dexterous demonstrations. Existing approaches can be broadly categorized into *decoupled* frameworks [2], which separate upper- and lower-body control for stability, and *unified* frameworks that learn a single full-body policy. Recent research increasingly favors the latter, as

coherent whole-body coordination is essential for natural and precise loco-manipulation.

However, the underactuated, floating-base dynamics of humanoid robots make high-fidelity unified control a multifaceted challenge spanning data curation [17, 25, 40], policy architecture [14, 17, 24, 42], and system-level infrastructure [41]. Despite extensive progress in each domain individually, a critical gap persists: existing systems remain tightly coupled to specific configurations and lack the robustness required for reliable real-world deployment across diverse operators and scenarios.

In this work, we address this gap from a system-engineering perspective, combining diagnostic benchmarking, an optimized data recipe, and a refined deployment infrastructure into a unified, affordable, and robust teleoperation framework.

III. THE OMNIBENCH BENCHMARK

We introduce **OmniBench**, a diagnostic evaluation suite designed to quantify the multifaceted capabilities of whole-body teleoperation policies. Below, we describe the benchmark construction and present baseline evaluations that expose the limitations of current systems.

A. Benchmark Construction

To ensure fair and rigorous evaluation on motions unseen during training, we leverage HY-Motion-1.0 [33] to generate a comprehensive set of reference motions, which are then retargeted to the humanoid morphology via General Motion Retargeting (GMR) [1]. The resulting evaluation set spans the full spectrum of humanoid capabilities, from basic locomotion to complex daily manipulation.

Specifically, we construct motions across six functional categories: ‘Loco-manipulation,’ ‘Manipulation,’ ‘Squatting,’ ‘Walking,’ ‘Running,’ and ‘Jumping,’ collectively covering the primary operational domains of teleoperated humanoid robots. Each category comprises 60 episodes with an average duration of 90 frames sampled at 30 Hz. To enable granular diagnosis, we further stratify these motions by *complexity* (High, Medium, Low) and *dynamic intensity* (Fast, Medium, Slow), yielding a fine-grained grid over both skill type and difficulty level. Further details on the evaluation dataset construction are provided in Sec. B-A.

B. Baseline Evaluation

We evaluate representative SOTA methods on **OmniBench**, with results shown in Fig. 1 and 2. The evaluation reveals a consistent pattern: existing whole-body teleoperation models remain narrowly specialized. While most systems achieve reasonable performance on manipulation tasks performed from a standard standing stance at medium-to-high heights, they exhibit significant degradation in other regimes—particularly tasks demanding lower-body agility, such as deep squatting or low-altitude loco-manipulation. These findings confirm that current aggregate evaluations mask critical skill imbalances, and provide empirical guidance for the policy and data improvements presented in the following sections.

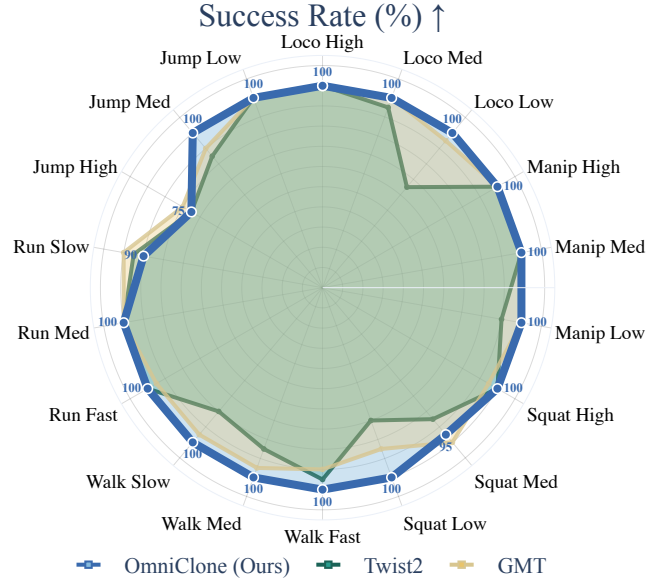


Fig. 2: **SR comparison on OmniBench reveals narrow skill specialization of existing methods.** While GMT and Twist2 suffer notable drops in lower-body agility tasks (deep squatting, low-altitude loco-manipulation) and high-dynamic maneuvers (jumping), **OmniClone** maintains near-perfect SR across all 18 stratified categories. Complementing the MPJPE radar map in Fig. 1, these results confirm that prior systems’ aggregate scores mask significant per-category failures.

IV. THE OMNICLEONE TELEOPERATION SYSTEM

We present **OmniClone**, a whole-body humanoid teleoperation system engineered for robust and versatile real-world performance. The framework comprises two integral components: (i) a transformer-based full-body control policy and (ii) an optimized whole-body teleoperation infrastructure. These components work synergistically to enable high-fidelity execution across a diverse spectrum of motions.

A. Transformer-Based Whole-Body Tracking Policy

As established in Sec. III, humanoid teleoperation demands versatile task execution characterized by diverse kinematic patterns. Commonly used Multi-layer Perceptrons (MLPs) architectures [11, 43] often struggle to capture the complex temporal dependencies inherent in such tasks. We therefore adopt a high-capacity Transformer architecture, whose ability to model long-range sequences is well-suited for processing histories of proprioceptive observations and actions. Its scalability further benefits learning from our large, high-variability training environments.

Building on the demonstrated success of Transformers in humanoid locomotion [10, 31], we implement both the teacher and student policies π using a Transformer backbone:

$$\pi : \mathbb{O} \rightarrow \mathbb{A}, \quad \mathbb{A} = \{\mathbf{a} \mid \mathbf{a} \in \mathbb{R}^n\}, \quad (1)$$

where $\mathbb{O}$ represents the observation space and $\mathbb{A}$ defines the action space in joint-space coordinates.

We train this policy using the data recipe distilled from our ablation study in Sec. V-B: 60% manipulation motions, with

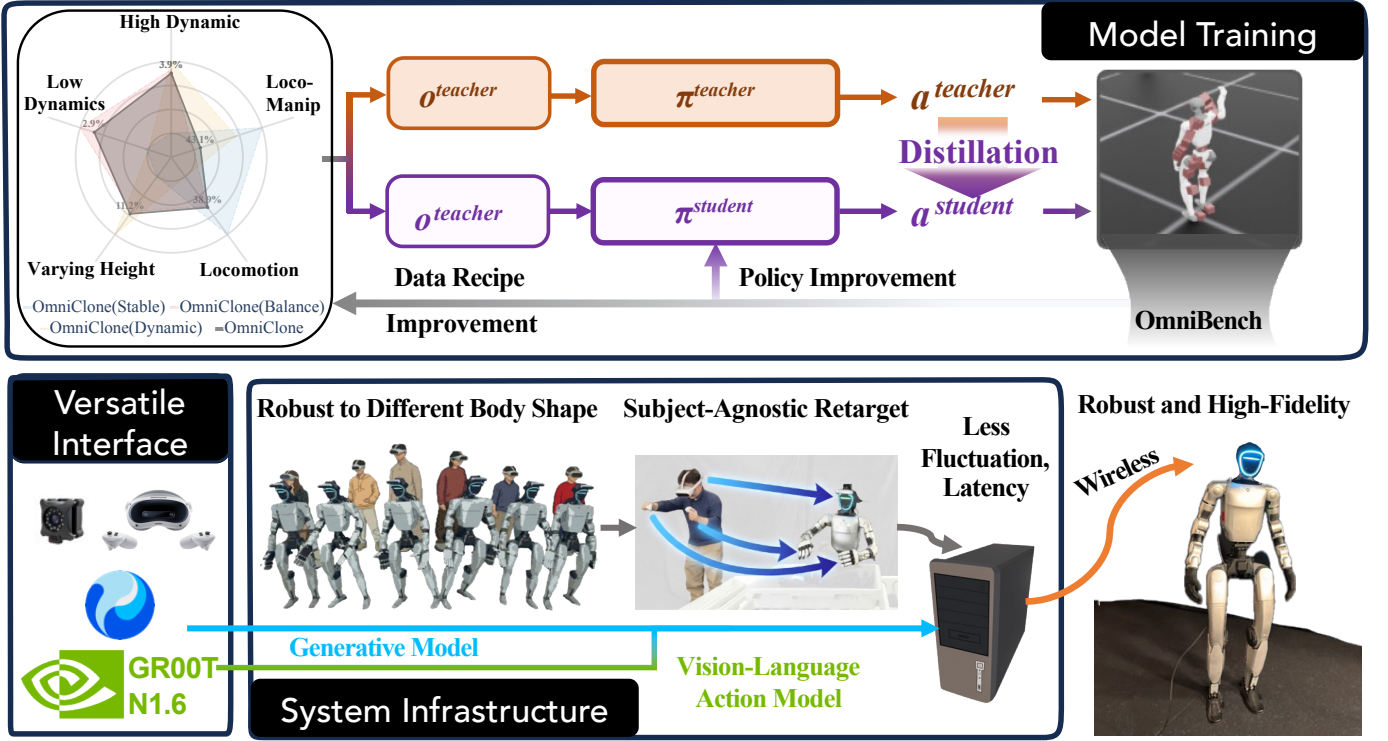


Fig. 3: **Overview of the OmniClone framework, comprising model training (top) and system infrastructure (bottom).** Top: a teacher-student distillation pipeline trained on an optimized data recipe, whose composition is iteratively refined via **OmniBench** diagnostics. Bottom: the deployment infrastructure features subject-agnostic retargeting and robust wireless communication, yielding a control-source-agnostic interface compatible with real-time teleoperation, generative motion models, and downstream VLA planners such as GR00T N1.6.

the remaining 40% balanced between dynamic maneuvers and stable locomotion.

Teacher policy: The teacher policy has access to a comprehensive set of joint state observations, including full body orientations and angular velocities, to facilitate high-fidelity behavior learning. The teacher observation vector $\mathbb{O}^{\text{teacher}}$ is defined as:

$$\mathbb{O}^{\text{teacher}} = [\mathbf{q}, \mathbf{p}, \mathbf{g}, \boldsymbol{\omega}^{\text{root}}, \mathbf{a}^{\text{last}}, \mathbf{q}^{\text{ref}}, \mathbf{p}^{\text{ref}}], \quad (2)$$

where $\mathbf{q}$ and $\mathbf{p}$ represent the robot’s full kinematic state, comprising joint states (positions and velocities) and body link states (positions, orientations, linear and angular velocities); $\mathbf{g}$ denotes the gravity vector; $\boldsymbol{\omega}^{\text{root}}$ is the angular velocity of the robot’s base (pelvis); and $\mathbf{a}^{\text{last}}$ is the previous action. $\mathbf{q}^{\text{ref}}$ and $\mathbf{p}^{\text{ref}}$ correspond to the reference motion’s joint states and body link kinematics. All observations are expressed in the robot’s local base frame to ensure spatial invariance.

Student policy: For robust real-world deployment, we exclude observations that are difficult to estimate reliably on physical hardware, such as the angular velocities of individual body links, and provide f frames of future reference motion. The student observation vector $\mathbb{O}^{\text{student}}$ is:

$$\mathbb{O}^{\text{student}} = [\mathbf{q}, \mathbf{g}, \boldsymbol{\omega}^{\text{root}}, \mathbf{a}^{\text{last}}, \mathbf{p}_{t:t+f}^{\text{ref}}], \quad (3)$$

where $\mathbf{p}^{\text{ref}}$ denotes the reference body state command, including the positions and orientations of key body links and the linear velocity of the robot base.

Network architecture and reward formulation: Details of the network architecture and reward formulation are provided in Sec. A-B.

B. Optimized Whole-Body Teleoperation Infrastructure

Beyond policy learning, robust real-world performance hinges on the deployment infrastructure itself. We identify two critical system-level factors that degrade teleoperation fidelity and address each with targeted mechanisms.

Retargeting discrepancies: Morphological inconsistencies introduce significant noise into the retargeting pipeline. These discrepancies arise from two sources: systematic variations across different MoCap hardware configurations and anthropometric variability among distinct human subjects using the same capture system (e.g., PICO [29]). We empirically quantify these discrepancies, observing maximum deviations of approximately 20 cm, which translate to an MPJPE increase of ≈ 20 mm when reproduced in simulation.

To address this, we propose a subject-agnostic refined retargeting pipeline. We compute a dynamic scaling factor from the initial calibration frame and the humanoid’s kinematics, then use it to rescale the raw MoCap data to match the humanoid’s morphology. This eliminates the need for per-operator manual calibration while effectively mitigating geometric discrepancies in the retargeting process.

Data fluctuation and latency: In real-world deployment, control signal fluctuations degrade motion smoothness, while

TABLE I: **Quantitative comparison with baselines and ablation on training data recipes, evaluated on OmniBench.** All results are stratified by motion category and difficulty level, reporting SR (%) and MPJPE (mm). **OmniClone** achieves the best overall balance across all categories, while the ablation variants illustrate how data composition steers policy behavior. Best MPJPE per category is in bold.

(a) Locomotion and manipulation tasks, stratified by workspace height (High, Medium, Low).

Method	Loco-Manip						Manip						Squat					
	High		Medium		Low		High		Medium		Low		High		Medium		Low	
	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓
GMT [3]	100	128.4	100	132.5	95	180.5	100	71.4	100	54.7	100	137.0	95	107.4	100	112.4	85	140.1
Twist2 [41]	100	188.0	95	168.9	65	210.5	100	151.1	100	156.3	90	174.6	100	176.9	85	178.5	70	181.1
OmniClone ^{MLP}	85	58.90	65	75.4	65	73.4	100	36.3	100	37.1	100	47.9	100	49.5	100	47.3	95	54.7
OmniClone	100	36.6	100	40.3	100	51.3	100	24.4	100	20.4	100	32.6	100	32.9	95	32.9	100	37.7
OmniClone ^{Balance}	100	35.3	100	39.9	100	48.1	100	19.0	100	18.4	100	33.3	100	32.0	100	31.4	90	37.1
OmniClone ^{Stable}	100	37.2	95	43.7	100	50.2	100	22.2	100	20.8	100	38.4	100	39.7	100	38.1	95	45.2
OmniClone ^{Dynamic}	30	78.1	20	96.8	10	92.8	75	53.9	75	53.7	25	75.4	35	72.9	45	65	35	71.6

(b) Agile motion tasks, stratified by dynamic intensity (Fast, Medium, Slow) for Walk and Run, and by height for Jump.

Method	Walk						Run						Jump					
	Fast		Medium		Slow		Fast		Medium		Slow		High		Medium		Low	
	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓	SR ↑	MPJPE ↓
GMT [3]	90	111.4	95	116.0	95	105.5	95	130.9	100	120.8	100	114.1	80	145.6	90	105.3	100	57.1
Twist2 [41]	95	176.9	85	171.2	80	181.6	100	180.4	100	176.9	95	173.4	75	193.8	85	177.2	100	173.0
OmniClone ^{MLP}	20	111.7	35	95.6	85	64.1	90	82.0	80	69.8	70	71.1	75	81.4	85	62.9	75	58.9
OmniClone	100	63.5	100	48.8	100	43.7	100	49.2	100	42.0	90	45.5	75	50.6	100	34.5	100	30.1
OmniClone ^{Balance}	100	63.0	95	46.4	95	40.8	90	45.7	100	38.3	100	41.6	90	50.0	100	35.3	100	29.8
OmniClone ^{Stable}	100	70.3	100	49.3	100	43.8	100	51.4	100	42.4	95	43.6	80	57.1	100	44.5	100	39.6
OmniClone ^{Dynamic}	90	64.8	100	48.6	100	44.1	100	47.7	100	38.3	95	41.8	85	48.8	10	74.9	45	63.9

network latency increases the cognitive load on the teleoperator. We address both challenges with a queue-based data manager that buffers f frames of retargeted motion in a FIFO queue at a fixed update frequency. The manager employs a zero-order hold strategy, retaining the previous observation during signal interruptions to ensure smooth execution. We further adopt UDP for communication between the retargeting server and policy server (supporting both remote and onboard configurations) for its superior transmission efficiency. Together, these mechanisms yield a low end-to-end latency of approximately 80 ms.

C. Implementation Details

We train our policy in Isaac Lab [27] on a single NVIDIA RTX 4090 GPU. The teacher policy converges in approximately 60 hours, spanning $\sim 33K$ Proximal policy optimization (PPO) steps with 4,096 parallel environments. The student policy is then distilled from the teacher via DAgger in approximately 22 hours under the same configuration. In total, the system requires only ~ 80 GPU hours on consumer hardware—orders of magnitude less than comparable methods—underscoring the affordability of **OmniClone**.

V. SIMULATIONS

We conduct comprehensive simulations in MuJoCo [34] to evaluate the motion tracking capabilities of **OmniClone** on the **OmniBench** suite. We report both SR and MPJPE (in mm) across a diverse spectrum of skills at varying difficulty levels. The primary objective is to quantitatively analyze how distinct training data recipes dictate policy performance, thereby distilling actionable insights for policy learning.

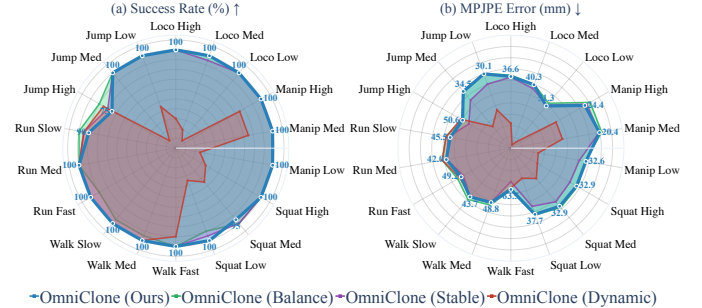


Fig. 4: **Ablations on training data recipes show how data composition steers policy behavior.** **OmniClone**^{Dynamic} excels at agile motions but collapses on manipulation and squatting; **OmniClone**^{Stable} recovers stability at the cost of dynamic agility; **OmniClone**^{Balance} improves breadth across both regimes. The final **OmniClone** recipe, obtained by filtering biased sequences from **OmniClone**^{Balance}, achieves the best overall trade-off in both (a) SR and (b) MPJPE.

A. Training Data Recipe

We curate our training dataset by aggregating motion data from the AMASS [26] and LAFAN [9] datasets. The composition of the final dataset is illustrated in Fig. 3. Since the majority of the data involves manipulation in a standing pose ($\approx 60\%$), the figure specifically highlights the distribution of distinct dynamic behaviors to better visualize the dataset’s diversity. A detailed breakdown of the data composition is provided in Sec. B-A.

B. Reference Motion Tracking

Baselines and ablation variants: We compare **OmniClone** against two SOTA full-body tracking baselines

using their pre-trained checkpoints: GMT [3] and Twist2 [41]. To validate our data recipe design, we further evaluate several ablation variants whose data compositions are visualized in Fig. 4:

- **OmniClone^{Dynamic}**: trained exclusively on high-dynamic maneuvers.
- **OmniClone^{Stable}**: augmented with stable motions such as standard locomotion and loco-manipulation.
- **OmniClone^{Balance}**: trained on a comprehensive mixture incorporating additional dynamic motion sources.
- **OmniClone**: our final model, trained on the optimized data mixture from **OmniClone^{Balance}** after rigorous noise filtering.
- **OmniClone^{MLP}**: an architectural ablation replacing the Transformer backbone with a standard MLPs policy.

Comparison with baselines: We evaluate motion tracking performance across diverse simulation tasks, with results presented in Tabs. Ia and Ib and Fig. 1 and 2. The quantitative metrics in Tabs. Ia and Ib demonstrate that **OmniClone** achieves superior tracking performance across most tasks. The radar maps in Fig. 1 and 2 further visualize this comparison in terms of SR and MPJPE, showing that **OmniClone** attains a well-balanced performance profile across multi-faceted tasks while substantially improving overall tracking fidelity relative to baselines.

Ablation on data recipe: We analyze the impact of training data composition on policy behavior, with the distinct recipes detailed in Fig. 3 and corresponding performance comparisons in Fig. 4.

OmniClone^{Dynamic}, trained on a dataset heavily skewed towards high-dynamic motions and varying heights, can execute explosive maneuvers such as jumping but lacks stable locomotion data. This coverage gap prevents the model from

learning robust stabilization strategies, resulting in significant instability during precise manipulation tasks. In contrast, **OmniClone^{Stable}** incorporates a larger proportion of locomotion and loco-manipulation data, yielding improved performance on stable-oriented tasks like squatting and loco-manipulation. By further increasing the proportion of dynamic data, **OmniClone^{Balance}** achieves more balanced performance across both dynamic and stable regimes, as shown in Fig. 4. However, we observed that certain raw dynamic sequences introduce tracking bias during real-world deployment. Filtering these from the **OmniClone^{Balance}** dataset yields the final, optimized recipe for **OmniClone**.

A key insight from this analysis is that, although standing manipulation constitutes the majority of the training data, the system’s behavioral characteristics are highly sensitive to the composition of the remaining fraction. Subtle adjustments to the proportions of dynamic and locomotion data, combined with filtering of biased sequences, can effectively steer the policy towards distinct behaviors and capabilities.

Ablation on model architecture: We quantify the performance differential between the MLPs and Transformer backbones. As shown in Tab. I, **OmniClone** consistently outperforms **OmniClone^{MLP}** across all metrics, achieving higher SR and lower MPJPE under identical training settings. We attribute this gain to the Transformer’s ability to model long-range temporal dependencies and its enhanced scalability when learning from massive, diverse skill datasets. Comparisons with baseline models further demonstrate that, even when controlling for the model backbone, our data recipe refined using **OmniBench** significantly enhances policy learning efficacy.

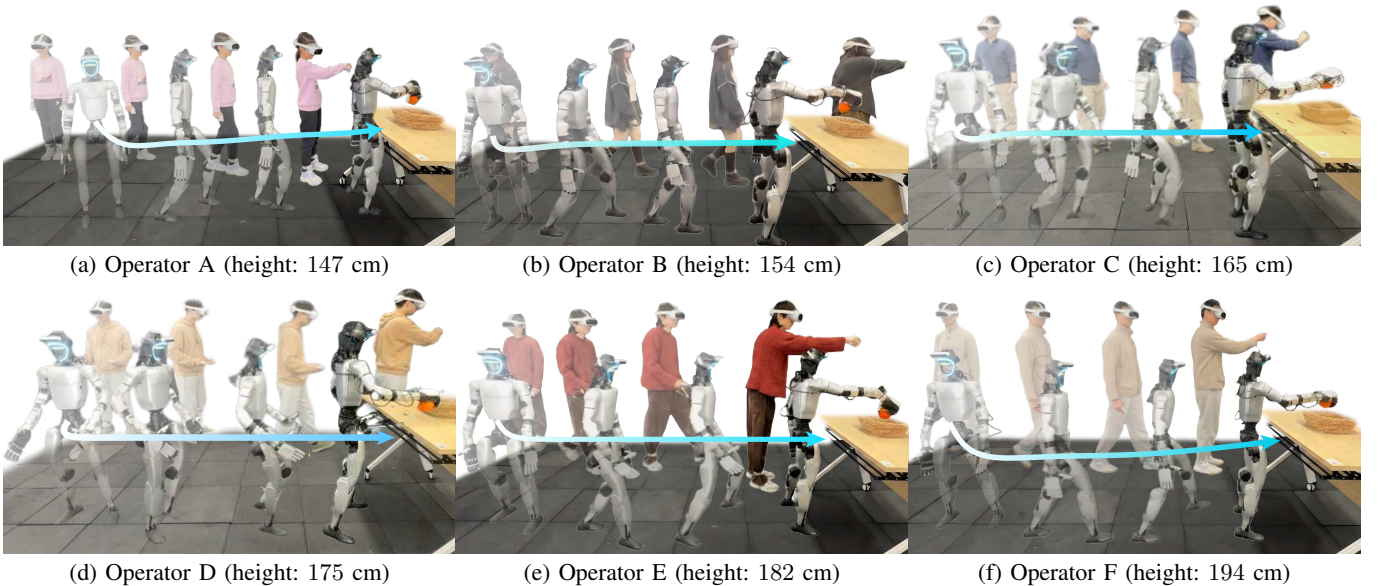


Fig. 5: **OmniClone generalizes reliably across teleoperators of vastly different heights.** Six participants (1.47 m–1.94 m) each perform a composite loco-manipulation task (walk, stabilize, pick-and-place). Despite a 47 cm height span, the system maintains consistent stability and task success throughout, with all novice operators completing the task within 5–7 practice trials.

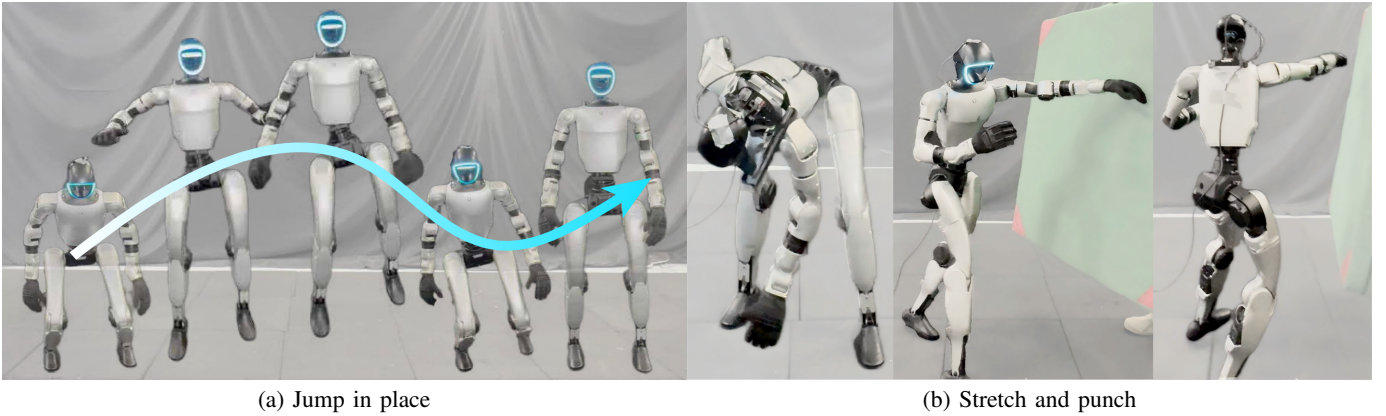


Fig. 6: **Zero-shot playback of generated motions establishes the performance upper bound of the control policy.** By replaying synthetic motions from HY-Motion-1.0 in the absence of system-induced noise (retargeting artifacts, network latency), we isolate the policy’s intrinsic tracking fidelity. Together with the sprint and long-jump results in the left panels of Fig. 1, these results confirm that the end-to-end system’s residual errors are primarily driven by upstream infrastructure bottlenecks rather than policy capacity.



(a) Long-horizon composite task: open cabinet, retrieve tissue, and place on table.



(b) Serving tea without spillage.

(c) Dynamic running.

Fig. 7: **A single unified policy handles the full spectrum of real-world teleoperation, from precise quasi-static manipulation to highly dynamic locomotion.** (a) The humanoid is teleoperated to open a cabinet, retrieve a tissue, and place it on a table, exercising articulated object interaction and coordinated whole-body motion. (b) Stable liquid transport with fine-grained force control is demonstrated by serving tea without spillage. (c) High-speed running further validates the policy’s dynamic agility. All tasks are executed via real-time teleoperation without policy switching.

VI. EXPERIMENTS

We conduct real-world experiments to validate **OmniClone** along two axes: (i) robustness and reliability across diverse operators and motion profiles, and (ii) data utility for training downstream autonomous policies.

A. Robustness and Reliability

We evaluate robustness through two complementary protocols: real-time teleoperation with diverse human operators, and zero-shot playback of generated motions unseen during training.

Real-time teleoperation: We recruit 6 participants of varying gender and height (1.47 m to 1.94 m), the majority of whom are novices with no prior teleoperation experience. Each participant performs a composite loco-manipulation task that simultaneously tests locomotion stability and manipulation precision: the operator teleoperates the humanoid to walk toward a target table, establish a stable stance, and execute a ‘pick-and-place’ maneuver.

As shown in Fig. 5, the system maintains consistent stability throughout the task despite substantial anthropometric differences among operators. All novice users achieved successful

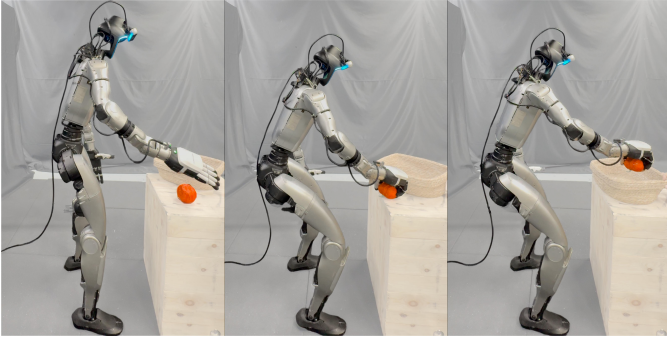


Fig. 8: **Enabling Autonomy via GR00T N1.6 [28] Post-Training.** We utilize high-fidelity data collected by our system to post-train the GR00T N1.6 [28] foundation model, demonstrating superior full-body control capability.

completion within only 5–7 practice trials, indicating that **OmniClone** generalizes effectively across diverse body proportions while offering an intuitive, low-barrier interface. We refer the reader to the supplementary video for comprehensive visualizations.

Zero-shot generalized motion playback: Since our full-body control policy is trained exclusively on standard MoCap data, evaluating it on unseen synthetic motions provides a rigorous test of generalization. We use HY-Motion-1.0 [33] to generate a diverse suite of test motions spanning stationary manipulation, loco-manipulation, and highly dynamic locomotion.

As depicted in the left panels of Fig. 1 and in Fig. 6, **OmniClone** tracks these unseen motion profiles with consistent fidelity, demonstrating robust zero-shot transfer to novel motion sources. This confirms that the system is control-source-agnostic: the same unified policy executes reliably regardless of whether the reference motion originates from a real-time teleoperator or a generative model.

Qualitative results: We further visualize real-time teleoperation across a broad range of behaviors on the right side of Fig. 1 and in Fig. 7. The system robustly executes complex, long-horizon maneuvers—‘Transporting Coffee Beans’ (Fig. 1), ‘Retrieving Tissue from a Low Cabinet’ (Fig. 7(a)), and ‘Serving Tea’ without spillage (Fig. 7(b))—spanning loco-manipulation at varying heights, articulated and deformable object handling, and precise liquid transport. As shown in Fig. 7(c), **OmniClone** also executes highly dynamic motions such as running. Crucially, this full spectrum of behaviors, from precise quasi-static manipulation to dynamic locomotion, is achieved with a single unified policy.

B. Data Utility for Autonomy

To validate the quality of demonstrations collected via **OmniClone**, we post-train GR00T N1.6 [28], a SOTA VLA foundation model, on 50 full-body demonstrations and evaluate it on two manipulation tasks: ‘Pick-and-Place’ (lifting a tomato and placing it into a basket; Fig. A.1) and ‘Squat to Pick-and-Place’ (performing the same task from a squatting pose; Fig. 8). Notably, **OmniClone** serves as both the

TABLE II: **Autonomous policy trained on OmniClone-collected demonstrations achieves high SR with only 50 demonstrations.** We post-train GR00T N1.6 on full-body data collected via **OmniClone** and evaluate on two whole-body manipulation tasks. The results validate that **OmniClone** produces kinetically feasible, high-fidelity demonstrations suitable for effective downstream policy learning in low-data regimes.

Task	Trials	SR
Pick-and-Place	21	85.71%
Squat to Pick-and-Place	5	80.00%

data collection interface and the low-level controller during autonomous inference, imposing stringent demands on the fidelity of the control policy and the stability of the system infrastructure. Implementation details are provided in Sec. C-B.

As reported in Tab. II, the autonomous policy achieves SR of 85.71% and 80.00% on the two tasks respectively, generalizing to minor variations in object position. These results confirm that **OmniClone** produces clean, consistent, and kinetically feasible demonstrations, enabling effective downstream policy learning even in low-data regimes.

VII. CONCLUSION

We have presented **OmniBench** and **OmniClone**, addressing the evaluation and system gaps that hinder practical whole-body humanoid teleoperation. **OmniBench** provides the first diagnostic benchmark that disentangles policy performance across stratified motion categories and difficulty levels on unseen motions, replacing coarse aggregate metrics with fine-grained analysis that exposes the narrow specialization of existing systems. Guided by these diagnostics, we identified an optimized training data recipe and demonstrated that the composition of even a small fraction of the training data can decisively steer policy behavior—a finding with broad implications for data curation in humanoid control.

Building on these insights, **OmniClone** integrates a transformer-based full-body tracking policy with system-level mechanisms—subject-agnostic retargeting and robust communication—into a unified framework that achieves high-fidelity teleoperation on a single consumer GPU with approximately 30 hours of motion data. The system generalizes across teleoperators of vastly different heights, operates as a control-source-agnostic platform compatible with real-time teleoperation, generated motion playback, and downstream VLA models, and produces kinetically feasible demonstrations that enable effective autonomous policy learning in low-data regimes.

Looking ahead, several promising directions remain. Scaling the data recipe to close the remaining gap in extreme dynamic regimes (e.g., high jumps) while preserving overall balance is a natural next step. Extending **OmniBench** with standardized real-world task protocols would further strengthen its diagnostic utility. Finally, the control-source-agnostic nature of **OmniClone** opens the possibility of closing the loop between generative motion models and long-horizon autonomous execution in the real world.

ACKNOWLEDGMENTS

This work is supported in part by the National Key Research and Development Program of China (2025YFE0218200), the National Natural Science Foundation of China (62172043 to W.L., 62376009 to Y.Z.), the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

REFERENCES

- [1] Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [2] Qingwei Ben, Feiyu Jia, Jia Zeng, Juntong Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. In *Robotics: Science and Systems (RSS)*, 2025.
- [3] Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. Gmt: General motion tracking for humanoid whole-body control. *arXiv preprint arXiv:2506.14770*, 2025.
- [4] Kourosh Darvish, Luigi Penco, Joao Ramos, Rafael Cisneros, Jerry Pratt, Eiichi Yoshida, Serena Ivaldi, and Daniele Pucci. Teleoperation of humanoid robots: A survey. *IEEE Transactions on Robotics (T-RO)*, 39(3): 1706–1727, 2023.
- [5] Yushi Du, Yixuan Li, Baoxiong Jia, Yutang Lin, Pei Zhou, Wei Liang, Yanchao Yang, and Siyuan Huang. Learning human-humanoid coordination for collaborative object carrying. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [6] Zipeng Fu, Xuxin Cheng, and Deepak Pathak. Deep whole-body control: learning a unified policy for manipulation and locomotion. In *Conference on Robot Learning (CoRL)*, pages 138–149, 2023.
- [7] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning (CoRL)*, 2024.
- [8] Seunghyeon Ha, Seongyong Kim, and Soo-Chul Lim. Prediction of delay-free scene for quadruped robot teleoperation: Integrating delayed data with user commands. *IEEE Robotics and Automation Letters (RA-L)*, 2025.
- [9] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2020.
- [10] Junzhe He, Chong Zhang, Fabian Jenelten, Ruben Grandia, Moritz Bäcker, and Marco Hutter. Attention-based map encoding for learning generalized legged locomotion. *Science Robotics*, 10(105):eadv3604, 2025.
- [11] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning (CoRL)*, 2024.
- [12] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.
- [13] Tairan He, Zi Wang, Haoru Xue, Qingwei Ben, Zhengyi Luo, Wenli Xiao, Ye Yuan, Xingye Da, Fernando Castañeda, Shankar Sastry, et al. Viral: Visual sim-to-real at scale for humanoid loco-manipulation. *arXiv preprint arXiv:2511.15200*, 2025.
- [14] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- [15] Kaizhe Hu, Haochen Shi, Yao He, Weizhuo Wang, C Karen Liu, and Shuran Song. Robot trains robot: Automatic real-world policy adaptation and learning for humanoids. In *Conference on Robot Learning (CoRL)*, 2025.
- [16] Intel Corporation. Intel® RealSense™ D455. <https://www.intel.com/content/www/us/en/products/sku/205847/intel-realsense-depth-camera-d455/specifications.html>, 2020.
- [17] Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [18] Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, et al. Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2026.
- [19] Nan Jiang, Zimo He, Wanhe Yu, Lexi Pang, Yunhao Li, Hongjie Li, Jieming Cui, Yuhao Li, Yizhou Wang, Yixin Zhu, et al. Uniact: Unified motion generation and action streaming for humanoid robots. *arXiv preprint arXiv:2512.24321*, 2025.
- [20] Jialong Li, Xuxin Cheng, Tianshu Huang, Shiqi Yang, Ri-Zhao Qiu, and Xiaolong Wang. Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control. In *Robotics: Science and Systems (RSS)*, 2025.
- [21] Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In *Conference on Robot Learning (CoRL)*, 2025.
- [22] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yu-

- man Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.
- [23] Yutang Lin, Jieming Cui, Yixuan Li, Baoxiong Jia, Yixin Zhu, and Siyuan Huang. Lessmimic: Long-horizon humanoid interaction with unified distance field representations. *arXiv preprint arXiv:2602.21723*, 2026.
- [24] Chenhao Lu, Xuxin Cheng, Jialong Li, Shiqi Yang, Mazeyu Ji, Chengjing Yuan, Ge Yang, Sha Yi, and Xiaolong Wang. Mobile-television: Predictive motion priors for humanoid whole-body control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- [25] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.
- [26] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [27] Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.
- [28] NVIDIA Corporation. Project GR00T: A Generalist Foundation Model for Humanoid Robots. <https://developer.nvidia.com/project-gr00t>, 2024.
- [29] PICO. PICO 4 Ultra. <https://www.picoxr.com/cn/products/pico4-ultra>, 2025.
- [30] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems (RSS)*, 2023.
- [31] Ilija Radosavovic, Tete Xiao, Bike Zhang, Trevor Darrell, Jitendra Malik, and Koushil Sreenath. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 9(89):eadi9579, 2024.
- [32] Zilin Si, Kevin Lee Zhang, Zeynep Temel, and Oliver Kroemer. Tilde: Teleoperation for dexterous in-hand manipulation learning with a deltahand. In *Robotics: Science and Systems (RSS)*, 2024.
- [33] Tencent Hunyuan 3D Digital Human Team. Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.
- [34] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.
- [35] Unitree Robotics. Unitree G1-Comp. <https://www.unitree.com/cn/robocup>, 2025.
- [36] Vicon Motion Systems Ltd. Vicon Motion Capture Systems. <https://www.vicon.com/>, 2026.
- [37] Huayi Wang, Wentao Zhang, Runyi Yu, Tao Huang, Junli Ren, Feiyu Jia, Zirui Wang, Xiaojie Niu, Xiao Chen, Jiahe Chen, et al. Physhsi: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025.
- [38] Yinhuai Wang, Qihan Zhao, Yuen Fui Lau, Runyi Yu, Hok Wai Tsui, Qifeng Chen, Jingbo Wang, Jiangmiao Pang, and Ping Tan. Humanx: Toward agile and generalizable humanoid interaction skills from human videos. *arXiv preprint arXiv:2602.02473*, 2026.
- [39] Haoru Xue, Tairan He, Zi Wang, Qingwei Ben, Wenli Xiao, Zhengyi Luo, Xingye Da, Fernando Castañeda, Guanya Shi, Shankar Sastry, et al. Opening the sim-to-real door for humanoid pixel-to-action policy transfer. *arXiv preprint arXiv:2512.01061*, 2025.
- [40] Yanjie Ze, Zixuan Chen, Joao Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C Karen Liu. Twist: Teleoperated whole-body imitation system. In *Conference on Robot Learning (CoRL)*, 2025.
- [41] Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C Karen Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [42] Zhikai Zhang, Chao Chen, Han Xue, Jilong Wang, Sikai Liang, Yun Liu, Zongzhang Zhang, He Wang, and Li Yi. Unleashing humanoid reaching potential via real-world-ready skill space. *IEEE Robotics and Automation Letters (RA-L)*, 11(2):2082–2089, 2026.
- [43] Zhikai Zhang, Jun Guo, Chao Chen, Jilong Wang, Chenghuai Lin, Yunrui Lian, Han Xue, Zhenrong Wang, Maoqi Liu, Jiangran Lyu, et al. Track any motions under any disturbances. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

APPENDIX A WHOLE-BODY TRACKING POLICY

A. Problem Formulation

We formulate whole-body humanoid teleoperation as a motion tracking problem, modeled as a Markov Decision Process (MDP) defined by the tuple $\mathbb{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$, where $\mathcal{S}$ encompasses the full system state including proprioceptive and task-oriented variables, $\mathcal{A} \subset \mathbb{R}^{29}$ represents the target joint angles for the humanoid’s actuators, $\mathcal{T}$ denotes the transition dynamics, and $\mathcal{R}$ is the reward function designed to encourage tracking fidelity and physical plausibility. We optimize a policy π using the PPO algorithm to maximize the expected cumulative reward.

Given a reference target trajectory, the policy minimizes the deviation between the robot’s current state and the reference while maintaining balance and stability:

$$\pi : \mathcal{O} \rightarrow \mathcal{A}, \quad \mathcal{A} = \{\mathbf{a} \mid \mathbf{a} \in \mathbb{R}^{29}\}, \quad (\text{A.4})$$

where $\mathcal{O}$ represents the observation space of the policy.

B. Model Structure

We employ a Transformer-based Actor-Critic architecture comprising three functional components: a linear input projection, a Transformer backbone, and MLPs output heads.

The teacher model uses a lightweight Transformer backbone with embedding dimension $d_{\text{model}} = 256$ and Feed-Forward Network (FFN) dimension $d_{\text{ff}} = 512$, processing observations from 4 input tokens. The student model is designed with higher representational capacity: the embedding dimension is doubled to $d_{\text{model}} = 512$ with a corresponding FFN dimension of $d_{\text{ff}} = 1024$, processing observations via 2 input tokens. Both models employ multi-head attention with 4 heads.

C. Reward Function and Domain Randomization

The reward function components and their respective weights are summarized in Tab. A.2. The domain randomization parameters applied during training are detailed in Tab. A.3.

APPENDIX B SIMULATIONS

A. Dataset

Training dataset: Our training dataset is partially derived from the TWIST2 [41] corpus, where we expand the retargeted motions from 23 to 29 Degree-of-Freedom (DoF) to match our robot’s configuration. To align the training data with the insights derived from **OmniBench**, we apply heuristic filtering criteria based on root state (position and velocity) and whole-body joint energy to selectively curate the reference motions. Following the data collection protocol established in TWIST2, we further augment this base dataset with additional reference motions captured directly via our teleoperation interface. The final dataset is predominantly composed of standing manipulation tasks (approximately 60%), totaling 30 hours of motion data. The distribution of the remaining motion categories is detailed in the data recipe presented in Fig. 3.

TABLE A.1: **Statistical overview of the generated evaluation corpus.** For locomotion tasks, we report the mean velocity alongside the 5th and 95th percentiles (P5/P95) to exclude transient phases such as startup and stopping transitions. For manipulation tasks, where metrics focus on root or hand height, we report the mean, min, and max values to illustrate the full extent of the reachable workspace. (**Med**: Medium; **Loco**: Loco-manipulation; **Manip**: Manipulation.)

(a) **Locomotion Speed (m/s)**

Motion	Min P5	Max P95	Mean
Walk Slow	0.618	1.704	1.026
Walk Med	0.793	2.193	1.364
Walk Fast	0.665	2.279	1.584
Run Slow	0.667	2.883	1.932
Run Med	0.644	4.151	2.386
Run Fast	0.685	5.046	2.956

(b) **Root Height (m)**

Motion	Min	Max	Mean
Jump Low	0.522	1.206	0.783
Jump Med	0.466	1.482	0.783
Jump High	0.351	1.638	0.783
Squat Low	0.260	0.954	0.610
Squat Med	0.338	0.831	0.627
Squat High	0.357	0.818	0.690

(c) **Hand Height (m)**

Motion	Min	Max	Mean
Loco Low	0.209	1.040	0.664
Loco Med	0.628	1.112	0.796
Loco High	0.626	1.470	0.929
Manip Low	0.162	1.007	0.698
Manip Med	0.661	1.103	0.849
Manip High	0.633	1.387	0.990

Evaluation dataset: To generate reference motions that comprehensively reflect the model’s capabilities on unseen distributions, we utilize HY-Motion-1 [33] to synthesize the evaluation dataset forming the core of **OmniBench**. Detailed specifications are provided in Tab. A.1. Crucially, this dataset encompasses a diverse array of tasks representative of daily activities, none of which are present in the training distribution. **OmniBench** therefore serves as a rigorous testbed for evaluating tracking precision and robustness across multifaceted capabilities.

APPENDIX C REAL-WORLD EXPERIMENTS

A. System Setup

Hardware setup: We evaluate **OmniClone** in real-world scenarios using the Unitree G1-Comp platform [35]. The robot is equipped with a 2-DoF actuated head featuring an Intel RealSense D455 depth camera [16], supporting a range of motion of $\pm 50^\circ$ in yaw and $[-90^\circ, +22.7^\circ]$ in pitch to ensure wide visual coverage for manipulation tasks.

Deployment architecture: The full-body control policy executes locally on the robot’s onboard computer (PC2). Motion data captured from the teleoperator—using either VR headsets such as PICO [29] or optical MoCap systems—is processed and retargeted on an external PC before being

TABLE A.2: **Reward function configuration.** Summary of reward components and their respective weights used for training. (**Pos.**: Position; **Rot.**: Rotation; **Lin.**: Linear; **Ang.**: Angular; **Vel.**: Velocity. The end-effectors consist of the ‘wrist yaw link’ and ‘ankle roll link.’)

Category	Reward Term	Weight
Regularization	Action Rate Penalty	−8.0
	Contact Air Time Penalty	−100.0
	Joint Acceleration Penalty	-1.0×10^{-7}
	Joint Position Limits	−10.0
	Velocity/Action Limits	−1.0
Tracking	Torso Global Pos. / Rot.	0.5
	Full-Body Global Lin. / Ang. Vel.	1.0
	Full-Body Relative Pos. / Rot.	1.0
	End-effector Relative Pos. / Rot. / Lin. / Ang. Vel.	0.5

TABLE A.3: **Domain randomization.** Specifications of the primary randomization parameters and their respective ranges applied during training. (**CoM**: Center of Mass; **RFI**: Random Force Injection; **nominal**: default values defined in the URDF. $\mathcal{U}(a, b)$ denotes a uniform distribution; $\delta_{x,y,z}$ for the CoM uses specific ranges per axis.)

Parameter	Sampling Distribution	Details
Action Delay (s)	$\mathcal{U}(0.0, 0.02)$	—
Action Noise (rad)	$\mathcal{U}(0.0, 0.02)$	—
Link Mass	$[0.9, 1.1] \times \text{nominal}$	‘torso,’ ‘shoulder yaw’
Torso CoM Offset (m)	$\delta_x \sim \mathcal{U}(-0.075, 0.075), \delta_{y,z} \sim \mathcal{U}(-0.1, 0.1)$	—
Torque RFI	$0.02 \times \text{limit (N·m)}$	—
Static Friction	$\mathcal{U}(0.3, 2.0)$	‘ankle roll,’ ‘pelvis,’ ‘hip roll,’ ‘knee,’ ‘elbow’
Dynamic Friction	$\mathcal{U}(0.3, 2.0)$	‘ankle roll,’ ‘pelvis,’ ‘hip roll,’ ‘knee,’ ‘elbow’
Stiffness Scale	$[0.95, 1.05] \times \text{nominal}$	—
Damping Scale	$[0.95, 1.05] \times \text{nominal}$	—
Armature Range	$[0.995, 1.015] \times \text{nominal}$	—

transmitted wirelessly to the robot via a router.

B. Data Utility for Autonomy

We detail the autonomy task specifications, data collection pipeline, and inference protocols used to validate the fidelity of full-body demonstrations acquired via **OmniClone**. The complete pipeline for autonomous data collection and task execution will be publicly released to facilitate reproducibility.

Autonomy task specifications: We utilize GR00T N1.6 [28], a widely adopted SOTA VLA foundation model that maps multimodal observations (natural language instructions and robot states) to executable actions. This setup provides a testbed for evaluating three key aspects: (i) the fidelity of demonstrations collected via our teleoperation pipeline, (ii) the reproducibility of our full-body control policy when functioning as the low-level controller for the VLA, and (iii) the robustness of the overall system infrastructure in executing the VLA’s high-level commands.

Data collection pipeline: We provide a whole-body data collection pipeline designed for seamless compatibility with GR00T N1.6 [28], recording the following data streams:

- **Visual observations:** Egocentric RGB streams captured via a head-mounted camera at 30 Hz, temporally interpolated to 50 Hz to synchronize with the high-frequency full-body control commands.

- **Proprioceptive states:** Comprehensive state **q** encompassing the 29 DoF whole-body joint positions, 12 DoF Inspire hand states, waist orientation, and the 2 DoF head pose.
- **Full-body control commands:** The corresponding command stream $\hat{\mathbf{p}}$ derived from the retargeted results, mirroring the structure of the proprioceptive states to ensure accurate demonstration replay and policy supervision.

VLA model training: We employ the GR00T N1.6 VLA model [28] as the high-level planner and post-train it on our collected dataset following standard protocols. Training is performed on four NVIDIA A100 GPUs for approximately 20k steps with a global batch size of 384 and a learning rate of 1×10^{-4} , keeping the backbone vision-language model frozen.

Inference protocols: At inference time, GR00T N1.6 processes the current visual observation, the natural language task description, and the robot’s proprioceptive state to predict an action chunk of 16 steps via a 4-step denoising process. **OmniClone** executes the first 8 actions in an iterative, receding-horizon fashion before the VLA output is refreshed. The VLA’s raw joint-space outputs are transformed into body link positions via Forward kinematics (FK), which then serve as observations for **OmniClone**—a process that inherently introduces potential noise to the control policy.

During autonomous inference, **OmniClone** functions as

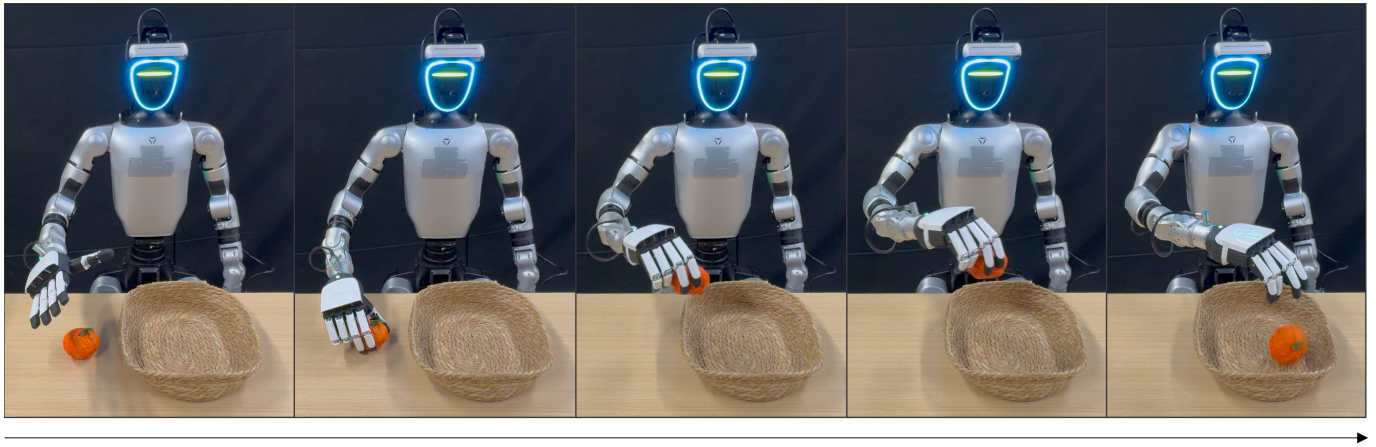


Fig. A.1: **Autonomous whole-body pick-and-place tasks executed by GR00T N1.6 trained on OmniClone-collected demonstrations.** The policy successfully performs full-body pick-and-place from a standing pose, validating the data quality, the fidelity of the control policy, and the stability of the system infrastructure.

the low-level controller while all communication and control command execution share the same system infrastructure as the teleoperation pipeline. This design imposes stringent demands on both the fidelity of the control policy and the stability of the infrastructure. The results in Appendix VI-B confirm that **OmniClone** provides reliable whole-body teleoperation supported by a stable system infrastructure for high-fidelity demonstration collection.