HuMiT: Low-Latency Whole-Body Humanoid Teleoperation via Minimal Reference Tracking

Anonymous

***Abstract*— Real-time whole-body humanoid teleoperation remains challenging due to the computational latency introduced by optimization-based retargeting pipelines and future-frame conditioning. Existing motion tracking policies typically require dense reference targets including full joint positions, body-part transformations, and even future frames, which tightly couples the controller to expensive upstream modules. In this work, we present *HuMiT*, a whole-body teleoperation system built on a unified tracking policy that requires only a *minimal* reference target, consisting only of root height, root velocity, and sparse keypoint positions at the current frame, yet achieves competitive or superior tracking fidelity compared to methods relying on more diverse reference states including joint positions, body transformations and even future frame. Experiments show that our policy achieves a communication-to-inference latency of $\sim$20 ms measured between headset streaming and policy output and a video-based optical flow latency of 80 ms (median) measured externally between operator and robot motion, significantly lower than the 200 ms+ latency typical of existing systems. We validate the system on a Unitree G1 humanoid, showcasing split-second reactive responses such as catching a tossed ball, as well as coordinated loco-manipulation capabilities. Project page: `https://humit-ral.github.io`.**

I. INTRODUCTION

Humanoid robots hold unique promise for operating in human-centric environments due to their morphology that naturally matches the workspaces and tools designed for people. Teleoperation is a practical paradigm for human-in-the-loop robot control. It provides a scalable mechanism for collecting diverse demonstration data to train autonomous policies, and also allows human operators to intervene in high-risk scenarios or situations where autonomous systems are likely to fail.

However, building an effective whole-body teleoperation system remains challenging. A primary issue is **large teleoperation latency**, which arises from two sources. First, optimization-based retargeting [1] introduces substantial per-frame delay and can suffer from numerical instability. Second, existing learning-based tracking policies rely heavily on *full reference states* as inputs, including complete joint positions, body-part transformations, and their velocities [2] [3]. Moreover, many of these methods condition the policy on *future* reference frames to enable anticipatory control [4] [5], which further exacerbates the latency problem. A secondary challenge is **limited whole-body coordination**, as many approaches decouple upper- and lower-body control [6]–[8], particularly restrictive for dynamic tasks. In addition, hardware-centric solutions based on exoskeletons [9]–[11] or marker-based motion capture [12] [13] further constrain

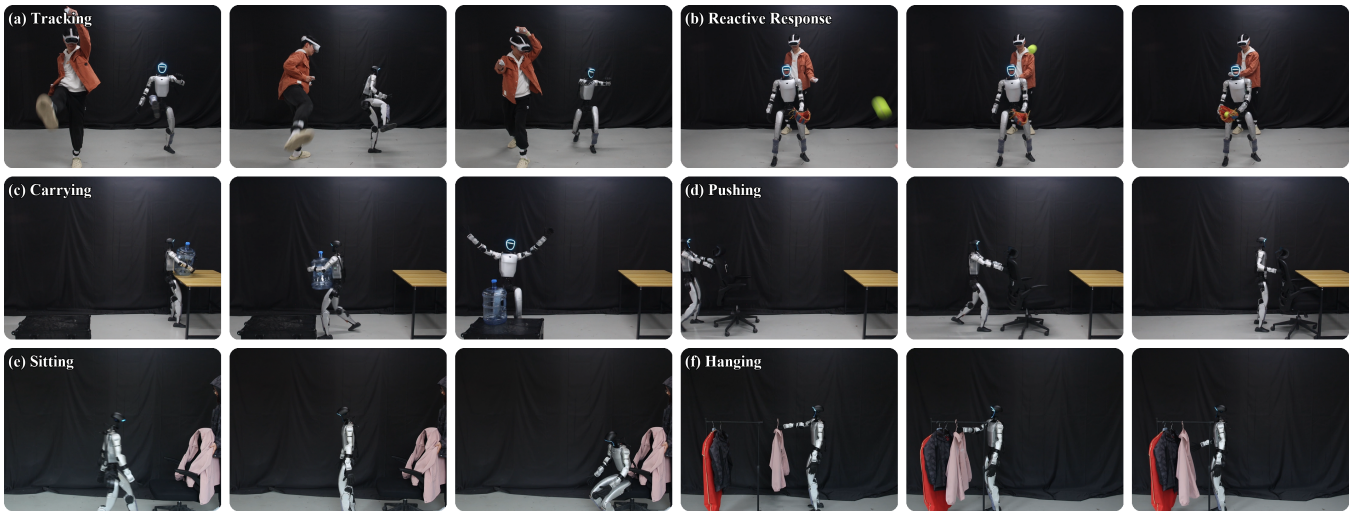

Fig. 1: Real-world teleoperation with HuMiT. By removing optimization-based retargeting and future-frame conditions, HuMiT supports a wide range of dynamic whole-body skills with a minimal-reference tracking policy: **(a)** Accurate whole-body *tracking* of operator motion, including front kick, side kick and boxing; **(b)** *reactive* ball catching enabled by low end-to-end latency; **(c)** *carrying* a water jug across the workspace; **(d)** *pushing* a wheeled chair while walking; **(e)** walking up to a chair, turning around and *sitting* down; **(f)** reaching overhead to *hang* a coat.

deployment flexibility.

In this work, we investigate a simple but consequential question: *is a full reference state truly necessary for high-fidelity whole-body motion tracking?* We find that a **minimal reference target**, consisting only of a root height, root linear and angular velocity commands, and sparse keypoint positions at the current frame, is sufficient to train a unified whole-body tracking policy that achieves competitive or superior tracking performance compared to methods relying on full reference states (i.e., complete joint positions, body-part orientations, their velocities, and future frames). This minimal reference design dramatically reduces the information that the upstream sensing module must provide, and it removes the retargeting pipeline from the online control loop entirely, resulting in significantly lower end-to-end latency.

We present **HuMiT**, a low-latency whole-body teleoperation system for humanoid robot. HuMiT employs lightweight human motion sensing to produce minimal reference targets, which are then tracked by a whole-body policy trained in massively parallel simulation. As shown in Figure. 1, HuMiT enables a wide range of dynamic whole-body skills that are difficult to achieve with high-latency or decoupled systems..

Our contributions are as follows:

- **Minimal reference suffices for whole-body tracking.** Through cross-method comparison and ablations under identical conditions, we show that a compact reference target without joint-level targets, body-part orientations, or future frames incurs no loss in tracking fidelity relative to full-reference policies.
- **Low-latency teleoperation system.** By removing optimization-based retargeting and future-frame conditioning from the online loop, HuMiT achieves a communication-to-inference latency of $\sim$20 ms measured between headset streaming and policy output and a video-based optical flow latency of 80 ms (median) measured externally between operator and robot motion, while maintaining robust tracking performance.
- **Dynamic whole-body teleoperation for humanoids.** We validate HuMiT on a Unitree G1 humanoid executing reactive catching and coordinated loco-manipulation, confirming that the unified minimal-reference design supports tightly coupled whole-body control under low-latency constraints.

II. Related Work

A. Learning-Based Humanoid Whole-Body Control

Decoupled vs. unified control. One line of work partitions the whole-body control problem into separate upper- and lower-body policies [6]–[8]. While this decomposition simplifies learning, it fundamentally limits coordinated movements that require coupling between limb motion and base stability. In contrast, unified policies integrate all degrees of freedom, enabling true loco-manipulation at the cost of increased training complexity.

Reference target design. Among unified tracking controllers, prior methods differs in what reference information the policy observes. BeyondMimic [2] conditions the policy on joint positions and joint velocities as reference targets. BFM [3] augments this with body-part local positions, providing denser spatial supervision. SONIC [4] and OmniClone [5] further extend the reference to include future frames for anticipatory control. Despite these differences in reference richness, all of these methods share the assumption that dense reference targets are necessary for high-fidelity tracking—an assumption that remains largely unexamined.

B. Humanoid Teleoperation Systems

Teleoperation enables humanoid robots to perform complex tasks by leveraging human intelligence [14]. Existing systems primarily differ in their sensing modality and in the resulting end-to-end latency.

Exoskeleton-based systems. Methods such as HOMIE [9], AirExo [10], and ACE [11] use wearable exoskeletons to capture the operator's upper-body motion, while the lower body is typically controlled via joysticks or pedals. This decoupled sensing naturally leads to decoupled control, limiting the robot's whole-body coordination.

Marker-based motion capture. TWIST [12] employs optical marker-based motion capture suites to provide high-accuracy full-body measurements, combined with a retargeting pipeline to produce robot reference states. While accurate, these systems are expensive, require extensive calibration, and restrict the operator to a confined capture volume.

Vision-based pose estimation. Camera-based methods estimate human body pose using models such as SMPL [6], [15]–[17], offering greater flexibility and lower hardware cost. However, the computational overhead of pose estimation followed by optimization-based retargeting makes it challenging to achieve the high-frequency, low-latency throughput required for responsive teleoperation. HumanPlus [18] and OmniH2O [19] adopt this approach but inherit its latency limitations.

Direct extremity control. A concurrent work, ExtremControl [20], operates directly on SE(3) extremity poses and incorporates velocity feedforward into the PD controller, reporting latency as low as 50 ms with optical motion capture and 60 ms with VR-based setup. However, it relies on teacher–student distillation. Meanwhile, for a robot with N degrees of freedom, $N/3$ link positions already provide sufficient constraints to determine the joint configuration; their use of full SE(3) transformations (position and orientation) per link introduces redundant reference information that is not necessary for resolving the kinematic state.

C. Motion Retargeting

Motion retargeting bridges the embodiment gap between the human operator and the target robot by converting captured human motion into robot-executable commands. Classical approaches solve per-frame inverse kinematics (IK) or nonlinear optimization problems [1] to map human joint positions onto the robot's kinematic structure. While these

methods produce geometrically accurate results, they introduce two practical drawbacks for teleoperation: (i) the per-frame optimization incurs non-negligible latency that accumulates in the control loop, and (ii) the numerical solvers are prone to instability (e.g., singularities, local minima), leading to jittery or infeasible reference commands.

In HuMiT, retargeting is performed *offline* only during training data preparation. At deployment, the tracking policy directly consumes scaled human keypoints as reference targets, completely removing retargeting from the online loop and eliminating its associated latency and instability.

III. Problem Formulation

We study low-latency whole-body humanoid teleoperation. Let $\mathbf{S}_t^h$ denote the human kinematic state at time t, obtained from an upstream motion sensing device (e.g., VR headset with body tracking). The state consists of root and body-part transformations in $SE(3)$:

$$\mathbf{S}_t^h = \{\mathbf{T}_{t,root}^h,\ \mathbf{T}_{t,j}^h\}_{j\in K},$$

where $\mathbf{T}_{t,root}^h = (\mathbf{R}_{t,root}^h, \mathbf{x}_{t,root}^h) \in SE(3)$ is the root (pelvis) transformation expressed in the world frame, with $\mathbf{R}_{t,root}^h \in SO(3)$ denoting its orientation and $\mathbf{x}_{t,root}^h \in \mathbb{R}^3$ its position. $\mathbf{T}_{t,j}^h = (\mathbf{R}_{t,j}^h, \mathbf{p}_{t,j}^h) \in SE(3)$ is the transformation of the j-th body part in the root frame, with $\mathbf{R}_{t,j}^h \in SO(3)$ its orientation and $\mathbf{p}_{t,j}^h \in \mathbb{R}^3$ its position. K denotes a predefined set of keypoints; in our implementation $|K| = 12$ (Section IV-B).

A teleoperation system must convert the human state stream $\mathbf{S}_{1:T}^h$ into robot-executable actions at a high frequency. We decompose this into two stages:

1) **Reference generation.** A mapping $\mathscr{F}$ converts the human state into a robot-frame reference target:
$$\mathbf{g}_t = \mathscr{F}(\mathbf{S}_t^h),$$
2) **Motion tracking.** A policy π maps onboard observations and the reference target to joint-space actions:
$$\mathbf{a}_t \sim \pi(\cdot \mid \mathbf{o}_t, \mathbf{g}_t).$$

The key design question we investigate is *how compact* $\mathbf{g}_t$ *can be* while still enabling high-fidelity whole-body tracking.

IV. Method

A. System overview.

Figure. 2 illustrates the HuMiT pipeline. Human motion is captured by a VR headset and converted into a minimal reference target (MRT) g_{mini} through a one-time scale calibration. The MRT is then consumed by a reinforcement-learning-based *whole-body tracking policy* π that directly outputs joint-space commands for the Unitree G1 humanoid ($N = 29$ DoF).

The central design choice in HuMiT is that π observes only the MRT, consisting of root height h_t^r, root linear velocity $\mathbf{v}_t^r$, root angular velocity ω_t^r, and body-part keypoint positions $\mathbf{p}_{t,j}^r$, $j \in K$ in the robot frame, without requiring joint-level references $\mathbf{q}_t^r$ and $\dot{\mathbf{q}}_t^r$, body-part orientations $\mathbf{R}_{t,j}^r$, $j \in K$ or any future-frame information.

This reduces the reference generation stage in Section III to two lightweight components: (1) a *motion sensing* module that captures the operator's body keypoints via VR headset (PICO 4 Ultra) with motion trackers, and (2) a *scale calibration* module that maps human keypoint positions to the robot's body positions in closed form, in place of an optimization-based retargeting loop.

B. From Human Motion to Minimal Reference Target

Motion sensing. We use a PICO 4 Ultra headset (as in Figure. 2) equipped with motion trackers to capture full-body motion at approximately 80 Hz. The system provides 6DoF tracking for 24 skeleton points, from which we extract positions of 12 bilateral keypoints ($|K| = 12$): hips, knees, ankles, shoulders, elbows, and wrists.

Scale calibration. To bridge the morphological difference between the human operator and the robot, we perform a one-time calibration following GMR [21]. At the start of each teleoperation session, the operator assumes a reference pose (e.g., T-pose). We compute per-limb scale factors α_j by comparing the operator's limb segment lengths to those of humanoid robot, and apply these factors to map incoming human keypoint positions into the robot's body frame:

$$\mathbf{p}_{t,j}^r = \alpha_j \cdot \mathbf{p}_{t,j}^h, \qquad j \in K.$$

The root height h_t^r is obtained as the z-component of the scaled root position, and the root velocity is computed via finite differences of the root position and orientation. All mappings are closed-form and add negligible latency ($<$1 ms).

C. Whole-Body Motion Tracking Policy

We train a unified whole-body tracking policy π via reinforcement learning in simulation. Training proceeds in two stages: (1) offline data preparation, in which human motion clips are retargeted to the robot via IK to produce full reference states g_{full}; and (2) online RL training, in which π learns to track the MRT using PPO with an asymmetric actor–critic architecture. The full reference g_{full} is used only to (i) initialize feasible simulator states, (ii) compute dense reward signals, and (iii) serve as an additional critic input. It is worth noting that the actor π itself never observes g_{full}.

Training data. We train on motion sequences from AMASS [22] ($\sim$ 35.7 hour) and LAFAN1 [23] [24] ($\sim$ 2.5 hour) after filtering out physically infeasible clips. For each human clip, we solve an optimization-based inverse kinematics problem to obtain a robot-frame reference trajectory

$$\{g_{full}\}_{t=1}^T = \{\mathbf{x}_{t,root}^r,\ \mathbf{R}_{t,root}^r,\ \mathbf{v}_t^r,\ \omega_t^r,\ \mathbf{p}_{t,j}^r,\ \mathbf{R}_{t,j}^r,\ \mathbf{q}_t^r,\ \dot{\mathbf{q}}_t^r\}_{t=1}^T, \tag{1}$$

which contains both link-level and joint-level references. We emphasize that this offline retargeting is performed *only once* during data preparation and is *never* part of the online teleoperation loop.

Observations. We employ an asymmetric actor–critic scheme. All quantities are expressed in the robot body frame unless otherwise stated, and symbols without the $(\cdot)^r$

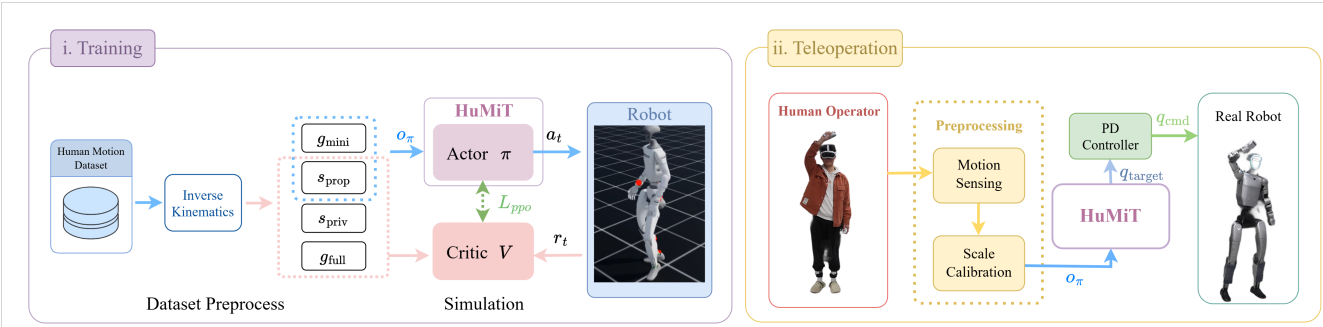

Fig. 2: Method overview. Human motion passes through scale calibration to produce a minimal reference target (g_{mini}); stages required by traditional pipelines (inverse kinematics, future-frame buffering) are eliminated. During training, the Actor π receives only the minimal reference g_{mini} and proprioception s_{prop}, and outputs actions a_t. The Critic additionally observes privileged states s_{priv} and full retargeted references g_{full}, along with dense reward r_t, enabling accurate value estimation. Both networks are updated jointly via PPO.

superscript refer to the robot's *measured* state (as opposed to the *reference* denoted by $(\cdot)^r$).

The observations are partitioned into four groups:

$$
\begin{aligned}
s_{prop} &= [\omega_t,\ \phi_t,\ \mathbf{q}_t,\ \dot{\mathbf{q}}_t,\ \mathbf{a}_{t-1}],\\
s_{priv} &= [\mathbf{v}_t,\ \mathbf{p}_{t,j},\ \mathbf{R}_{t,j}],\\
g_{full} &= [\mathbf{x}^r_{t,root},\ \mathbf{R}^r_{t,root},\ \mathbf{v}^r_t,\ \omega^r_t,\ \mathbf{p}^r_{t,j},\ \mathbf{R}^r_{t,j},\ \mathbf{q}^r_t,\ \dot{\mathbf{q}}^r_t],\\
g_{mini} &= [h^r_t,\ \mathbf{v}^r_t,\ \omega^r_t,\ \mathbf{p}^r_{t,j}] \subset g_{full},
\end{aligned}
$$

where s_{prop} is the robot proprioceptive state (ω_t is the measured root angular velocity, ϕ_t the projected gravity vector, $\mathbf{q}_t, \dot{\mathbf{q}}_t$ the measured joint positions and velocities, and $\mathbf{a}_{t-1}$ the previous action); s_{priv} contains ground-truth body states available only in simulation; g_{full} is the complete retargeted reference as defined in Eq. (1); and g_{mini} is a strict subset of g_{full} that retains only the root height, root velocities, and sparse keypoint positions—omitting joint-level targets, body-part orientations, and all future frames.

The actor and critic observations are defined as:

$$
\begin{aligned}
o_\pi &:= [s_{prop},\ g_{mini}],\\
o_v &:= [s_{prop},\ s_{priv},\ g_{full}].
\end{aligned}
$$

Action and low-level control. The policy outputs a joint-space offset $\mathbf{a}_t \in \mathbb{R}^N$:

$$
\mathbf{q}_{target} = \mathbf{q}_0 + \sigma\,\mathbf{a}_t,
$$

where $\mathbf{q}_0$ is the nominal standing configuration and σ scales the output to the feasible joint range. A PD controller computes motor torques:

$$
\tau = K_p(\mathbf{q}_{target} - \mathbf{q}_t) - K_d\,\dot{\mathbf{q}}_t.
$$

Reward. The tracking reward combines exponential tracking terms with regularization penalties:

$$
r_{\text{track}} = \sum_i w_i \exp\left(-\frac{e_i}{\delta_i^2}\right),
$$

where e_i is the i-th tracking error and w_i is its weight. Table I lists all reward terms; the "anchor" terms refer to the robot root (pelvis). Notably, the reward function uses the full retargeted reference (body positions, body orientations, velocities) for computing dense error signals during training, even though π itself only observes the minimal reference. This asymmetry between the reward supervision and the policy observation is key: it provides rich learning gradients while keeping the deployment interface compact.

TABLE I: Reward Terms. Rotations are compared using the axis–angle residual ($\ominus$).

Reward Term	Error Term	Weight
Anchor position	$\mathbf{x}^r_{t,root} - \mathbf{x}_{t,root}$	1.0
Anchor orientation	$\mathbf{R}^r_{t,root} \ominus \mathbf{R}_{t,root}$	1.0
Body position	$\mathbf{p}^r_{t,j} - \mathbf{p}_{t,j}$	2.0
Body orientation	$\mathbf{R}^r_{t,j} \ominus \mathbf{R}_{t,j}$	2.0
Body linear velocity	$\dot{\mathbf{p}}^r_{t,j} - \dot{\mathbf{p}}_{t,j}$	2.0
Body angular velocity	$\omega^r_{t,j} - \omega_{t,j}$	2.0
Feet slip (penalty)	$\sum_{j\in\text{feet}} \mathbf{v}_{t,j} \cdot \mathbb{I}_{contact}$	-1.0
Action rate (reg.)	$\mathbf{a}_t - \mathbf{a}_{t-1}$	-0.2

Early termination and domain randomizations. We terminate episodes when the root translation deviation exceeds 0.4 m, root orientation deviation exceeds 0.8, or keypoint deviation exceeds 0.4 m. To improve sim-to-real transfer, we randomize external pushes, contact friction and restitution, link masses (with stronger perturbations on end-effectors to approximate small payload variations), motor gains, armature parameters, and initial joint states. We note that these payload perturbations have limited magnitudes; heavy loads are not covered by our randomization and are discussed as a limitation in Section VI.

Training and deployment. We train with PPO [25] across 8192 parallel environments in Isaac Sim [26]. For real-world deployment on the Unitree G1, the trained Actor policy receives proprioceptive feedback from onboard sensors and the MRT streamed from the PICO headset, with no additional processing required.

V. Results

We design experiments to investigate three questions:

- **(Q1) Sufficiency of minimal reference:** Does the MRT

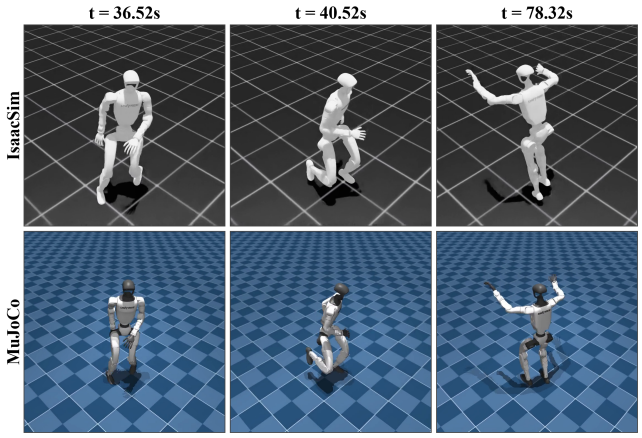

Fig. 3: Simulation Tracking Visualization. We do sim-to-sim transfer validation. The three columns show three successive time steps of the same motion clip rendered in Isaac Sim [26] and MuJoCo [27]. The policy, trained entirely in Isaac Sim, transfers to MuJoCo without retraining, producing consistent whole-body poses across both simulators.

suffice for high-fidelity whole-body tracking, compared to policies that observe full reference states?

- **(Q2) System latency:** What is the latency of HuMiT between human operator and robot motion, and how does it compare to state-of-the-art systems that rely on optimization-based retargeting and future frames?
- **(Q3) Real-world teleoperation:** Can HuMiT achieve dynamic whole-body teleportation in real, including tasks that require coordinated loco-manipulation and fast reactive response?

A. *Sufficiency of Minimal Reference Targets (**Q1**)*

We evaluate whether the MRT is sufficient for accurate whole-body tracking through two complementary comparisons: (1) against existing methods that use richer reference observations under different training frameworks, and (2) an ablation within HuMiT itself that isolates the effect of reference richness.

1) Metrics: We report three complementary metrics, including mean-per-keypoint position error E_{kp} (mm), which measures whole-body geometric tracking accuracy, root linear velocity error E_{lin} (m/s), and root angular velocity error E_{ang} (rad/s), which together capture locomotion fidelity and dynamic stability.

2) Baselines and Ablation: We compare against two methods with increasing reference richness:

- **BFM** [3]: It uses joint positions and body-part local positions as reference targets, and learns a *single unified policy across the dataset*. We use numbers reported from their original paper.
- **BeyondMimic** [2]: It uses joint positions and velocities as reference targets. The released codebase trains a *separate policy per motion clip*; We reproduce results using the official codebase and train on LAFAN1.

To isolate the effect of reference richness from differences in architecture and training, we additionally train a **full-reference** variant of HuMiT (**HuMiT-Full**) in which the policy π observes g_{full} instead of g_{mini} in addition to proprioception s_{prop}. All other components, including network architecture, reward function, domain randomization, and training procedure, remain identical.

TABLE II: Motion tracking evaluation. †: numbers from original paper. ‡: reproduced with released codebase.

Dataset	Method	$E_{kp}\downarrow$ (mm)	$E_{lin}\downarrow$ (m/s)	$E_{ang}\downarrow$ (rad/s)
AMASS (Test)	BFM†	61.12	0.30	1.43
	HuMiT-Full	52.74	**0.22**	**1.15**
	HuMiT	**48.63**	0.28	1.29
LAFAN1	BeyondMimic‡	**54.47**	0.38	1.38
	HuMiT-Full	57.84	0.41	1.34
	HuMiT	59.62	**0.33**	**1.34**

3) Quantitative Results: Results are in Table II.

Baseline comparisons. On the AMASS Test, HuMiT reduces E_{kp} by $\sim$20% relative to BFM despite observing significantly less reference information. On LAFAN1, BeyondMimic achieves a slightly lower E_{kp} (54.47 vs. 59.62 mm), but this comparison favors BeyondMimic in two ways: it has direct access to joint-level references, and its released codebase trains a separate policy per motion clip while HuMiT trains a single policy over the entire dataset. Despite this asymmetry, HuMiT closes the E_{kp} gap to $\sim$9% and achieves lower root-velocity errors on both linear and angular components.

Ablation: minimal vs. full reference. Under identical training conditions, HuMiT and HuMiT-Full achieve comparable overall performance. Neither variant consistently dominates: HuMiT is better on AMASS E_{kp}, while HuMiT-Full is better on LAFAN1 E_{kp}, and the two trade off closely on velocity metrics. This confirms that the additional reference information available to HuMiT-Full does not yield meaningful improvements in tracking accuracy. However, the two variants differ substantially in deployment cost: HuMiT-Full requires online inverse kinematics to produce its joint-level references, introducing $\sim$15ms latency gap.

These overall results support our central finding: a minimal reference target, without joint positions, body orientations, or future frames, is sufficient for training a whole-body tracking policy with competitive fidelity across different datasets and baselines.

B. *End-to-End Latency (Q2)*

A key advantage of HuMiT is the removal of retargeting and future-frame conditioning from the online control loop. For context, SONIC [4] conditions its policy on 10 future reference frames, introducing a buffering delay of at least $10 \times 20\,\text{ms} = 200\,\text{ms}$. TWIST2 [13] performs real-time IK retargeting at every control step, adding over 100 ms per-frame overhead. HuMiT eliminates both sources: the policy requires only the current-frame MRT, with no future look-ahead and no online IK.

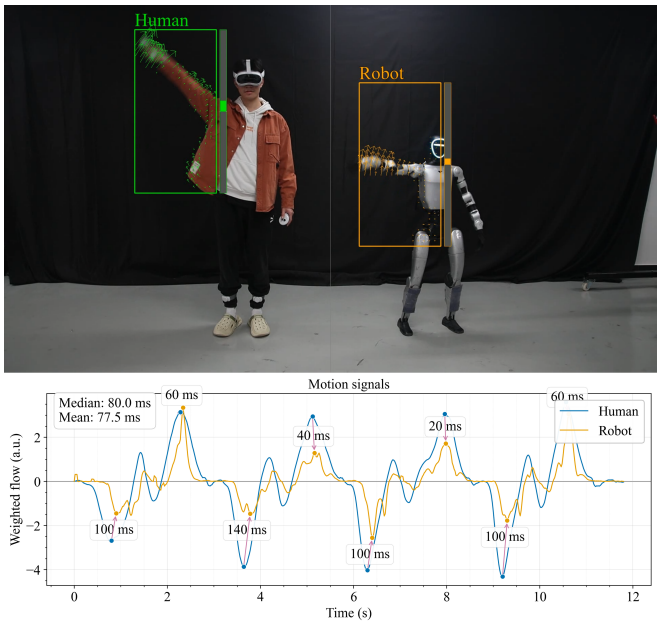

Fig. 4: Video-based latency measurement. Human (blue) and robot (orange) motion signals are extracted via optical flow from a 50 fps recording. Arrows indicate matched peak pairs with per-peak latency annotations. Upward peaks (20–60 ms) reflect the true pipeline latency; downward peaks (100–140 ms) are inflated by the G1 shoulder's kinematic constraints during arm lowering.

Measurement protocol. To verify this advantage empirically, we measure end-to-end latency using a video-based optical flow analysis that captures the externally observable delay between human and robot motion. We record the operator and robot simultaneously at 50 fps during a repetitive waving arm motion, compute dense optical flow projected onto the vertical axis to produce a magnitude-weighted motion signal for each half of the frame, and estimate latency by matching corresponding peaks between the two signals (Figure. 4).

Interpretation on the measurements. We observe a systematic asymmetry between the two phases of the waving motion. Upward (positive) peaks consistently exhibit lower latency (20–60 ms), while downward (negative) peaks show higher latency (100–140 ms). This is attributable to the kinematic structure of the Unitree G1 shoulder: its three revolute joints permit fast arm raising but create a more constrained return path during lowering, which slows execution and inflates the apparent latency. The upward-phase measurements are therefore a more faithful indicator of the underlying communication-and-inference latency of 20 ms.

Overall latency. Aggregating across the 8 matched peak pairs (both phases), the end-to-end latency has a **median of 80 ms** (mean 77.5 ± 36.7 ms). Even the worst-case measurement (140 ms) remains well below SONIC's buffering delay ($\geq$200 ms) and is comparable to TWIST2's retargeting overhead ($>$100 ms), while the median and best-case values are substantially lower than both.

C. Real-World Teleoperation (*Q3*)

We deploy HuMiT on a Unitree G1 to evaluate three capabilities that are directly enabled by its low-latency, unified-policy design: accurate real-time tracking, fast reactive responses, and coordinated loco-manipulation. Snapshots of all real-world demonstrations are shown in Figure. 1.

1) Real-Time Tracking: We first stream live operator motion to the robot and evaluate qualitatively how closely the G1 mirrors the human (Figure. 1a). The robot reproduces a wide range of whole-body poses—including front kick, side kick, boxing, and quick directional changes—driven by the minimal reference target alone.

2) Reactive Catching: To validate the low-latency advantage, we conduct a task as shown in Figure. 1b: A ball is thrown toward the robot, who mirrors the operator's motion in real time and catches the ball with its glove. This task is inherently difficult for high-latency systems, the success of this demonstration provides direct, qualitative evidence that HuMiT's latency is low enough for time-critical interaction.

3) Coordinated Loco-Manipulation: We further evaluate four loco-manipulation tasks that require tight coordination between locomotion and manipulation (Fig. 1c–f): *carrying* a water jug across the workspace, *pushing* a wheeled chair while walking, walking up to and *sitting* on a chair, and reaching overhead to *hang* a coat. In all four tasks, the unified policy maintains balance during locomotion without degrading manipulation accuracy. These behaviors are difficult to obtain with decoupled upper-/lower-body controllers, where the lower body is typically driven by joystick or pedal commands that cannot be tightly synchronized with arm motions.

VI. CONCLUSIONS AND LIMITATIONS

We presented HuMiT, showing that a minimal reference target is sufficient for training a whole-body tracking policy with competitive fidelity, achieving a communication-to-inference latency of $\sim$20 ms and an externally measured end-to-end latency of 80 ms (median)—significantly lower than existing systems. Real-world experiments on the Unitree G1 demonstrate that this latency enables reactive tasks such as catching a tossed ball, as well as coordinated loco-manipulation, validating the approach.

HuMiT also has several limitations. First, the training data is dominated by upright locomotion and upper-body gestures; motions involving ground contact such as kneeling or crawling are underrepresented, leading to degraded tracking accuracy or failure on these behaviors. Second, the morphological mismatch between the human and the G1 reduces the precision of fine wrist control, limiting performance on manipulation tasks that demand accurate end-effector positioning. Third, current hardware and teleoperation policy design do not support high-payload loco-manipulation (e.g., carrying heavy objects while walking), as the policy has not been trained under such loading conditions. Addressing these limitations through broader training distributions, operator-adaptive calibration, and load-aware policy learning are promising directions for future work.

REFERENCES

[1] Z. Luo, J. Cao, K. Kitani, W. Xu, *et al.*, "Perpetual humanoid control for real-time simulated avatars," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 895–10 904.

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

[3] W. Zeng, S. Lu, K. Yin, X. Niu, M. Dai, J. Wang, and J. Pang, "Behavior foundation model for humanoid robots," *arXiv preprint arXiv:2509.13780*, 2025.

[4] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castaneda, Z.-A. Cao, J. Li, D. Minor, Q. Ben, *et al.*, "Sonic: Supersizing motion tracking for natural humanoid whole-body control," *arXiv preprint arXiv:2511.07820*, 2025.

[5] K. Yin, W. Zeng, K. Fan, M. Dai, Z. Wang, Q. Zhang, Z. Tian, J. Wang, J. Pang, and W. Zhang, "Unitracker: Learning universal whole-body motion tracker for humanoid robots," *arXiv preprint arXiv:2507.07356*, 2025.

[6] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.

[7] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.

[8] J. Li, X. Cheng, T. Huang, S. Yang, R.-Z. Qiu, and X. Wang, "Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control," *arXiv preprint arXiv:2505.03738*, 2025.

[9] 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.

[10] H. Fang, C. Wang, Y. Wang, J. Chen, S. Xia, J. Lv, Z. He, X. Yi, Y. Guo, X. Zhan, *et al.*, "Airexo-2: Scaling up generalizable robotic imitation learning with low-cost exoskeletons," *arXiv preprint arXiv:2503.03081*, 2025.

[11] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang, "Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation," *arXiv preprint arXiv:2408.11805*, 2024.

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

[13] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, "Twist2: Scalable, portable, and holistic humanoid data collection system," *arXiv preprint arXiv:2511.02832*, 2025.

[14] K. Darvish, L. Penco, J. Ramos, R. Cisneros, J. Pratt, E. Yoshida, S. Ivaldi, and D. Pucci, "Teleoperation of humanoid robots: A survey," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1706–1727, 2023.

[15] J. Li, J. Cao, H. Zhang, D. Rempe, J. Kautz, U. Iqbal, and Y. Yuan, "Genmo: A generalist model for human motion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 11 766–11 776.

[16] Z. Shen, H. Pi, Y. Xia, Z. Cen, S. Peng, Z. Hu, H. Bao, R. Hu, and X. Zhou, "World-grounded human motion recovery via gravity-view coordinates," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[17] Y. Wang, Z. Wang, L. Liu, and K. Daniilidis, "Tram: Global trajectory and motion of 3d humans from in-the-wild videos," in *European Conference on Computer Vision*. Springer, 2024, pp. 467–487.

[18] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[19] 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.

[20] Z. Xiong, L. Fang, J. Huang, K. Yamazaki, H. Zhang, and C. Gan, "Extremcontrol: Low-latency humanoid teleoperation with direct extremity control," *arXiv preprint arXiv:2602.11321*, 2026.

[21] Y. Ma, H. Yu, J. Xie, C. Lv, Q. Luo, C. Zhang, Y. Yin, B. Xing, X. Ren, and D. Zheng, "Robust and generalized humanoid motion tracking," *arXiv preprint arXiv:2601.23080*, 2026.

[22] 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.

[23] F. G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. Pal, "Robust motion in-betweening," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 60–1, 2020.

[24] lvhaidong, "Lafan1 retargeting dataset," https://huggingface.co/datasets/lvhaidong/LAFAN1_Retargeting_Dataset, 2025, hugging Face dataset; accessed 2026-04-24.

[25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[26] M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du, A. Serrano-Munoz, X. Yao, R. Zurbrügg, N. Rudin, *et al.*, "Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning," *arXiv preprint arXiv:2511.04831*, 2025.

[27] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.