

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

Liu Cao
Tsinghua IIIS

Botian Xu
The Chinese University of Hong Kong

Ruoqu Chen
Tsinghua IIIS

Jiajun Liu
Renmin University of China

Mingzhi Pei
Tsinghua College AI

Zihao Wang
Tsinghua Weiyang College

Zhongyu Li
The Chinese University of Hong Kong

Mengdi Xu
Tsinghua IIIS

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 *compact* reference target, consisting of root height, root velocity, and the SE(3) poses of four sparse keypoints at the current frame, yet achieves competitive tracking fidelity compared to methods relying on more diverse reference states including complete joint positions, dense body-part transformations, and even future frames. Experiments show that our policy achieves a communication-to-inference latency of ~ 20 ms measured between headset streaming and policy output and a video-based optical flow latency of 65 ms (median) across different motion types. 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://xiaohu-art.github.io/Track/>.

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],

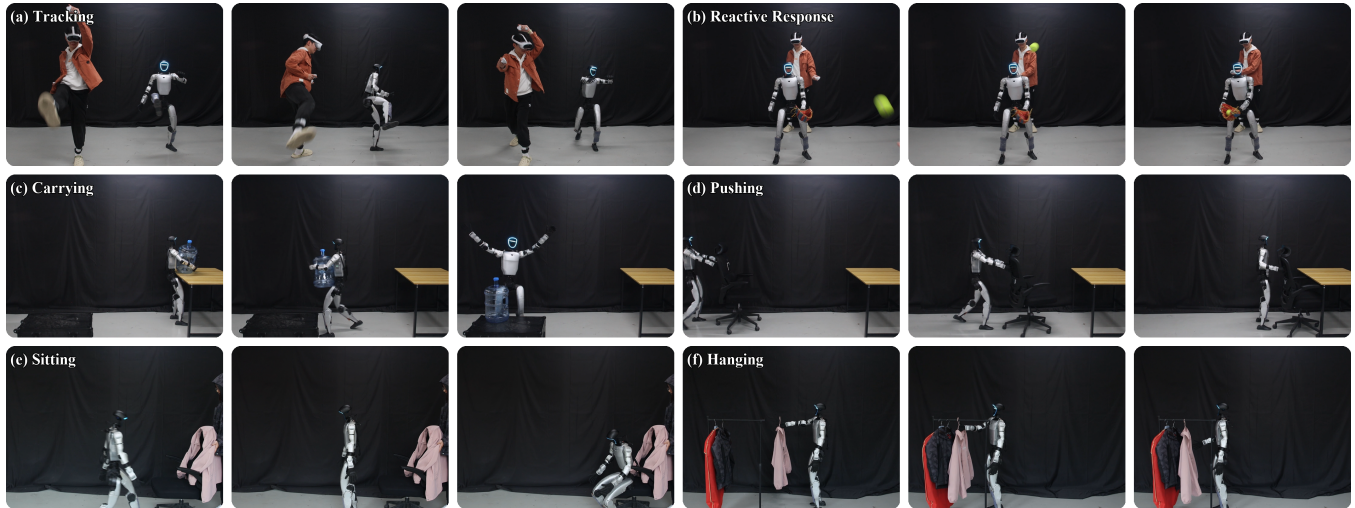


Fig. 1: Real-world teleoperation with HuMiT. By removing optimization-based retargeting and future-frame conditions, HuMiT supports a wide range of dynamic whole-body skills with a compact-reference tracking policy: (a) Accurate whole-body *tracking* of operator motion, including front kick, side kick 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.

[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], which is particularly restrictive for dynamic tasks. In addition, hardware-centric solutions based on exoskeletons [9]–[11] or marker-based motion capture [12], [13] further constrain deployment flexibility.

In this work, we investigate a simple but consequential question: *does high-fidelity whole-body motion tracking truly require a full reference state at deployment?* We find that a **compact reference target (CRT)**, consisting only of a root height, root linear and angular velocity commands, and the SE(3) poses of four sparse keypoints (bilateral wrists and ankles) at the current frame, is sufficient as the deployment-time input of a unified whole-body tracking policy that achieves competitive tracking performance compared to methods relying on full reference states (i.e., complete joint positions, dense body-part states, their velocities, and future frames). This compact 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 robots. HuMiT employs lightweight human motion sensing to produce compact reference targets, which are then tracked by a whole-body policy trained in parallel simulation. As shown in Fig. 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:

- **Compact reference suffices at deployment.** Through cross-method comparison and ablations under identical conditions, we show that a compact reference target matches 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 ~ 20 ms measured between headset streaming and policy output and a video-based optical flow latency of 65 ms (median) 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 locomanipulation, confirming that the unified compact-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. Partitioning whole-body control into separate upper- and lower-body policies [6]–[8] simplifies learning but fundamentally limits movements

that couple limb motion and base stability, whereas unified policies integrate all degrees of freedom, enabling true locomanipulation at the cost of increased training complexity.

Reference target design. Among unified tracking controllers, prior methods differ 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

Existing teleoperation systems [14] differ mainly in their sensing modality and in the resulting end-to-end latency.

Hardware-centric sensing. Exoskeleton-based systems [9]–[11] capture the operator’s upper body while the lower body is driven by joysticks or pedals, so sensing leads to decoupled control. Optical marker-based motion capture suits [12] are accurate but expensive, require extensive calibration, and restrict the operator to a constrained workspace.

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, but the computational overhead of pose estimation followed by optimization-based retargeting makes the high-frequency, low-latency throughput required for responsive teleoperation hard to reach. HumanPlus [18] and OmniH2O [19] adopt this approach and inherit its latency limitations.

Direct extremity control. A concurrent work, Extremity Control [20], operates directly on the SE(3) poses of six links (pelvis, torso, hands, and feet) and incorporates velocity feedforward into the PD controller, reporting latency as low as 50 ms with optical motion capture and 60 ms with a VR-based setup. Its deployable policy, however, is distilled from a privileged teacher that observes up to 32 future reference steps, and its reference is anchored by the operator’s global pelvis and torso poses. In contrast, HuMiT is trained in a single stage without distillation or any future-frame supervision, and replaces global pose anchoring with root height and velocity commands.

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], [21] to map human joint positions onto the robot’s kinematic structure. While geometrically accurate, 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.,

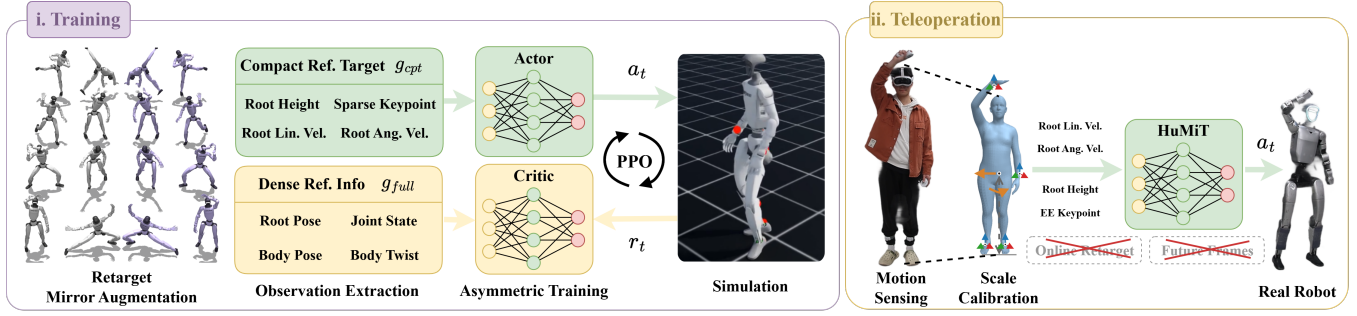


Fig. 2: Method overview. Human motion passes through scale calibration to produce a compact reference target (g_{cpt}). Stages required by traditional pipelines (inverse kinematics, future-frame buffering) are eliminated. During training, the Actor π receives only the compact reference g_{cpt} and proprioception 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.

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 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 $SE(3)$: $S_t^h = \{\mathbf{T}_{t,root}^h, \mathbf{T}_{t,j}^h\}_{j \in K}$, 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 setup, $|K| = 12$, including the bilateral hips, knees, ankles, shoulders, elbows, and wrists.

A teleoperation system must convert the human state stream $S_{1:T}^h$ into robot-executable actions at a high frequency. We decompose this into two stages: **reference generation**, in which a mapping $\mathcal{F}$ converts the human state into a robot-frame reference target $\mathbf{g}_t = \mathcal{F}(S_t^h)$, and **motion tracking**, in which a policy π maps onboard observations and the reference target to joint-space actions $\mathbf{a}_t \sim \pi(\cdot | \mathbf{o}_t, \mathbf{g}_t)$. 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

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

The central design choice in HuMiT is that π observes only the CRT, consisting of root height h_t^r , root linear velocity $\mathbf{v}_t^r$, root angular velocity $\boldsymbol{\omega}_t^r$, and the root-relative $SE(3)$ poses $(\mathbf{R}_{t,j}^r, \mathbf{p}_{t,j}^r)$ of a sparse keypoint subset $j \in K_c \subset K$ (bilateral wrists and ankles, $|K_c| = 4$), without requiring joint-level references $\mathbf{q}_t^r$ and $\dot{\mathbf{q}}_t^r$, the states of the remaining body parts, 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 Compact Reference Target

Motion sensing. Our sensing front-end is entirely consumer-grade as in Fig. 2: a PICO 4 Ultra VR headset, two handheld PICO controllers, and two PICO Motion Trackers strapped to the operator’s lower legs. From this setup, the PICO body-tracking runtime infers a full-body skeleton of 24 points with 6 DoF each at approximately 80 Hz and streams it to the control host over Wi-Fi. From this stream we extract the $SE(3)$ poses of the four keypoints in $K_c \subset K$. Putting on the entire wearable set takes less than a minute, and the system imposes no workspace constraint beyond Wi-Fi coverage, in contrast to exoskeleton or optical motion-capture pipelines.

Scale calibration. To bridge the morphological difference between the human operator and the robot, we perform a one-time calibration following GMR [22]. 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 the humanoid robot, and apply these factors to map incoming human keypoint positions into the robot’s body frame, while keypoint orientations require no morphological rescaling and are passed through unchanged:

$$\mathbf{p}_{t,j}^r = \alpha_j \cdot \mathbf{p}_{t,j}^h, \quad \mathbf{R}_{t,j}^r = \mathbf{R}_{t,j}^h, \quad j \in K_c.$$

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 transformation. 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 CRT 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 [23], LAFAN1 [24], [25], and BONES-SEED [4] after filtering out physically infeasible clips. Every clip is further augmented with its sagittal-plane mirror, which swaps the left and right joints and reflects the root trajectory about the vertical plane through the initial heading. For each human clip, we solve an optimization-based inverse kinematics problem to obtain a robot-frame reference trajectory

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

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. Symbols without the $(\cdot)^r$ superscript refer to the robot's *measured* state as opposed to the *reference* $(\cdot)^r$.

The observations are partitioned into four groups:

$$\begin{aligned} s_{prop} &= [\omega_t, \phi_t, \mathbf{q}_t, \dot{\mathbf{q}}_t, \mathbf{a}_{t-1}], \\ s_{priv} &= [\mathbf{v}_t, \mathbf{p}_{t,j}, \mathbf{R}_{t,j}], \\ g_{full} &= [\mathbf{x}_{t,root}^r, \mathbf{R}_{t,root}^r, \mathbf{v}_t^r, \omega_t^r, \mathbf{p}_{t,j}^r, \mathbf{R}_{t,j}^r, \mathbf{q}_t^r, \dot{\mathbf{q}}_t^r], \\ g_{cpt} &= [h_t^r, \mathbf{v}_t^r, \omega_t^r, \{\mathbf{p}_{t,j}^r, \mathbf{R}_{t,j}^r\}_{j \in K_c}] \subset g_{full}, \end{aligned}$$

where s_{prop} is the robot proprioceptive state (ω_t is the measured root angular velocity in the root frame, ϕ_t the gravity vector projected into the root frame, $\mathbf{q}_t, \dot{\mathbf{q}}_t$ the measured joint positions and velocities, and $\mathbf{a}_{t-1}$ the previous action). The privileged group s_{priv} contains ground-truth body states available only in simulation (the root linear velocity $\mathbf{v}_t$ in the root frame, and the root-relative body poses $\mathbf{p}_{t,j}, \mathbf{R}_{t,j}$). The full reference g_{full} is defined in Eq. (IV-C), and g_{cpt} is a strict subset of g_{full} that retains only the root height (the world-frame z -coordinate of the root above the ground), root velocities (in the root frame), and the root-relative sparse keypoint poses over K_c (orientations encoded in the 6D rotation representation), omitting joint-level targets, the dense states of the remaining body parts, and future frames.

The actor and critic observations are defined as:

$$o_\pi := [s_{prop}, g_{cpt}], \quad o_v := [s_{prop}, s_{priv}, g_{full}].$$

Action and low-level control. The policy outputs a joint-space offset $\mathbf{a}_t \in \mathbb{R}^N$ that yields the joint targets $\mathbf{q}_{target} = \mathbf{q}_0 + \sigma \mathbf{a}_t$, where $\mathbf{q}_0$ is the nominal standing configuration and σ scales the output to the feasible joint range. A PD controller then computes motor torques $\tau = K_p(\mathbf{q}_{target} - \mathbf{q}_t) - K_d \dot{\mathbf{q}}_t$.

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

Reward Term	Error Term	Weight
Anchor position	$\mathbf{x}_{t,root}^r - \mathbf{x}_t$	1.0
Anchor orientation	$\mathbf{R}_{t,root}^r \ominus \mathbf{R}_t$	1.0
Body position	$\mathbf{p}_{t,j}^r - \mathbf{p}_{t,j}$	2.0
Body orientation	$\mathbf{R}_{t,j}^r \ominus \mathbf{R}_{t,j}$	2.0
Body linear velocity	$\dot{\mathbf{p}}_{t,j}^r - \dot{\mathbf{p}}_{t,j}$	2.0
Body angular velocity	$\omega_{t,j}^r - \omega_{t,j}$	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

Reward. The tracking reward combines exponential tracking terms with regularization penalties: $r_{track} = \sum_i w_i \exp\left(-\frac{e_i}{\delta_i^2}\right)$, where e_i is the i -th tracking error and w_i is its weight. Table I lists all reward terms, where the ‘‘anchor’’ terms refer to the robot root (pelvis). Notably, the reward function uses the full retargeted reference (body positions, body orientations, velocities) for dense error signals during training, even though π itself only observes the compact reference. This asymmetry between the reward supervision and the policy observation is key: it provides rich learning gradients while keeping the deployment interface compact.

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.

Training and deployment. We train with PPO [26] across 8192 parallel environments in Isaac Sim [27]. For real-world deployment on the Unitree G1, the trained Actor policy receives proprioceptive feedback from onboard sensors and the CRT streamed from the PICO headset, with no additional processing required.

V. RESULTS

We design experiments to investigate three questions:

- **(Q1) 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?
- **(Q2) Sufficiency of compact reference:** Does the CRT suffice as the policy input at deployment, compared to policies that observe full reference states?
- **(Q3) Real-world teleoperation:** Can HuMiT achieve dynamic whole-body teleoperation in the real world, including tasks that require coordinated loco-manipulation and fast reactive response?

A. End-to-End Latency (Q1)

Conditioning a policy on k future reference frames delays at least $k\Delta t$ of reference buffering (≥ 200 ms for 10 frames at 20 ms spacing).

Internal timestamped measurement. We record timestamps from the arrival of a skeleton packet streamed by the headset to the policy’s action, covering wireless transmission, CRT assembly, and policy inference. This yields a communication-to-inference latency of ~ 20 ms, excluding the sensing latency and the robot’s actuation response.

External optical-flow estimation. To capture the full operator-to-robot delay including physical execution, we record the operator and the robot in the same camera frame, compute dense optical flow between consecutive frames within an operator region and a robot region, and reduce each region to one standardized motion signal by averaging its flow along the vertical direction. We then detect prominent peaks in the operator signal and match each to the robot extremum of the same sign, and the per-peak latency is the time offset of the matched pair (Fig. 3). We record three motion types at 200 fps (5 ms temporal resolution): punching, squatting, and waving, yielding median (mean) latencies of 78 (91) ms with $n=4$ matched pairs, 60 (63) ms with $n=5$, and 85 (91) ms with $n=7$, respectively. The pooled median (mean) over all 16 pairs is 65 (83) ms.

Baseline comparisons. Table II applies the same protocol to the systems we can deploy with our sensing front-end. Shortening the lookahead of SONIC from 10 to 4 future frames (SONIC-LL) drops the delay from 245 to 145 ms, in line with the 120 ms of buffering removed. TWIST2 buffers nothing but retargets online and measures 120 ms, while HuMiT, which does neither, measures 65 ms.

TABLE II: End-to-end teleoperation latency, measured with the external optical-flow protocol.

Method	Buffering (ms)	Online retarget.	End-to-end (ms)
SONIC [4]	200	×	245
SONIC-LL	80	×	145
TWIST2 [13]	0	✓	120
HuMiT (ours)	0	×	65

B. Sufficiency of Compact Reference Targets (Q2)

We evaluate whether the CRT is sufficient as the policy input at deployment 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 five complementary metrics over T evaluated frames, the keypoint set K ($|K| = 12$), and the $N = 29$ actuated joints:

$$E_{kp} = \frac{1}{T|K|} \sum_{t,j} \|\mathbf{x}_{t,j}^r - \mathbf{x}_{t,j}\|_2, \quad E_{jpe} = \frac{1}{TN} \sum_t \|\mathbf{q}_t^r - \mathbf{q}_t\|_1,$$

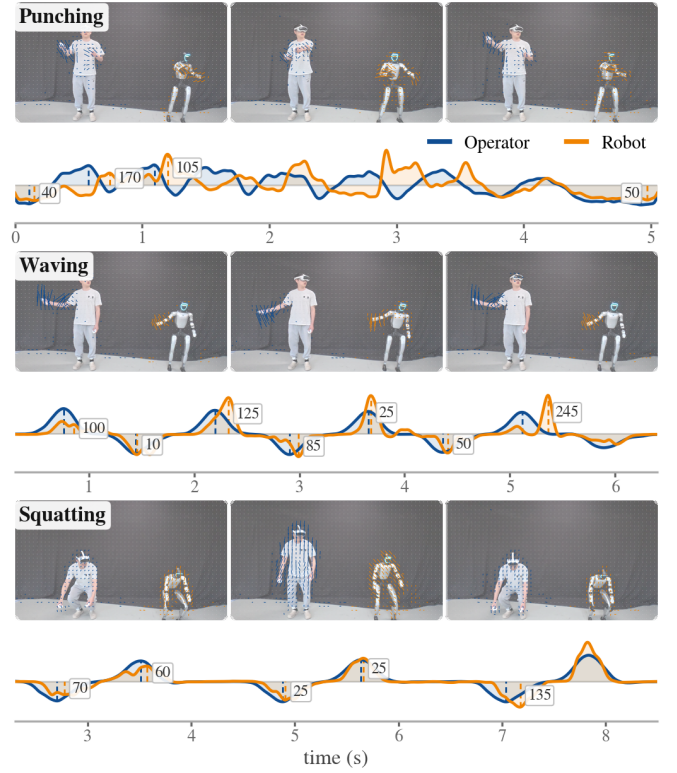


Fig. 3: End-to-end latency across motion types. For each motion, the flow-overlay frames of the operator (blue) and the robot (orange) are shown above. White boxes give per-pair delays in ms.

where $\mathbf{x}_{t,j}$ is the world-frame position of keypoint j , in contrast to the root-relative $\mathbf{p}_{t,j}$, so that the mean per-keypoint error E_{kp} (mm) measures whole-body geometric accuracy and the mean joint position error E_{jpe} (rad) joint-level accuracy. The root velocity errors $E_{lin} = \frac{1}{T} \sum_t \|\mathbf{v}_t^r - \mathbf{v}_t\|_2$ (m/s) and $E_{ang} = \frac{1}{T} \sum_t \|\omega_t^r - \omega_t\|_2$ (rad/s) capture root tracking fidelity, and the success rate (SR, %) is the fraction of rollouts that reach the end of the clip without triggering the early-termination criteria of Section IV-C.

2) Baselines: We compare against five external baselines spanning a spectrum of reference richness, plus our controlled full-reference variant. **SONIC** [4] is a unified tracking policy trained on BONES-SEED, conditioned on dense reference states over *10 future frames* (~ 200 ms of lookahead), and **SONIC-LL** is its low-latency version, which shortens the reference lookahead to *4 future frames* (~ 80 ms) for more responsive teleoperation. **ScaleBFM (M/XL)** [28] are behavior foundation models built on [3] that track masked root-relative target poses of body links over *6 future frames* (~ 120 ms). We evaluate both officially released Humanoid Transformer checkpoints (M: 3.0M, XL: 9.9M parameters). **ExtremControl** [20] is a concurrent low-latency method tracking the SE(3) poses of six links (pelvis, torso, hands, and feet), obtained via privileged teacher–student distillation.

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

TABLE III: Whole-body motion tracking evaluation. E_{kp} (mm), E_{jpe} (rad), E_{lin} (m/s), E_{ang} (rad/s), SR (%). A compact reference target, without joint-level targets, dense per-link states, or future frames, is sufficient as the deployment-time input of a whole-body tracking policy with competitive fidelity.

Method	AMASS					LAFAN1					BONES-SEED				
	$E_{kp} \downarrow$	$E_{jpe} \downarrow$	$E_{lin} \downarrow$	$E_{ang} \downarrow$	SR $\uparrow$	$E_{kp} \downarrow$	$E_{jpe} \downarrow$	$E_{lin} \downarrow$	$E_{ang} \downarrow$	SR $\uparrow$	$E_{kp} \downarrow$	$E_{jpe} \downarrow$	$E_{lin} \downarrow$	$E_{ang} \downarrow$	SR $\uparrow$
SONIC [4]	316.0	0.116	0.141	0.312	79.8	1492.6	0.097	0.221	0.394	20.0	143.1	0.098	0.112	0.295	98.2
SONIC-LL	400.5	0.113	0.177	0.397	59.5	1524.5	0.110	0.245	0.537	15.0	229.9	0.098	0.136	0.384	96.5
ScaleBFM-M [28]	73.2	0.123	0.145	0.444	92.6	82.1	0.102	0.226	0.699	27.5	60.2	0.079	0.136	0.435	97.7
ScaleBFM-XL [28]	71.8	0.125	0.137	0.407	94.0	82.8	0.097	0.224	0.621	32.5	58.5	0.081	0.128	0.403	98.7
ExtremControl [20]	329.5	0.246	0.259	0.675	60.1	191.0	0.203	0.244	0.781	0.0	220.2	0.138	0.203	0.638	66.3
HuMiT-Full (ours)	307.9	0.132	0.146	0.399	87.5	132.1	0.054	0.370	0.732	62.5	154.6	0.092	0.089	0.212	96.6
HuMiT (ours)	313.5	0.133	0.145	0.391	87.6	148.0	0.035	0.370	0.716	52.5	155.0	0.126	0.090	0.203	96.2

TABLE IV: Real-world teleoperation reported as successful trials over total trials with a single operator. SONIC-LL and TWIST2 never catch the ball (0/10).

Task	Succ./trials	Dominant failure mode
Reactive catching	2 / 10	Arm workspace differs from operator's
Carrying a water jug	9 / 10	Jug tips out of the grasp
Pushing a wheeled chair	5 / 10	Chair heading hard to steer
Sitting on a chair	7 / 10	Chair out of view after turning
Hanging a coat	6 / 10	Hanger orientation uncontrolled

policy π observes g_{full} instead of g_{cpt} in addition to proprioception s_{prop} . All other components, including network architecture, reward function, domain randomization, and training procedure, remain identical, making HuMiT vs. HuMiT-Full the controlled comparison that isolates reference richness with the training methodology held fixed.

3) *Quantitative Results:* Table III shows that HuMiT attains competitive tracking fidelity across all three datasets while observing the least reference information among the compared methods and no future frames.

Baseline comparisons. On AMASS, HuMiT tracks at $E_{kp} = 313.5$ mm with an 87.6% success rate, and its root-velocity errors ($E_{lin} = 0.145$ m/s, $E_{ang} = 0.391$ rad/s) are on par with the strongest baselines. On LAFAN1, the most dynamic dataset, it attains the highest success rate among all external baselines (52.5%) and the lowest joint error of all compared methods ($E_{jpe} = 0.035$ rad). On BONES-SEED, it reaches a 96.2% success rate with the lowest root angular-velocity error of all compared methods ($E_{ang} = 0.203$ rad/s) and an E_{lin} of 0.090 m/s within 0.001 m/s of the best. The world-frame E_{kp} , however, favors ScaleBFM, which observes global root-position targets and corrects drift online.

Ablation: compact vs. full reference. Under identical training conditions, HuMiT and HuMiT-Full are nearly indistinguishable on AMASS and BONES-SEED. The most notable gap is on LAFAN1, indicating that a dense reference target still helps for tracking highly dynamic motion.

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 Fig. 1, and Table IV reports the outcome of ten trials per task together with the dominant failure mode we observed.

Real-time tracking. We first stream live operator motion to the robot and evaluate qualitatively how closely the G1 mirrors the human (Fig. 1a). The robot reproduces a wide range of whole-body poses driven by the compact reference target alone.

Reactive catching. To validate the low-latency advantage, we conduct a task as shown in Fig. 1b: A ball is thrown toward the robot, which mirrors the operator's motion in real time and catches the ball with its glove. This task is inherently difficult for high-latency systems, since the robot has to react within the flight time of the ball. HuMiT catches the ball in 2 of 10 trials, whereas SONIC-LL and TWIST2 never succeed, and its remaining failures come from the mismatch between the operator's and the robot's arm workspace rather than from a delayed response.

Coordinated loco-manipulation. We further evaluate four loco-manipulation tasks that require tight coordination between locomotion and manipulation (Fig. 1c-f). In all of them, the unified policy maintains balance during locomotion without degrading manipulation accuracy. SONIC-LL and TWIST2 remain stable on these four tasks, but their higher latency requires the operator to pause and correct after each motion, which makes teleoperated data collection substantially slower. 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 compact reference target is sufficient as the deployment-time input of a whole-body tracking policy with competitive fidelity, achieving a communication-to-inference latency of ~ 20 ms and an externally measured end-to-end latency of 65 ms (median). Real-world experiments 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, so motions involving ground contact such as kneeling or crawling are underrepresented, leading to degraded tracking accuracy or failure on these behaviors. Second, 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 and load-aware policy learning is a promising direction 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] Y. Li, L. Ma, Y. Lin, Y. Du, M. Liu, K. Hu, J. Cui, Y. Zhu, W. Liang, B. Jia, and S. Huang, “Omniclone: Engineering a robust, all-rounder whole-body humanoid teleoperation system,” *arXiv preprint arXiv:2603.14327*, 2026.
- [6] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, “Expressive whole-body control for humanoid robots,” in *Robotics: Science and Systems (RSS)*, 2024.
- [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,” in *Robotics: Science and Systems (RSS)*, 2025.
- [9] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, “Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit,” in *Robotics: Science and Systems (RSS)*, 2025.
- [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,” in *Conference on Robot Learning (CoRL)*. PMLR, 2025, pp. 198–220.
- [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,” in *Conference on Robot Learning (CoRL)*, 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,” in *Conference on Robot Learning (CoRL)*. PMLR, 2025, pp. 2143–2154.
- [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,” in *Conference on Robot Learning (CoRL)*, 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,” in *Conference on Robot Learning (CoRL)*, 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] L. Penco, B. Clément, V. Modugno, E. M. Hoffman, G. Nava, D. Pucci, N. G. Tsagarakis, J.-B. Mouret, and S. Ivaldi, “Robust real-time whole-body motion retargeting from human to humanoid,” in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2018, pp. 425–432.
- [22] J. P. Araújo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, “Retargeting matters: General motion retargeting for humanoid motion tracking,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [23] 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.
- [24] 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.
- [25] Ivhaidong, “Lafan1 retargeting dataset,” https://huggingface.co/datasets/ivhaidong/LAFAN1_Retargeting_Dataset, 2025, hugging Face dataset; accessed 2026-04-24.
- [26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [27] M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du, A. Serrano-Munoz, X. Yao, R. Zurbrugg, N. Rudin, *et al.*, “Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning,” *arXiv preprint arXiv:2511.04831*, 2025.
- [28] W. Zeng, K. Yin, X. Niu, S. Lu, W. Zhong, J. Chen, F. Jia, X. Chen, Z. Wang, F. Xu, M. Zhou, K. Li, W. Zhang, H. Wang, L. Yi, D. Lin, J. Pang, and J. Wang, “Scaling behavior foundation model for humanoid robots,” *arXiv preprint arXiv:2607.15163*, 2026.