PHT: Predictive Whole-Body Humanoid Teleoperation

Chengrui Zhu[1,§,*], Liangjun Xing[2,§,*], Lin Li[3], Yong Liu[1,†], and Yinghao Xu[3,†]

***Abstract*— Future-reference conditioning enables humanoid controllers to anticipate whole-body motion, but conventional retargeting cannot supply such references online before the corresponding human observations arrive. We present PHT, a predictive whole-body humanoid teleoperation framework that generates explicit robot-reference horizons from causal, irregularly timed human observations and executes them with a generalist tracker. Its x-prediction conditional-flow retargeter combines real-valued timestamp encoding with temporal-derivative supervision to generate accurate and temporally coherent references. A shared cluster-balanced prior mitigates long-tailed imbalance in both modules, while mastery-aware sampling and a softly routed residual mixture-of-experts enable a single tracker to learn a broad motion repertoire from scratch. Experiments demonstrate accurate future-horizon prediction, the lowest tracking errors among evaluated released policies, and smoother, more trackable references from irregular live inputs. By converting the predicted lookahead into earlier robot responses, PHT achieves high-fidelity, low-latency whole-body tracking in closed-loop teleoperation and supports autonomous loco-manipulation. Code will be publicly available.**

***Index Terms*— Humanoid robots, teleoperation, motion retargeting, whole-body motion tracking, reinforcement learning.**

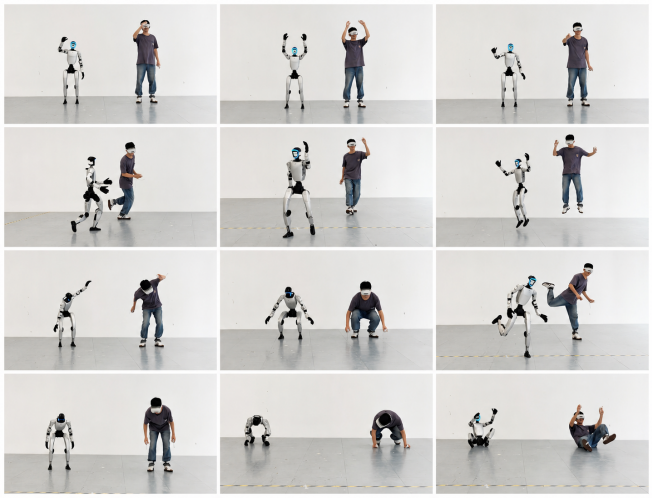


Fig. 1. Whole-body humanoid teleoperation with PHT. Representative operator–robot pose pairs show PHT's versatility across upper-body gestures, locomotion, jumping, squatting, and contact-rich whole-body motions.

I. Introduction

Whole-body teleoperation offers a direct way to command a humanoid across locomotion, expressive behavior, and coordinated loco-manipulation. A common design for learning-based human-to-humanoid systems is a sensing–retargeting–tracking pipeline [1], [2]. A human-motion sensor first captures the operator's body state; a retargeter resolves embodiment differences and generates robot-space reference motion; and a whole-body controller tracks this reference in closed loop. This decomposition lets the retargeter focus on motion fidelity and the tracker on robust physical execution.

Recent work has advanced each stage of this pipeline. H2O couples monocular pose estimation with an RL motion imitator for real-time whole-body teleoperation [3], whereas OmniH2O uses kinematic pose as a universal interface and supports downstream learning from teleoperated demonstrations [4]. TWIST pairs online retargeting with a single controller spanning coordinated skills [1], while TWIST2 integrates portable sensing, holistic retargeting, and general tracking for scalable demonstration collection [2]. GMR improves robot-reference fidelity through whole-body kinematic retargeting [5]; MoReFlow learns unpaired cross-morphology correspondences via flow matching [6]; and NMR learns sequence-level mappings from physics-refined motion pairs to improve trackability [7]. For tracking, ExBody2 distills a privileged teacher into a deployable generalist [8]; GMT distills a soft-MoE teacher trained with performance-based sampling [9]; and Humanoid-GPT distills hundreds of RL experts into a causal Transformer for zero-shot tracking [10]. Departing from this modular pipeline, SONIC uses a shared human–robot latent space to drive a common policy [11]. Recent work also shows that reference lookahead helps controllers prepare for changes in whole-body coordination and dynamics [1], [9], [11], [12].

Such lookahead is readily available from recorded trajectories during offline training but unavailable during live teleoperation, where future human motion has yet to occur. Several methods eliminate the need for lookahead at deployment by distilling future-conditioned teacher policies into student policies without lookahead [1], [4], whereas TeleGate infers an implicit prior over future motion from the history of reference motions [13]. These alternatives alter the deployed policy interface and forgo the tracking benefit afforded by explicit lookahead. Preserving the lookahead-conditioned interface and this benefit requires supplying future references online without delaying execution.

We introduce PHT, a modular whole-body teleoperation system that links an independently trained predictive retargeter and generalist motion tracker through a one-way interface carrying timestamped robot-reference horizons. Whereas conventional retargeting maps observed human motion to contemporaneous robot references, we model a conditional distribution over robot trajectories that extends beyond the latest observation. The extrapolated states supply the tracker

[1]Zhejiang University. [2]Tsinghua University. [3]Robbyant. [§]Work done during internship at Robbyant. [*]Equal contribution. [†]Corresponding authors.

with online lookahead, enabling responsive teleoperation across a broad motion repertoire shown in Fig. 1. This capability also supports collecting high-quality, smooth loco-manipulation demonstrations for training downstream whole-body policies. Through the same horizon interface, the retargeter and downstream policies share a whole-body execution layer, thereby preserving consistent and robust control from teleoperation to autonomous deployment.

We parameterize this trajectory distribution with an x-prediction conditional flow. Real-valued timestamp encoding handles asynchronous human observations and robot queries without resampling, while temporal-derivative supervision promotes coherent trajectories. At the system level, both modules contend with a long-tailed motion distribution, where frequent modes can dominate training and degrade performance on tail motions. We therefore construct a shared cluster-balanced prior over robot-space motion segments to balance base exposure across motion modes in both modules. For tracker training, mastery-aware sampling adapts this prior toward segments with high mastery deficits or insufficient exposure, while a softly routed residual mixture-of-experts (MoE) specializes action means and exploration variances across heterogeneous motion regimes. Together, the sampling prior, motion curriculum, and specialization enable a single generalist tracker to be trained from scratch over the full repertoire, without specialist pretraining or distillation.

The contributions are fourfold:

- We formulate predictive retargeting as conditional robot-trajectory generation from causally available and irregular human observations, covering both contemporaneous retargeting and future extrapolation.
- We introduce a shared cluster-balanced prior that balances long-tail exposure in training processes.
- We train a single generalist whole-body tracker from scratch over the full training repertoire, without specialist pretraining or distillation, using mastery-aware sampling and specialization of control and exploration.
- We validate the retargeter, tracker, and integrated PHT system, and demonstrate the composability in downstream autonomous loco-manipulation tasks.

II. Method

Fig. 2 provides an overview of PHT, which connects two independently trained modules through a robot-reference-horizon interface: a predictive motion retargeter and a generalist whole-body tracker. A shared cluster-balanced prior defines the base motion distribution used to train both modules. From causally available human observations, the retargeter predicts a future robot reference horizon without waiting for the corresponding future measurements, and the tracker executes this horizon in closed loop. Together, PHT preserves the anticipatory benefit of future-conditioned control and enables responsive whole-body humanoid teleoperation.

A. Problem Formulation

We formulate whole-body humanoid teleoperation as two coupled problems: human-to-robot motion retargeting and whole-body motion tracking. We model motion tracking as a partially observable Markov decision process (POMDP) with observation $\mathbf{o}_t \in \mathcal{O}$ derived from the environment state $\mathbf{s}_t \in \mathcal{S}$, and action $\mathbf{a}_t \in \mathcal{A}$. The policy samples $\mathbf{a}_t \sim \pi(\cdot \mid \mathbf{o}_t)$ and receives reward r_{t+1}. Given a discount factor $\gamma \in [0, 1)$, the policy maximizes the expected discounted return

$$\mathcal{J}(\pi) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \right]. \tag{1}$$

For future-reference-conditioned tracking, $\mathbf{o}_t$ includes a robot reference horizon $\mathbf{X}_t$, allowing the policy to prepare for upcoming motion [1], [9], [11]. During policy training, prerecorded trajectories provide this horizon; constructing it by direct retargeting in live teleoperation requires waiting for the corresponding human observations. Let $\mathcal{H}_t = \left\{ (\mathbf{h}_i, \tau_i^h) \mid \tau_i^h \le t \right\}$ denote the causally available human observations, where τ_i^h is the acquisition time of observation $\mathbf{h}_i$. The predictive retargeter models the conditional reference-horizon distribution under the available human observations, $\mathbf{X}_t \sim q_\theta(\cdot \mid \mathcal{H}_t)$. By predicting those targets directly from $\mathcal{H}_t$, PHT supplies online lookahead without delaying execution.

B. Cluster-Balanced Sampling Prior

Large-scale motion corpora often exhibit a long-tailed frequency distribution, and uniform sampling reproduces this imbalance, allowing common modes to dominate the learning process. We instead construct a sampling prior that balances exposure across the motion repertoire while preserving within-mode diversity. We partition each robot-space trajectory into segments and represent each segment with a whole-body motion descriptor spanning joint-space motion statistics, root states, and end-effector kinematics. Applying k-means to the descriptors yields C clusters of segments with similar motion characteristics. For segment b, the prior is defined as

$$p_0(b) = \frac{1}{C\, n_{c(b)}}, \tag{2}$$

where $c(b)$ is the cluster containing segment b and $n_{c(b)}$ is the number of segments in that cluster. Consequently, every cluster receives probability mass $1/C$, distributed uniformly among its segments. The resulting prior provides the base sampling distribution for both retargeter and policy training, reducing the dominance of frequent modes and increasing exposure to underrepresented ones. Cluster identities affect sampling only and are not provided as inputs to either model.

C. Predictive Retargeting

Given a timestamped sequence of human observations and a set of robot query timestamps, the predictive retargeter models a conditional distribution over the corresponding robot trajectories. The human observation at index i is defined as

$$\mathbf{h}_i = \left[\mathbf{R}_B^h \;\; \mathbf{v}_B^h \;\; \boldsymbol{\omega}_B^h \;\; \left\{ {}^B\mathbf{R}_j^h, {}^B\mathbf{p}_j^h \right\}_{j \in \mathcal{J}_h} \right]_i, \tag{3}$$

where B denotes the base frame, $\mathcal{J}_h$ is the set of observed human joints, and $\mathbf{R}_B^h$, $\mathbf{v}_B^h$, and $\boldsymbol{\omega}_B^h$ denote the base orientation, linear velocity, and angular velocity, respectively.

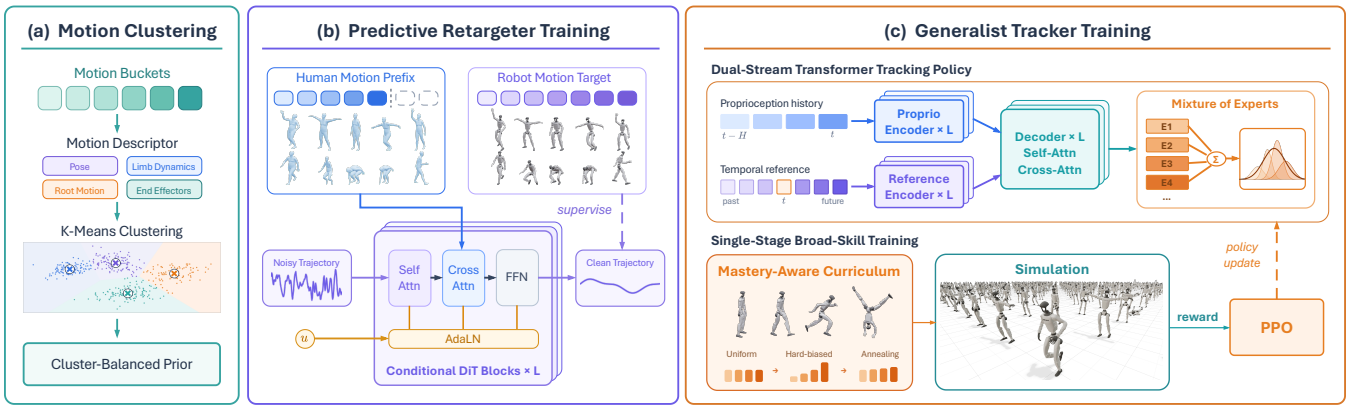


Fig. 2. Overview of PHT. (a) Motion segments are represented and clustered to construct a shared cluster-balanced sampling prior. (b) Predictive retargeting maps causally available human motion to robot trajectories by x-prediction flow matching. (c) A single generalist motion-tracking policy is learned over the full repertoire with a mastery-aware motion curriculum and a MoE mechanism specializing actions and exploration.

For each $j \in \mathcal{J}_h$, ${}^B\mathbf{R}_j^h$ and ${}^B\mathbf{p}_j^h$ denote the rotation and position of joint j expressed in the base frame. The robot reference state at query index k is represented as

$$\mathbf{x}_k = \left[\mathbf{R}_B^r \ \mathbf{v}_B^r \ \boldsymbol{\omega}_B^r \ \mathbf{q}^r\right]_k, \tag{4}$$

where $\mathbf{R}_B^r$ is the 6D representation of the robot root orientation, $\mathbf{v}_B^r$ and $\boldsymbol{\omega}_B^r$ are its linear and angular velocities, and $\mathbf{q}^r$ is the joint configuration. Let $\mathcal{H} = \{(\mathbf{h}_i, \tau_i^h)\}_{i=0}^{N_h-1}$ denote the input human observation sequence and its acquisition times, and let $\mathcal{T}^r = \{\tau_k^r\}_{k=0}^{N_r-1}$ denote the robot query timestamps. The requested robot trajectory is modeled as

$$\mathbf{X} \sim q_\theta\left(\cdot \mid \mathcal{H}, \mathcal{T}^r\right), \quad \mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_{N_r-1}]. \tag{5}$$

Conditional Flow. To parameterize q_θ in (5), we define a conditional flow [14] along a linear noise-to-data path. For a training triplet $(\mathcal{H}, \mathcal{T}^r, \mathbf{X}^\star)$, the path is defined as

$$\mathbf{X}_u = (1-u)\boldsymbol{\epsilon} + u\mathbf{X}^\star, \ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \ u \sim \mathcal{U}(0,1), \tag{6}$$

where $\mathbf{X}^\star$ is the clean target trajectory, $\boldsymbol{\epsilon}$ is a Gaussian noise sequence, and $\mathbf{X}_u$ is the interpolated trajectory at flow time u. The predictor estimates the clean endpoint and calculates the flow velocity along the path:

$$\widehat{\mathbf{X}} = f\left(\mathbf{X}_u, u, \mathcal{H}, \mathcal{T}^r\right), \quad \frac{d\mathbf{X}_u}{du} = \frac{\mathbf{X}^\star - \mathbf{X}_u}{1-u}, \tag{7}$$

Training Objective. Unlike the usual v-prediction of flow matching, we train f with the x-prediction parameterization. For a fixed training triplet, the same target supervises all sampled $(u, \boldsymbol{\epsilon})$, and every forward pass predicts the complete trajectory. To explicitly promote temporal consistency, we additionally supervise changes between consecutive states:

$$\mathcal{L}_{\text{retarget}} = \mathbb{E}_{u,\boldsymbol{\epsilon}}\left[\ell(\widehat{\mathbf{X}}, \mathbf{X}^\star) + \boldsymbol{\lambda}_\Delta^\top \ell\left(\Delta\widehat{\mathbf{X}}, \Delta\mathbf{X}^\star\right)\right], \tag{8}$$

where ℓ denotes mean-squared error, Δ computes temporal derivatives, and $\boldsymbol{\lambda}_\Delta$ denotes their corresponding weights.

DiT with Real-Valued Timestamp Encoding. We implement f with a diffusion Transformer (DiT) [15]. Each block applies self-attention to robot tokens derived from $\mathbf{X}_u$ and cross-attention to human tokens embedded from $\mathcal{H}$, while adaptive layer normalization conditions the block on flow time u. Instead of resampling the two streams onto the same timeline, we evaluate rotary position embeddings (RoPE) [16] at the real-valued human timestamps τ_i^h and robot query timestamps τ_k^r. This places both asynchronous streams on a shared physical-time axis, allowing the model to directly accommodate dual asynchronous temporal streams. During training, we synthesize irregular human observation streams by randomizing sampling rate and phase and applying timestamp jitter, observation dropout, and causal cutoffs. Robot states queried beyond cutoffs serve as future predictions.

D. Generalist Motion Tracking

A generalist motion tracker must maintain tracking fidelity across motions with substantially different patterns and dynamics. Such a policy is trained from scratch over the full repertoire, without specialist pretraining or distillation.

Environment Design. We adopt the asymmetric actor–critic formulation [17]. At step t, the observation concatenates a proprioceptive-history and a reference-motion window

$$\mathbf{o}_t = \left[\mathbf{p}_{t-H_o+1} \cdots \mathbf{p}_t \ \ \mathbf{c}_{t,-H_p+1} \cdots \mathbf{c}_{t,H_f}\right], \tag{9}$$

$$\mathbf{p}_t = \left[\boldsymbol{\omega}_t \ \mathbf{g}_t \ \mathbf{q}_t \ \dot{\mathbf{q}}_t \ \mathbf{a}_{t-1}\right], \ \mathbf{c}_{t,k} = \left[\mathbf{q}_t^r \ \boldsymbol{\rho}_t^{\text{rel}} \ \mathbf{v}_t^r \ \boldsymbol{\omega}_t^r\right]_k, \tag{10}$$

with lengths H_o and $H_p + H_f$, respectively. Here, $\boldsymbol{\omega}_t$ and $\mathbf{g}_t$ denote the base angular velocity and projected gravity; $\mathbf{q}_t$, $\dot{\mathbf{q}}_t$, and $\mathbf{a}_{t-1}$ are the joint positions, joint velocities, and previous action. At relative index k, $\mathbf{q}_t^r$ is the target joint configuration, $\boldsymbol{\rho}_t^{\text{rel}}$ is the relative torso orientation, and $\mathbf{v}_t^r$ and $\boldsymbol{\omega}_t^r$ are the target linear and angular velocities. The critic additionally receives ground-truth base linear velocity, anchor-relative tracked-body poses, and torso displacement error. The action is mapped to joint-position targets, which are tracked by a low-level proportional–derivative (PD) controller. The reward combines root and whole-body pose and velocity tracking with regularization on action rate, joint-limit violations, and self-collisions. During training, we employ a domain-randomization curriculum that transitions from a narrow to a broader randomization range following Wang et al. [18].

Mastery-Aware Repertoire Sampling. Although the cluster-balanced prior p_0 promotes broad repertoire coverage, it does

not reflect the tracker's evolving competence. We therefore adapt the rollout distribution using recent closed-loop tracking performance. Each rollout window for segment b yields a deficit based on how far the normalized tracking errors rise above the minima attained within that window. This relative measure captures deterioration in tracking quality rather than absolute error alone. We smooth these signals across repeated visits to obtain a bounded mastery-deficit score $D_{b,n}$ at curriculum update n. To account for uncertainty in sparsely visited segments, an under-exposure score $L_{b,n}$ compares the segment's discounted visitation with its expected visitation under the base prior. The motion segments are sampled according to

$$p_n(b) \propto p_0(b) \left[D_{b,n}^2 + \lambda_L L_{b,n} + \varepsilon \right], \tag{11}$$

where $p_0(b)$ supplies the cluster-balanced base weighting, $D_{b,n}^2$ accentuates differences in mastery deficit, λ_L weights the under-exposure correction, and $\varepsilon > 0$ preserves full support over the repertoire.

Actor–Critic Architecture with Adaptive Specialization. The actor and critic use independent dual-stream Transformers. In each network, the proprioceptive-history and reference-window streams are encoded separately and then fused with cross-attention. The critic maps its fused representation to a scalar value estimate with a linear layer. To accommodate heterogeneous motion regimes within a single actor, we introduce a residual MoE design for adaptive action-distribution specialization. A state-dependent router computes softmax weights over expert heads, each applying residual adjustments to a shared action mean and exploration variance. All heads are evaluated and softly aggregated into a single diagonal Gaussian via per-action-dimension moment matching. The resulting policy retains standard Gaussian log-probability and entropy calculations, while soft routing avoids the likelihood discontinuities associated with hard expert switching.

Regularized Policy Optimization. We optimize the actor and critic using proximal policy optimization (PPO) [19], augmented by symmetry and routing regularization:

$$\mathcal{L}_{\text{track}} = \mathcal{L}_{\text{PPO}} + \lambda_{\text{sym}}\mathcal{L}_{\text{sym}} + \lambda_{\text{bal}}\mathcal{L}_{\text{bal}} + \lambda_{\text{shp}}\mathcal{L}_{\text{shp}}, \tag{12}$$

$$\mathcal{L}_{\text{bal}} = D_{\text{KL}}(\mathcal{U}_E \,\|\, \mathbb{E}[\mathbf{w}]), \quad \mathcal{L}_{\text{shp}} = \frac{\mathbb{E}[\text{H}(\mathbf{w})]}{\log E}, \tag{13}$$

where $\mathcal{L}_{\text{PPO}}$ is the standard PPO loss, $\mathcal{L}_{\text{sym}}$ is the mirror symmetry loss [20], $\mathcal{L}_{\text{bal}}$ and $\mathcal{L}_{\text{shp}}$ are the routing load-balancing and sharpness losses, and λ_{sym}, λ_{bal}, and λ_{shp} are their weights, respectively; $\mathbf{w}$ denotes the routing weights over the E expert heads, $\mathcal{U}_E$ is the uniform distribution over these heads, and H denotes categorical entropy. The load-balancing loss promotes balanced utilization across inputs, whereas the sharpness loss encourages selective routing for individual inputs. Together, they support an emergent division of labor across motion regimes without predefined expert roles.

III. Experiments

We first describe the training and deployment setup. We then address the following research questions:

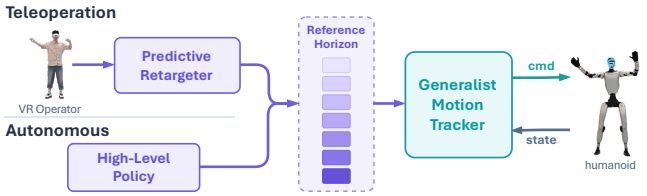


Fig. 3. Deployment pipeline. The predictive retargeter or a high-level policy supplies reference horizons to the same generalist motion tracker.

- How are the training motions distributed, and why is a cluster-balanced prior important for learning?
- Can predictive retargeting generate accurate and temporally coherent robot-reference horizons from causally available human observations?
- Can a generalist tracker accurately track a broad motion repertoire, and how do key design components contribute to its performance?
- How does predictive retargeting affect closed-loop responsiveness, motion fidelity, and reference trackability?
- Can upstream policies reuse the same reference horizon interface for autonomous loco-manipulation?

A. Implementation Details

Training. We use 66,340 flat-ground motions from BONES-SEED [21] together with their mirrored counterparts, yielding 132,680 training trajectories and approximately 270 trajectory-hours, and retarget the original SOMA motions to the SMPL kinematic skeleton [22] as the human representation. The retargeter is composed of 8 DiT blocks of width 768 with 8 heads and trained for 15 epochs, with $\boldsymbol{\lambda}_\Delta$ set to 10^{-3} for root rotational velocity and 10^{-4} for joint velocity. The tracker uses 2 layers of 256-dimensional, 4-head Transformers for each encoder and decoder, with $E = 8$ expert heads, $H_o = 10$ proprioceptive-history frames, $H_p = 10$ past-to-current reference frames, $H_f = 10$ future reference frames, introducing a lookahead of 200 ms. For its mastery-aware sampler, $\lambda_L = 0.1$ and $\varepsilon = 10^{-3}$; for policy regularization, $\lambda_{\text{sym}} = 0.1$, $\lambda_{\text{bal}} = 0.01$, and $\lambda_{\text{shp}} = 0.005$. It is trained for 15,000 PPO iterations in an mjlab environment [23], with a total of 2^{18} parallel environments and 48-step rollouts. The control frequency of the tracking policy is 50 Hz.

Deployment. The deployment pipeline is illustrated in Fig. 3. The retargeter receives live human-motion observations with acquisition timestamps from a PICO 4 Ultra at approximately 90 Hz and uses a 2-step ($J = 2$) Euler sampler to generate reference horizons for the tracking policy. The initial noise for each queried timestep is fixed to avoid reference discontinuities caused by stochastic resampling. Robot query timestamps are defined on the tracker's 50-Hz control timeline, decoupling policy execution from the upstream observation. Both networks run on an NVIDIA Jetson Orin NX onboard the G1. Policy inference and retargeter sampling take 1.15 ms and 12.53 ms, respectively, keeping their combined inference time within the 20-ms control period. To avoid delaying action execution, each cycle first runs the policy on the latest available horizon and issues the current action; the retargeter

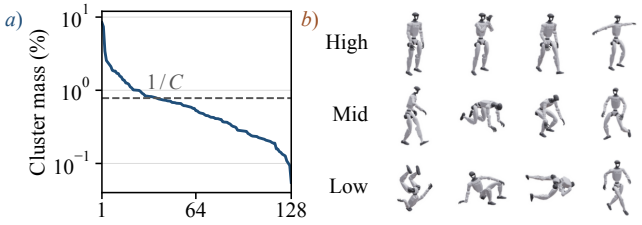


Fig. 4. Motion-cluster analysis. *a*) Ranked empirical cluster masses; the dashed line denotes the balanced target $1/C$. *b*) Snapshots of four distinct clusters from each of the high-, mid-, and low-frequency regions.

TABLE I

RETARGETING ERROR ABLATION.

Method	MPKPE (cm)	ROE (10^{-3}rad)	RLVE (m/s)	RAVE (rad/s)	MPJPE (10^{-3}rad)	MPJVE (rad/s)
w/o flow	0.324	8.2	0.028	0.109	9.53	0.136
v-pred. ($J = 2$)	0.874	20.4	0.078	0.318	19.76	1.053
v-pred. ($J = 8$)	0.327	9.3	0.029	0.114	9.80	0.194
Ours ($J = 2$)	0.321	7.9	**0.027**	0.106	9.46	0.135
Ours ($J = 8$)	**0.316**	**7.7**	**0.027**	**0.104**	**9.37**	**0.133**
w/o irr. timing	0.401	10.2	0.035	0.132	11.50	0.162
w/o deriv. loss	0.327	8.7	0.028	0.113	9.43	0.143
w/o bal. prior	0.334	8.2	0.030	0.113	9.90	0.140

then generates the horizon for the next cycle. With a fully populated horizon, the nominal observation-to-execution delay is two control cycles (40 ms): one to update the horizon and one to execute the resulting action.

B. Motion Cluster Analysis

We divide the robot-space training trajectories into 1-s segments and assign the resulting 1,005,194 segments to $C = 128$ clusters. Cluster sizes range from 554 to 84,522 (median 5,273), yielding a 153× max-to-min ratio. The 13 least-populated clusters contain only 1.73% of the segments; the balanced prior instead assigns them $13/128 = 10.16\%$ of the sampling mass, increasing their collective exposure by 5.87×. Fig. 4 shows the ranked distribution together with snapshots from distinct clusters spanning high-, mid-, and low-frequency regions of the distribution. The snapshots illustrate the breadth of body configurations captured by the partition rather than predefined semantic action classes.

C. Retargeter Evaluation

Metric Definitions. We use the following metrics to evaluate the design choices of the predictive retargeter. Mean per-keypoint position error (MPKPE) measures root-relative error over keypoints, while root orientation error (ROE) is the geodesic distance between root orientations on $\mathrm{SO}(3)$. Root linear- and angular-velocity errors (RLVE and RAVE) use Euclidean distance; mean per-joint position and velocity errors (MPJPE and MPJVE) use mean absolute joint angle and velocity errors, respectively. We sample 8,192 trajectories for testing and use others for training.

Results. Table I evaluates the retargeter components and flow parameterizations, with the w/o flow baseline directly regressing the queried trajectory from the same causal inputs;

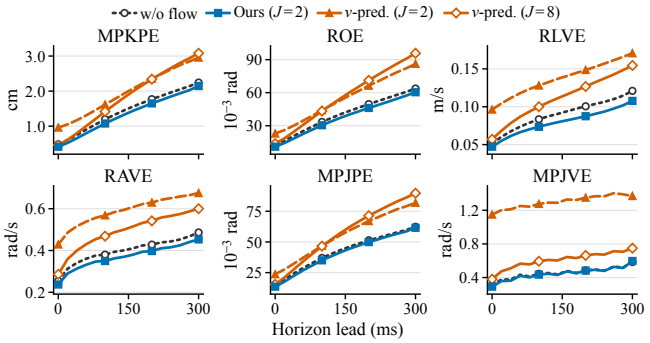


Fig. 5. Comparison of future prediction errors across horizon leads.

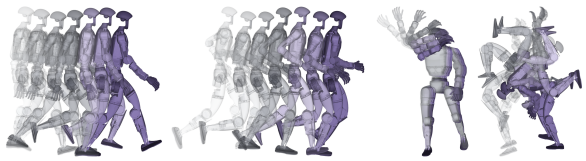


Fig. 6. Representative predictive-retargeting trajectories for walking, running, waving, and flipping motions (left to right), with the observation-supported interval shown in gray and the predicted horizon in violet.

J denotes the number of Euler updates. Overall, the x-prediction flow provides the best accuracy–efficiency trade-off: its 2-update variant outperforms the deterministic regressor on all 6 metrics and remains within 2.3% of the 8-update variant, which achieves the best or tied-best value in every column. By contrast, reducing J from 8 to 2 incurs a substantially larger accuracy loss for v-prediction. Among the ablations, retaining native irregular timestamps has the largest effect, reducing errors by 16.5–22.9% relative to synchronized 50-Hz inputs. Derivative consistency further improves five metrics, including reductions of 5.4–9.2% in ROE, RAVE, and MPJVE, while leaving MPJPE essentially unchanged. The cluster-balanced prior complements these improvements by reducing every error by 3.6–8.8%. Fig. 5 further shows that our method reduces all future-prediction errors by 1.6–11.5% relative to w/o flow. These results identify the proposed model as the best accuracy–efficiency trade-off among the evaluated instantiations of predictive retargeting. Fig. 6 demonstrates coherent extrapolation across diverse motions.

D. Tracker Evaluation

Tracking Performance. We evaluate the behavioral breadth and tracking fidelity of our generalist tracker independently of the retargeter through simulation and real-world experiments. As shown in Fig. 7, a single policy executes locomotion, dancing, and highly dynamic acrobatic skills with stable and coordinated whole-body control. Table II further compares closed-loop tracking fidelity with released policies on LAFAN1, AMASS, and SEED, where BPE is the world-frame position error of the torso link, while A-MPKPE and A-MPBOE are the heading-aligned link-position and orientation errors. Our tracker achieves the lowest error across every dataset and metric, outperforming the competing policy by 24.7–74.7% on LAFAN1, 6.0–34.5% on AMASS, 27.3–

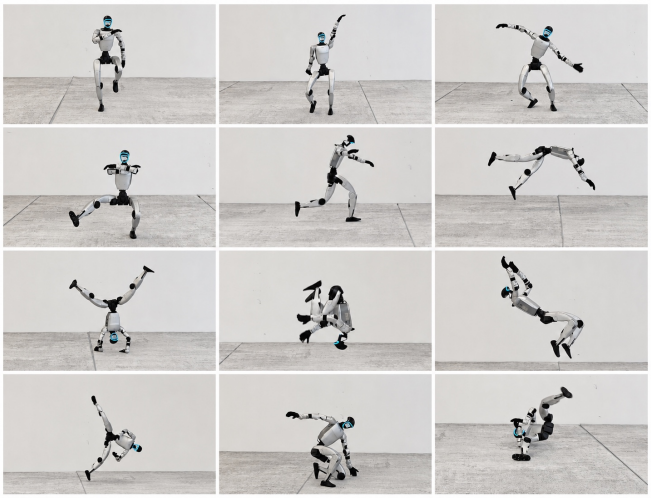


Fig. 7. Dynamic motions executed on the G1 by a single generalist tracking policy, including running, kicking, dancing, jumping, cartwheels, and flips.

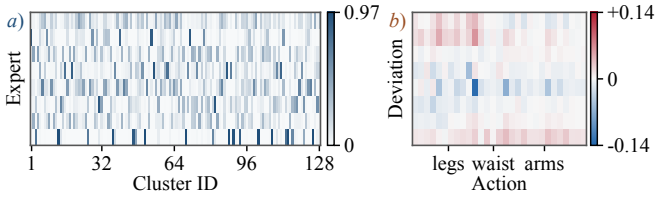


Fig. 8. Expert routing and exploration specialization. *a*) Routing weights by motion-segment cluster. *b*) Per-expert deviations of the exploration scale from the across-expert mean for each action dimension.

52.7% on SEED, and 19.4–59.9% on SEED Hard, a subset of SEED comprising challenging motion trajectories such as flips, jumping, and cartwheels. The consistent advantage on LAFAN1 and AMASS, both excluded from training, indicates strong cross-dataset generalization of our tracker.

Ablation Analysis. While Table II benchmarks released policies, Table III isolates the contributions of our tracker components. The lookahead comparison validates the role of future references in providing the policy with the anticipatory information needed to coordinate upcoming pose changes and contact transitions. The cluster-balanced prior maintains similar performance on SEED while improving every metric on SEED Hard, indicating better coverage of challenging motions. The motion curriculum improves both evaluation sets especially on the harder set, by prioritizing motions that have not yet been sufficiently mastered. Finally, the MoE improves most tracking metrics, showing that motion-dependent specialization facilitates learning across a diverse repertoire. Fig. 8 shows that the MoE develops pronounced preferences across motion clusters and action dimensions. The normalized mutual information between motion cluster and expert assignment reaches 0.441, although cluster identities are not provided to the router. These results support a meaningful division of labor across motion regimes.

E. Teleoperation Evaluation

Reference Trackability and Execution Regularity. We evaluate the benefits of PHT on a real PICO whole-body motion stream. The evaluated stream consists of 4 segments totaling 373 s and spans diverse motions, including locomotion, upper-body gestures, and contact-rich motions. It averages 89.9 Hz but has irregular inter-arrival intervals, typically 7.8–14.4 ms with occasional gaps up to 52.4 ms. It also exhibits realistic sensing imperfections and irregular drift in the estimated global root velocity. We compare PHT with GMR [5], providing both with the PICO stream and tracking both resulting reference streams with our tracker. We therefore assess each reference stream by how accurately and smoothly it can be executed by the shared controller. As shown in Table IV, PHT yields lower joint-reference jerk and less reference-foot drift than GMR. This improved reference quality translates into lower time-aligned MPJPE, smoother joint actions, and less robot-foot motion under the shared controller. These gains can be attributed to our

TABLE II

TRACKING BENCHMARK OF RELEASED POLICIES.

Method	Train ?‡	MPJPE (rad)	MPJVE (rad/s)	RLVE (m/s)	BPE (m)	A-MPKPE (m)	A-MPBOE (rad)
LAFAN1 [24]: 40 trajectories (2.45 h)							
GMT†	✓	0.225	1.049	0.743	5.601	0.374	1.200
TWIST2†	×	0.284	1.133	0.731	4.793	0.402	1.273
SONIC†	×	0.244	0.946	0.414	2.488	0.174	0.678
HoloMotion†	✓	0.108	0.639	0.317	2.046	0.149	0.509
Humanoid-GPT†	✓	0.162	1.149	0.572	3.143	0.196	0.701
Ours	×	**0.081**	**0.480**	**0.190**	**1.087**	**0.038**	**0.132**
AMASS [25]: 4,096 trajectories (13.96 h)							
GMT	✓	0.136	0.752	0.410	0.959	0.138	0.479
TWIST2	✓	0.132	0.834	0.349	0.659	0.104	0.372
SONIC	×	0.148	0.606	0.232	0.289	0.075	0.297
HoloMotion	✓	0.099	0.452	0.182	0.262	0.077	0.248
Humanoid-GPT	✓	0.098	0.715	0.280	0.365	0.078	0.283
Ours	×	**0.083**	**0.385**	**0.171**	**0.211**	**0.049**	**0.171**
SEED: 4,096 paired (original & mirrored) trajectories (21.44 h)							
GMT	×	0.138	0.660	0.282	0.569	0.096	0.434
TWIST2	×	0.145	0.864	0.286	0.529	0.093	0.417
SONIC	✓	0.124	0.490	0.143	0.159	0.048	0.230
HoloMotion	×	0.080	0.377	0.118	0.218	0.048	0.190
Humanoid-GPT	×	0.111	0.838	0.253	0.428	0.078	0.329
Ours	✓	**0.054**	**0.265**	**0.086**	**0.075**	**0.025**	**0.090**
SEED Hard: 128 paired trajectories (0.28 h)							
GMT	×	0.186	1.910	1.013	0.901	0.210	0.672
TWIST2	×	0.222	2.377	1.127	1.009	0.208	0.689
SONIC	✓	0.205	1.623	0.610	0.482	0.125	0.459
HoloMotion	×	0.120	1.257	0.621	0.566	0.109	0.314
Humanoid-GPT	×	0.193	2.211	1.014	0.899	0.177	0.612
Ours	✓	**0.097**	**0.912**	**0.337**	**0.193**	**0.060**	**0.165**

† We evaluate the public checkpoints of SONIC [11], GMT [9], TWIST2 [2], Humanoid-GPT [10], and HoloMotion v1.4 [12], while retaining their native interfaces. All evaluated policies complete every trajectory without early termination.
‡ "Train?" indicates whether each method reports using the dataset for training.

TABLE III

TRACKING ERROR DEVIATION OF ABLATION SETTINGS (%).

Method	MPJPE	MPJVE	RLVE	BPE	A-MPKPE	A-MPBOE
SEED: 4,096 paired trajectories						
w/o lookahead	+18.5	+17.0	+10.7	+14.7	+10.4	+19.6
w/o bal. prior	+0.0	+1.5	−0.7	−2.2	−3.3	−0.2
w/o MoE	+3.4	+1.4	+1.1	+7.2	+3.2	+3.3
w/o curriculum	+3.8	+5.6	+6.5	+6.5	+2.2	+3.9
SEED Hard: 128 paired trajectories						
w/o lookahead	+18.6	+12.1	−4.3	−4.2	+2.4	+17.9
w/o bal. prior	+3.3	+3.8	+7.5	+16.4	+2.2	+3.7
w/o MoE	+4.0	+1.7	−1.6	+2.6	−0.9	+1.5
w/o curriculum	+14.6	+11.4	+24.2	+50.6	+19.4	+18.9

Percentage changes relative to the full model; positive values denotes higher errors.

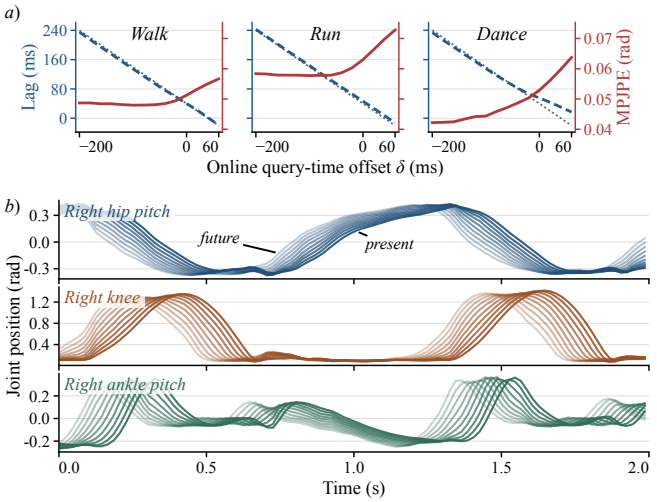


Fig. 9. Closed-loop analysis of predictive retargeting. a) Robot-response lag (blue) and time-aligned MPJPE (red) for walking, running, and dancing; dotted lines denote the nominal lag. b) Joint references at ten 20-ms-spaced prediction leads during walking; lighter curves indicate longer lookahead.

TABLE IV

TELEOPERATION COMPARISON.

Metric	GMR	Ours	Δ
Mean absolute joint reference jerk (10^3 rad/s^3)	1.57	**1.11**	−29.5%
Mean absolute reference foot drift speed (m/s)	0.277	**0.190**	−31.4%
Time-aligned MPJPE (rad)	0.0719	**0.0622**	−13.5%
Mean absolute joint action jerk (10^3 rad/s^3)	1.88	**1.09**	−42.1%
Mean absolute robot foot slip speed (m/s)	0.0810	**0.0696**	−14.0%

timestamp-conditioned neural retargeter, whose trajectory-level prediction converts the irregular PICO stream into smoother and more trackable robot references.

Closed-Loop Response. We characterize the closed-loop teleoperation system along two dimensions: temporal responsiveness and motion fidelity, measured by response lag and time-aligned MPJPE, respectively. Response lag is the temporal shift that minimizes joint-space MPJPE between the reference and feedback trajectories, and time-aligned MPJPE is the corresponding residual error. We control this trade-off with an online query-time offset δ, which advances the timestamps requested from the predictive retargeter. Increasing δ therefore requires longer-horizon prediction while advancing the commanded motion. With the tracker's 200-ms future window, the maximum prediction lead is $200\,\text{ms} + \delta$: $\delta = -200$ ms yields a fully observed horizon, whereas $\delta = 40$ ms compensates the nominal observation-to-execution delay. We sweep δ from -200 to 60 ms for 3 representative sequences of walking, running, and dancing. Fig. 9(a) shows that increasing δ consistently advances the robot response, with a gradual fidelity cost as the query time extends farther into the future. Increasing δ from 0 to 40 ms reduces the average response lag from 47.3 to 13.3 ms, while time-aligned MPJPE increases moderately from 0.0557 to 0.0615 rad. Further advancing the reference continues to reduce response lag, but with diminishing returns and an increasingly steep rise in tracking errors. The lag reduction is most pronounced for temporally regular sequences, whereas the more dynamic dancing sequence retains residual lag because its future evolution is harder to predict reliably. Fig. 9(b) shows that the joint references evolve smoothly across multiple lookahead steps within the predicted horizon, illustrating the coherent multi-step prediction that enables this responsiveness–fidelity trade-off. Together, these results show that predicted lookahead substantially reduces closed-loop response lag at a controllable cost in tracking fidelity.

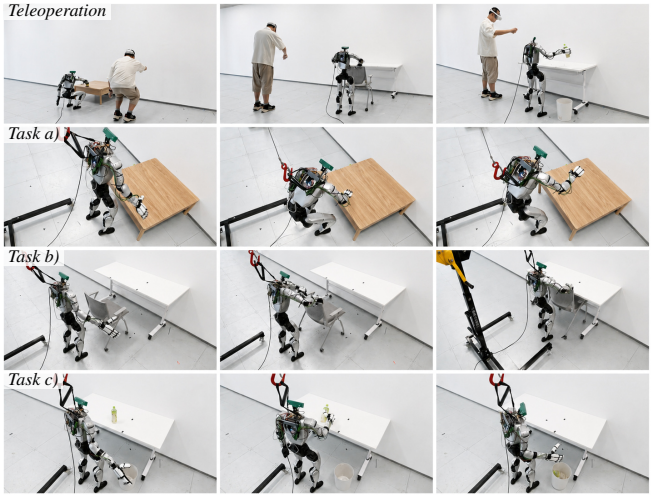


Fig. 10. Representative teleoperated and autonomous loco-manipulation executions. The top row shows teleoperation for a) bread pickup, b) chair pushing, and c) bottle throwing, and the following rows show their autonomous executions, respectively. The overhead tether serves only as a passive safety device and provides no support or control assistance.

F. Loco-Manipulation Evaluation

Experiment Setup. We conduct the evaluation on a G1 humanoid equipped with two actuated head DoFs (pitch and yaw) and a pair of ROBOTERA XHAND 1 dexterous hands. During teleoperated demonstration collection, the retargeter produces body-reference chunks, while the PICO 4 Ultra headset orientation and controller inputs are mapped to head and hand targets, respectively. We evaluate three whole-body loco-manipulation tasks: a) squatting to pick up bread and standing up; b) holding a chair by its backrest and pushing it next to a table; and c) grasping a bottle and throwing it into a trash bin. Together, these tasks span lower-body posture transitions, locomotion, and object manipulation. For each task, we collect 50 demonstrations with PHT and fine-tune a separate $\pi_{0.5}$ vision–language–action (VLA) policy on the resulting trajectories [26]. At deployment, the policy predicts action chunks whose body-reference states are supplied to the tracker as reference horizons, while the head and hand targets are issued through separate control channels.

Results. As shown in Fig. 10, the learned policies produce coordinated whole-body behaviors across all three tasks. In task a), the robot couples a deep squat with reaching and grasping, and returns to an upright posture. In task b), it grasps the chair backrest and moves forward to push the chair toward the table. In task c), it grasps and throws a bottle into the trash bin. The policies achieve success rates of $6/10$, $9/10$, and $8/10$ on tasks a), b), and c), respectively.

No trial, successful or not, involved a fall or control failure. These results demonstrate that autonomous policies can drive diverse loco-manipulation tasks with our tracker.

IV. Conclusion and Discussion

We presented PHT, a modular whole-body teleoperation framework that connects a predictive retargeter and a generalist tracker through explicit robot-reference horizons. The retargeter transforms causally available, irregularly timed human observations into coherent present and future robot references, providing online lookahead without waiting for future measurements. The tracker executes these horizons with high fidelity across a broad motion repertoire, generalizes to motion datasets excluded from training, and supports stable whole-body behaviors on the G1. Together, the predictive retargeter and generalist tracker provide a unified execution layer for teleoperation and autonomous loco-manipulation.

The robot-reference horizon decouples motion generation from physical execution, allowing the predictive retargeter and downstream policies to share the same whole-body tracker. Predictive references improve responsiveness, but ambiguous or highly dynamic motions can produce inaccurate futures, leaving residual delay or increasing tracking error. Independent module training may also introduce a distribution gap between the offline motions used to train the tracker and the generated horizons encountered during deployment. Incorporating generated references and their uncertainty during training could further improve deployment robustness.

References

[1] 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 *Proceedings of the 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 305. PMLR, 2025, pp. 2143–2154. [Online]. Available: https://proceedings.mlr.press/v305/ze25a.html

[2] 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. [Online]. Available: https://arxiv.org/abs/2511.02832

[3] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024. [Online]. Available: https://arxiv.org/abs/2403.04436

[4] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi, "OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *Proceedings of the 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2025, pp. 1516–1540. [Online]. Available: https://proceedings.mlr.press/v270/he25b.html

[5] J. P. Araújo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025. [Online]. Available: https://arxiv.org/abs/2510.02252

[6] W. Kim, T. Li, and S. Ha, "MoReFlow: Motion retargeting learning through unsupervised flow matching," *arXiv preprint arXiv:2509.25600*, 2025. [Online]. Available: https://arxiv.org/abs/2509.25600

[7] Q. Zhao, K. Yang, X. Wang, S. Zhao, Y. Lu, X. Zhang, W. Yin, Q. Shen, X.-X. Long, and X. Cao, "Make tracking easy: Neural motion retargeting for humanoid whole-body control," *arXiv preprint arXiv:2603.22201*, 2026. [Online]. Available: https://arxiv.org/abs/2603.22201

[8] 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. [Online]. Available: https://arxiv.org/abs/2412.13196

[9] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang, "GMT: General motion tracking for humanoid whole-body control," *arXiv preprint arXiv:2506.14770*, 2025. [Online]. Available: https://arxiv.org/abs/2506.14770

[10] Z. Qi, X. Chen, J. Wang, C. Lin, Y. Lian, W. Zhang, X. Yu, H. Wang, and L. Yi, "Humanoid generative pre-training for zero-shot motion tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2026, pp. 20 834–20 844.

[11] Z. Luo, Y. Yuan, T. Wang, C. Li, F. Castañeda, S. Chen, Z.-A. Cao, J. Li, D. Minor, Q. Ben, J. Park, D. Sami, Z. Wang, X. Da, R. Ding, C. Hogg, L. Song, E. Lim, E. Jeong, T. He, H. Xue, W. Xiao, S. Yuen, J. Kautz, Y. Chang, U. Iqbal, L. J. Fan, and Y. Zhu, "SONIC: Supersizing motion tracking for natural humanoid whole-body control," *Science Robotics*, vol. 11, no. 117, p. eaed4592, 2026. [Online]. Available: https://www.science.org/doi/10.1126/scirobotics.aed4592

[12] M. Chen, K. Wang, B. Zhang, X. Ma, Z. Yang, Y. Ren, Q. Huang, Z. Zhu, Y. Wang, and Z. Su, "HoloMotion-1 technical report," *arXiv preprint arXiv:2605.15336*, 2026. [Online]. Available: https://arxiv.org/abs/2605.15336

[13] J. Li, B. Tang, F. Wu, and R. Cao, "TeleGate: Whole-body humanoid teleoperation via gated expert selection with motion prior," *arXiv preprint arXiv:2602.09628*, 2026. [Online]. Available: https://arxiv.org/abs/2602.09628

[14] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=PqvMRDCJT9t

[15] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[16] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[17] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel, "Asymmetric actor critic for image-based robot learning," in *Proceedings of Robotics: Science and Systems*, June 2018.

[18] Y. Wang, S. Zhu, P. Zhi, Y. Li, J. Li, Y.-L. Li, Y. Xiao, X. Wang, B. Jia, and S. Huang, "OmniXtreme: Breaking the generality barrier in high-dynamic humanoid control," *arXiv preprint arXiv:2602.23843*, 2026. [Online]. Available: https://arxiv.org/abs/2602.23843

[19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[20] W. Yu, G. Turk, and C. K. Liu, "Learning symmetric and low-energy locomotion," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 144:1–144:12, August 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201397

[21] Bones Studio, "BONES-SEED: Skeletal everyday embodiment dataset," https://bones.studio/datasets/seed, 2026, motion data by Bones Studio.

[22] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[23] K. Zakka, Q. Liao, B. Yi, L. Le Lay, K. Sreenath, and P. Abbeel, "mjlab: A lightweight framework for gpu-accelerated robot learning," *arXiv preprint arXiv:2601.22074*, 2026. [Online]. Available: https://arxiv.org/abs/2601.22074

[24] F. G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. Pal, "Robust motion in-betweening," *ACM Transactions on Graphics*, vol. 39, no. 4, 2020.

[25] 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.

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