LEARNING BEYOND BODIES: JOINT LEARNING AND TRANSFER FOR ROBUST HUMANOID MOTION TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Humanoid motion trackers often use robot-specific motion references, raising the question of how to share motion knowledge across bodies and reuse it when adapting to a new embodiment with limited motion data. We study a common human-motion reference interface that separates shared motion conditioning from embodiment-specific control. A shared encoder takes body-centric human keypoints as input, while robot-specific heads combine its features with proprioception and reference orientation. Retargeted trajectories remain in use for training supervision. On our primary motion benchmark, this input yields average tracking success rate and position accuracy comparable to robot joint-reference conditioning. A G1 comparison additionally shows lower source-position error near screened reference discontinuities. The common input enables joint reinforcement learning across heterogeneous humanoids, improving primary-benchmark position tracking over corresponding single-robot runs. For adaptation, an embodiment excluded from pretraining reuses the encoder and learns a new control head. In a low-data setting, freezing the transferred encoder improves position tracking on an external motion dataset relative to training from scratch. Deployment on two physical humanoids further demonstrates tracking with the same human-keypoint reference. These results support a common reference interface for joint learning and the transfer of motion knowledge to new embodiments.

1 INTRODUCTION

General-purpose humanoid robots must perform diverse tasks reliably in real-world environments. A common design pattern connects task understanding and action generation to low-level physical execution (NVIDIA, 2025; Luo et al., 2026; Ze et al., 2026). Figure 1 summarizes these functional roles as System 2–1–0 (Figure AI, 2026), although module boundaries vary across architectures. Whole-body control (System 0) provides the motor foundation: its tracking accuracy and balance regulation underpin the precise coordination required for locomotion and manipulation. Tracking human motion offers a scalable way to learn this capability (Liao et al., 2025; Luo et al., 2026).

Alongside advances in high-level models and large-scale tracking, cross-robot interface reuse remains a challenge. Robot-specific references define different motion-command spaces across bodies (Figure 1(a)). Adapting System 1 may then require remapping action outputs and demonstration targets, adding data-processing and training effort. Motion-reference features are also relearned when trackers are trained separately (Yang et al., 2026). Joint references can further carry retargeting discontinuities along continuous human motion (Araujo et al., 2025; Zhao et al., 2026). Building on cross-embodiment sharing and tracker adaptation (Zhang et al., 2026; Yang et al., 2026), we ask: *can a shared human-reference encoder support accurate tracking, joint learning across bodies, and reuse by a new embodiment's controller?*

We study this question using body-centric human keypoints as the common interface (Figure 1(b)). Joint reinforcement learning trains a shared reference encoder across robots, while separate control heads learn to execute the motion under each robot's dynamics. Adaptation pairs this encoder with a newly trained target control head; we evaluate feature reuse beyond the source bodies by freezing the encoder. Retargeted trajectories still provide supervision, initialization, and reference orientation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

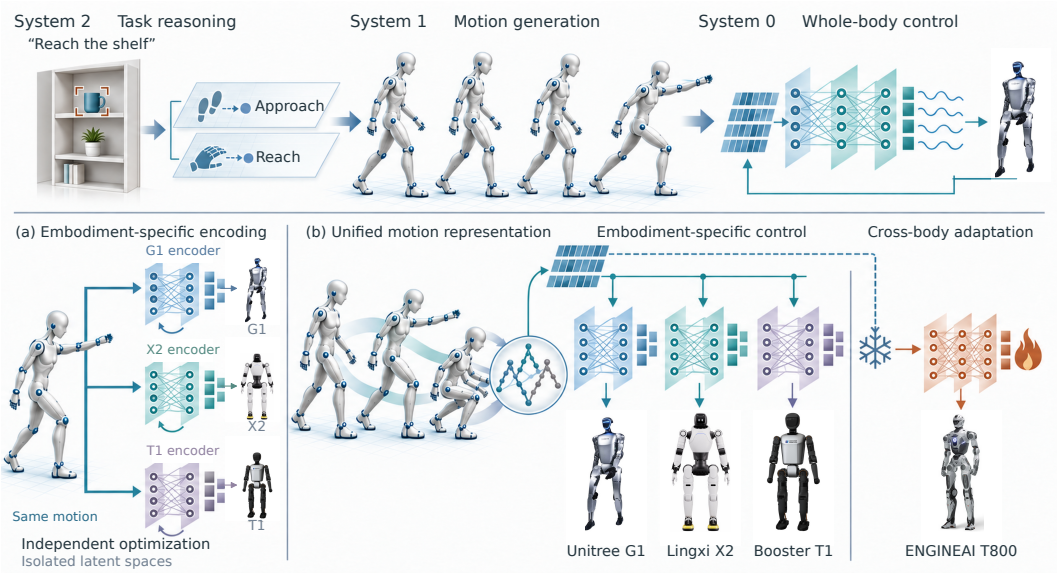

Figure 1: A shared motion-reference interface across bodies. (a) Robot-specific references from the same human motion train separate encoders, repeating feature learning. (b) Human keypoints support a shared representation, joint learning across bodies, and new-embodiment adaptation. Matching colors identify the same embodiment. The snowflake and flame denote a frozen encoder and a newly trained target head. Poses and networks are conceptual. G1, X2, T1 (23 DoF), and T800 product images illustrate embodiments, not experimental source–target assignments; upstream modules are contextual. Retargeting supplies supervision, initialization, and reference orientation.

Our contributions connect the reference interface to joint learning and transfer. **First**, human-keypoint conditioning achieves comparable average success rate and position accuracy to robot joint-reference conditioning on the primary benchmark, with lower source-position error near screened reference jumps in the G1 comparison. **Second**, joint training of the reference encoder across four humanoids reduces primary-benchmark position error by 3.3–4.6% relative to corresponding solo runs. **Third**, on two embodiments excluded from pretraining, a frozen encoder yields the highest or tied-highest success rate across eight target–regime–dataset comparisons against from-scratch and fine-tuned alternatives. We further validate the interface on two physical humanoids, showing that common human-keypoint references support real-world motion tracking.

2 Related Work

Humanoid motion tracking. DeepMimic established reference-driven skill learning (Peng et al., 2018a). ExBody2, ASAP, and TWIST advance expressive control, sim-to-real tracking, and teleoperation, respectively (Ji et al., 2024; He et al., 2025; Ze et al., 2025). BeyondMimic and SONIC build on motion tracking to enable versatile humanoid control (Liao et al., 2025; Luo et al., 2026). Behavioral foundation models offer a complementary approach, learning reusable behaviors through unsupervised RL (Tirinzoni et al., 2025; Li et al., 2026).

Human-motion reference interfaces. SMPL/SMPL-X and AMASS provide human-body representations and motion data (Loper et al., 2015; Pavlakos et al., 2019; Mahmood et al., 2019); GMR and NMR convert human motions into robot references (Araujo et al., 2025; Zhao et al., 2026). SONIC supports human commands through an encoder of 3D joint positions (Luo et al., 2026). We distinguish policy conditioning from training supervision: our reference input uses human keypoints, while retargeted robot trajectories still provide tracking targets.

Cross-embodiment motion tracking. Earlier work shares task modules across robots (Devin et al., 2017) or conditions control on robot-body descriptions (Gupta et al., 2022). XHugWBC supports upper-body intervention with lower-body balance compensation (Xue et al., 2026). Like SONIC, X-WBC explicitly aligns reference encodings through an auxiliary MSE loss across human,

robot, and sparse-VR tokens for the same motion (Luo et al., 2026; Zhang et al., 2026). We instead use a common human-keypoint input and learn shared features directly through tracking, without auxiliary alignment losses. Our source-referenced evaluation complements robot-reference tracking metrics by examining fidelity to human motion near retargeting discontinuities.

Transferring motion knowledge to a new robot. Any2Any reuses a source tracker through input/output dimension matching, joint-level kinematic alignment, and low-rank network adaptation of robot-motion-conditioned policies (Yang et al., 2026). Our transfer boundary is the reference encoder: a common human-motion description passes to a newly trained, two-hidden-layer control head in the target's native state and action spaces. This avoids mapping target joints into a source controller's layout while allowing the target to learn its own execution strategy. We evaluate frozen and fine-tuned encoders against training from scratch under restricted target-motion data.

3 Problem Setup

Robot r has d_r controlled joints and a tracking MDP $\mathcal{M}_r = (\mathcal{S}_r, \mathcal{A}_r, \mathcal{T}_r, R_r, \gamma)$, with robot-specific dynamics and a common tracking-reward formulation. A human motion supplies joint positions $P_t \in \mathbb{R}^{22\times 3}$, pelvis position p_t^{pel}, and rotation matrix R_t^{pel}. The pelvis defines the local coordinate frame for the human-keypoint reference (Section 4.1). Retargeted trajectories $\hat{s}_t^{(r)}$ provide robot-specific tracking targets, initialization, and reference orientation. The actor combines the motion reference with local observations $o_t^{(r)}$ and outputs joint-position offsets $a_t^{(r)} \in \mathbb{R}^{d_r}$ for PD control.

We consider two settings. Joint training learns a shared motion-reference encoder across source robots $\mathcal{R}$, with a separate controller for each body. Transfer reuses this encoder on a target $r^\star \notin \mathcal{R}$ that was excluded from pretraining. Each robot retains its own observation and action definitions and dimensions. The shared interface therefore describes the desired human motion; it does not require identical robot poses, joint correspondences, or a common action space.

4 Method

Our framework learns a shared motion-reference encoder while retaining robot-specific control (Figure 2). Human keypoints describe the desired motion independently of robot joint conventions. Source robots jointly train the encoder through tracking, while separate heads learn execution under their own dynamics. To adapt to a new embodiment, we reuse the frozen reference encoder and train a robot-specific control head from scratch.

4.1 Body-Centric Human-Keypoint Reference

We construct the reference from the first 22 joints of a neutral SMPL-X model with zero shape coefficients. Each frame is expressed in its own pelvis coordinates to remove global translation and rotation, then paired with a near-future frame:

$$\tilde{p}_{t,j} = (R_t^{\mathrm{pel}})^\top (p_{t,j} - p_t^{\mathrm{pel}}), \qquad h_t = \mathrm{vec}(\tilde{P}_t, \tilde{P}_{t+5}) \in \mathbb{R}^{132}. \tag{1}$$

Here $\tilde{P}_t$ stacks the transformed joints. The two frames provide body configuration and 100 ms of look-ahead at 50 Hz, yielding the same 132D reference for every robot. During training, positions receive uniform noise in ± 0.05 m. Excluding retargeted joint positions and velocities from the reference encoder removes their direct input pathway for discontinuities. Retargeted trajectories still supply rewards, initialization, and reference orientation.

4.2 Shared Encoder and Robot-Specific Control

A shared normalizer N_h and MLP f_θ produce $z_t \in \mathbb{R}^{256}$. Each robot's MLP head g_{ϕ_r} combines z_t with observations normalized by N_r. Local observations $o_t^{(r)} \in \mathbb{R}^{3d_r+9}$ comprise joint positions relative to the default pose, joint velocities, the previous action, base angular velocity, and a six-dimensional relative reference-anchor orientation (Zhou et al., 2019).

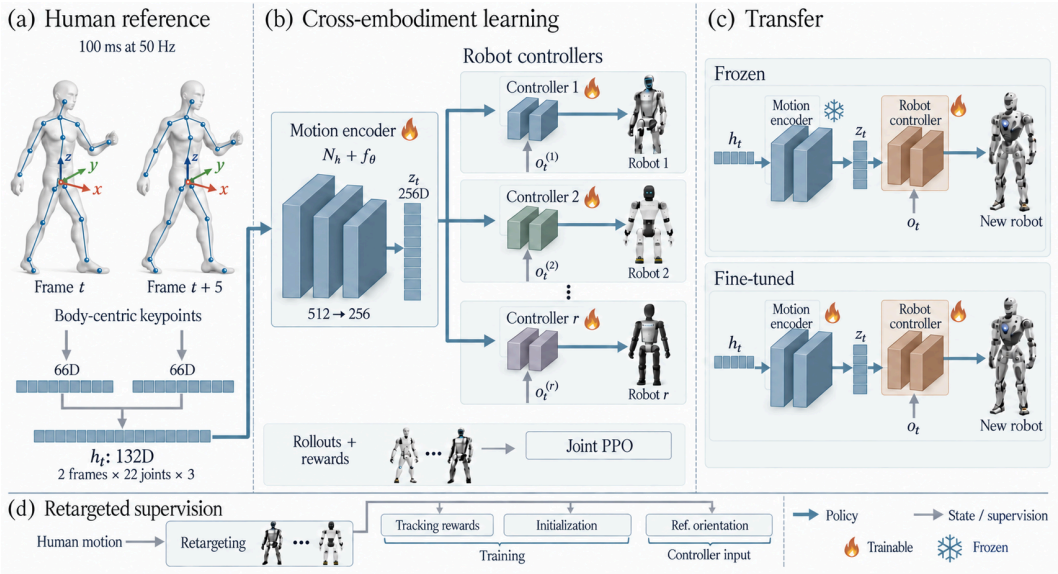

Figure 2: Overview of our framework. (a) Two frames of body-centric keypoints form the human reference. (b) Source robots jointly train a motion encoder and robot-specific controllers. (c) A robot excluded from pretraining learns a new controller with the encoder frozen or fine-tuned in separate runs from the same initialization; freezing also fixes the reference normalizer. Flames and snowflakes indicate trainable and frozen modules. (d) Retargeting supplies rewards, initialization, and reference orientation; o_t includes robot state and reference orientation.

The orientation term retains robot-reference information outside the shared encoder. The actor is

$$z_t = f_\theta(N_h(h_t)), \qquad \mu_t^{(r)} = g_{\phi_r}\left([z_t, N_r(o_t^{(r)})]\right). \tag{2}$$

Each head defines a diagonal Gaussian policy π_r in its robot's native action space, without joint padding; ϕ_r includes head and distribution parameters. Actions are scaled into offsets from the default joint pose for PD control (Appendix A.2).

For a fixed reference, z_t does not depend on the robot's state. Each head learns state-dependent execution with separate normalization, exploration parameters, and a critic, enabling encoder reuse without transferring a source controller. This separation assigns motion description to the shared encoder and feedback control to the embodiment-specific head. The same feature can therefore lead to different actions as balance, joint limits, and the current tracking state differ across robots.

4.3 Joint Training across Embodiments

We use PPO (Schulman et al., 2017) with generalized advantage estimation (Schulman et al., 2016), computed from each robot's own rollouts and critic. The joint actor objective is

$$\max_{\theta,\{\phi_r\}} \quad \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} \left(\mathcal{J}_r + \beta\,\mathbb{E}_t[\mathcal{H}(\pi_r)]\right). \tag{3}$$

Here $\mathcal{J}_r$ is the clipped PPO surrogate, $\mathcal{H}$ is policy entropy, and β is its weight. Every robot contributes gradients to f_θ, whereas each control head receives gradients from its own robot's objective. Actor gradients are averaged across ranks and reference-normalization statistics are pooled; robot-state statistics and value targets remain separate. Robots sample motions independently, so shared feature learning requires neither synchronized reference frames nor an auxiliary cross-robot alignment loss.

We use BeyondMimic-style rewards (Liao et al., 2025), reference-state initialization (Peng et al., 2018a), and failure-aware motion sampling with robot-specific retargeted supervision. Appendix A details the rewards, optimization, and environment differences.

4.4 Adaptation to New Robot Embodiments

For a target $r^\star \notin \mathcal{R}$, we reuse the reference encoder f_θ and normalizer N_h with both held fixed. The same human-keypoint reference therefore produces the same features across embodiments, allowing the target to reuse the learned motion description without retraining the reference encoder.

The target receives a newly initialized control head, state normalizer, action distribution, and critic. Its controller learns in the target's native observation and action spaces to execute the reference under its own dynamics, without mapping joints into a source controller's layout. PPO training uses target-specific retargeted trajectories for tracking supervision and initialization.

Freezing the reference normalizer is part of this transfer: changing its statistics would change the encoder's effective input even with fixed network weights. In the fine-tuned alternative, the pretrained encoder is updated alongside the new head. Both alternatives start from the same pretrained motion features; training from scratch tests the same target-learning task without those features. Adaptation still requires target interaction and supervision, so encoder reuse is distinct from deploying a pretrained controller without further training.

5 Experiments

Our experiments examine whether a common human-motion reference can support accurate tracking, joint learning across robots, and adaptation to new embodiments. We first compare Human Keypoints with Robot Motion on BONES-SEED and AMASS, and assess source-position tracking near retargeting discontinuities. We then evaluate cross-embodiment training and encoder transfer under limited motion data or diversity. Finally, deployments on G1 and K1 demonstrate the use of the same human-keypoint reference on physical humanoids.

5.1 Experimental Setup

The evaluation covers seven humanoids with 21–31 controlled joints: G1, X2, T1, K1, T800, PM01, and Kuavo. Training uses mjlab with MuJoCo (Todorov et al., 2012), 50 Hz control, 4,096 environments per robot, and 24-step rollouts (Appendix A). From our processed BONES-SEED (Bones Studio, 2026) pool, we use 100,000 clips for training and 5,000 held-out clips for evaluation. We also evaluate on 2,000 external AMASS (Mahmood et al., 2019) clips unused in training.

Following prior work (He et al., 2025; Luo et al., 2026), we evaluate motion tracking using pose errors and success rate. The main comparisons emphasize mean per-keybody position error (MPKPE) and success rate to assess tracking with a shared reference interface. MPKPE averages position errors over the tracked robot bodies after aligning the reference with the robot's horizontal anchor position and heading while retaining the reference anchor height. Success rate is the percentage of clips with MPKPE at or below 0.2 m throughout the full rollout. Table 1 also reports orientation, end-effector, and joint errors; a separate analysis evaluates source-position tracking near reference jumps.

Within each comparison, methods use the same evaluation clips and are compared after 150k training iterations. Cross-robot averages give each robot equal weight. The following subsections specify the robot groups used for cross-embodiment learning and transfer.

5.2 Tracking with Human-Keypoint References

We first compare Human Keypoints with Robot Motion, which conditions the policy on current retargeted joint positions and velocities. Both retain robot state, reference orientation, and retargeted supervision. Human Keypoints also uses look-ahead and reference noise; the comparison evaluates these choices together with the reference representation.

Human Keypoints achieves comparable average success rate and position accuracy on BONES-SEED: 91.45% versus 91.57% success rate and 0.0710 versus 0.0697 m MPKPE (Table 1). On AMASS, Human Keypoints has lower MPKPE (0.0923 versus 0.1117 m) but a slightly lower success rate (85.03% versus 85.74%). MPKPE improves on T1, K1, T800, and Kuavo, while Robot Motion performs better on G1, X2, and PM01 (Appendix B.1). These results support a shared

Table 1: Full reference-input metrics at 150k: cross-robot arithmetic means, with n paired robots. Arrows indicate the preferred direction; no statistical equivalence is implied.

	BONES-SEED ($n = 7$)		AMASS ($n = 7$)	
Metric	Human Keypoints	Robot Motion	Human Keypoints	Robot Motion
Success rate (%) ↑	91.45	91.57	85.03	85.74
MPKPE (m) ↓	0.0710	0.0697	0.0923	0.1117
MPKOE (rad) ↓	0.3098	0.2674	0.3852	0.3673
End-effector position error (m) ↓	0.1073	0.1030	0.1295	0.1424
End-effector orientation error (rad) ↓	0.3925	0.3169	0.4844	0.4246
Joint position error (rad) ↓	0.1840	0.1599	0.2187	0.1811
Joint velocity error (rad/s) ↓	0.6179	0.5860	0.6535	0.6208

human-reference interface across the evaluated bodies. Per-robot results appear in Appendix B.1; Appendix B.5 provides the BONES-SEED learning curves.

5.3 Source-Position Tracking around Reference Jumps

Retargeting can introduce abrupt joint-angle changes that are absent from the source human motion (Figure 3). We compare how closely Robot Motion and Human Keypoints follow source human positions near these reference discontinuities on the same G1 motions.

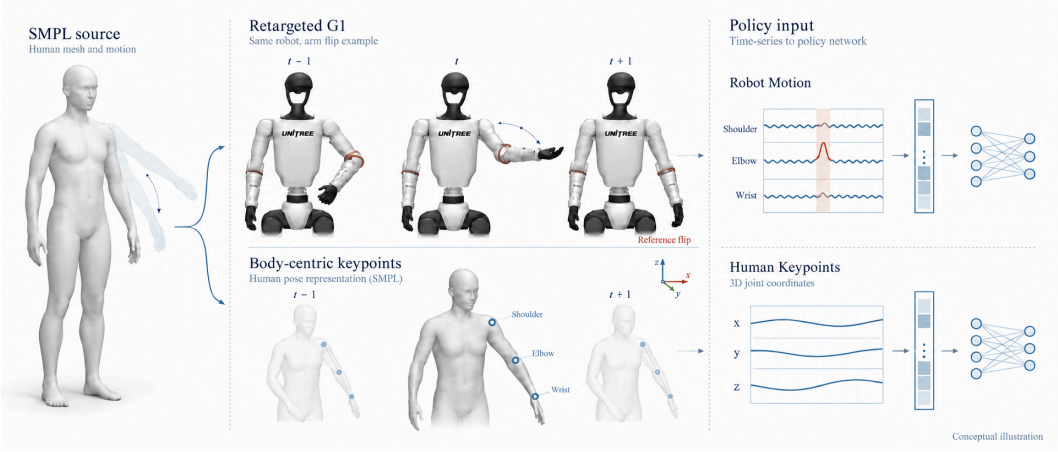

Figure 3: **Two reference pathways near a retargeting discontinuity.** Robot Motion consumes retargeted joint positions and velocities, which can contain discontinuities absent from the source human motion. Human Keypoints consumes body-centric source positions. Both policies retain robot state, reference orientation, retargeted rewards, and initialization. Poses, human surfaces, traces, and networks are conceptual illustrations; measured examples appear in Figure 4.

We evaluate both 150k G1 policies on a screened cohort of BONES-SEED training clips containing pronounced retargeting discontinuities absent from the corresponding source human motion. Selection uses only source and retargeted reference poses, retaining the largest qualifying discontinuity per clip with ± 0.5 s of context (96 clips in 64 mirror-grouped families). With a fixed pre-IK GMR mapping (Araujo et al., 2025) defining the source reference $\tilde{p}^*$, we report the mean source-position error $E_p = \frac{1}{13|W|} \sum_{t \in W, j} \| \tilde{p}^{\text{rollout}}_{t,j} - \tilde{p}^*_{t,j} \|_2$ over the 13 tracked non-root bodies within the ± 200 ms window W of each event, with poses pelvis-centered and yaw-aligned (excluding global root translation) and clips weighted equally. Human Keypoints reduces this error by 8.0% relative to Robot Motion (paired bootstrap -1.09 cm, 95% interval $[-2.13, -0.35]$). Figure 4 and Appendix B.7 show separate illustrative cases outside this aggregate.

The lower source-position error supports improved positional fidelity on this screened cohort. Look-ahead and input-noise differences prevent attributing the gain solely to the keypoint representation, and this selected cohort does not establish robustness across all motions.

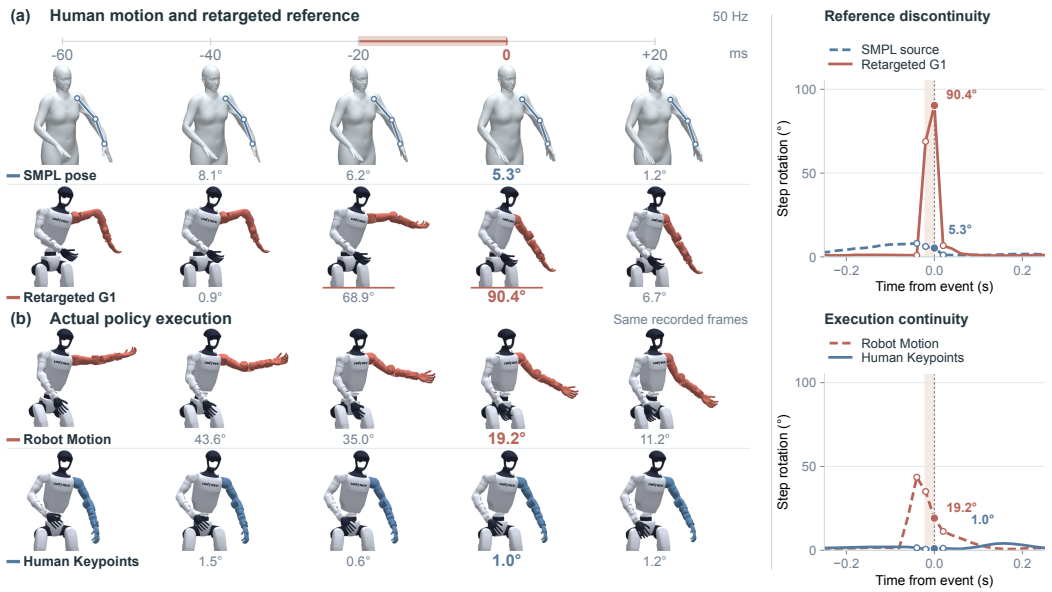

Figure 4: **Reference discontinuity and execution continuity.** In this selected sequence, the retargeted reference jumps while the SMPL source remains smooth; Human Keypoints yields smoother execution than Robot Motion. Values report torso-relative forearm rotation over the preceding 20 ms (50 Hz). Snapshots are centered for pose comparison. Human surfaces are illustrative, with standardized hand orientations; curves use recorded rotations. See Appendix B.6.

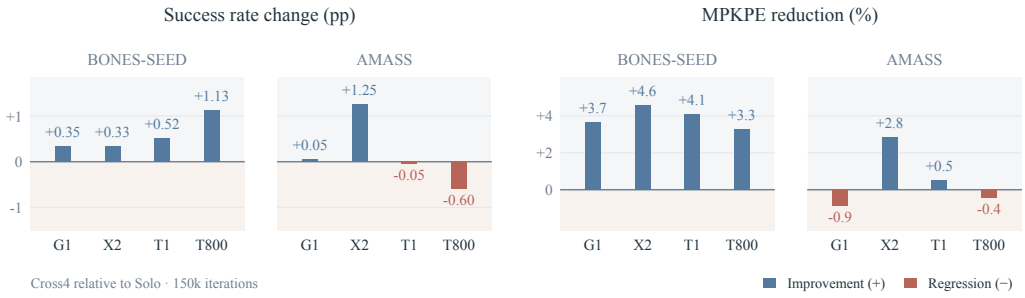

Figure 5: **Cross-embodiment learning.** Cross4 relative to Solo at 150k iterations: success rate change in percentage points (left pair) and relative MPKPE reduction (right pair). Positive values indicate improvement for both metrics. Scales are shared within each metric. Compute and interaction budgets are unmatched.

5.4 Cross-Embodiment Learning

We evaluate cross-embodiment learning on G1, X2, T1, and T800. These robots share a reference encoder but retain separate control heads (Cross4); each is compared with its single-embodiment baseline (Solo). On BONES-SEED, Cross4 reduces MPKPE by 3.7%, 4.6%, 4.1%, and 3.3%, respectively, and improves success rate for all four (Figure 5). Appendix B.4 provides the BONES-SEED learning curves.

The gains are largest on BONES-SEED and muted on AMASS: X2 still improves in both metrics, while G1, T1, and T800 sit essentially at parity with Solo, with success rate and MPKPE each differing by under one point and under one percent (Table 6). Solo denotes the single-embodiment Human Keypoints policies of Section 5.2, evaluated with the same per-motion-mean MPKPE aggregation as Cross4. Cross4 nonetheless aggregates more total robot interactions than a single-embodiment run. The comparison therefore demonstrates feasible cross-embodiment learning—with clear in-distribution gains and out-of-distribution parity—but does not isolate embodiment diversity from additional training experience or establish compute efficiency.

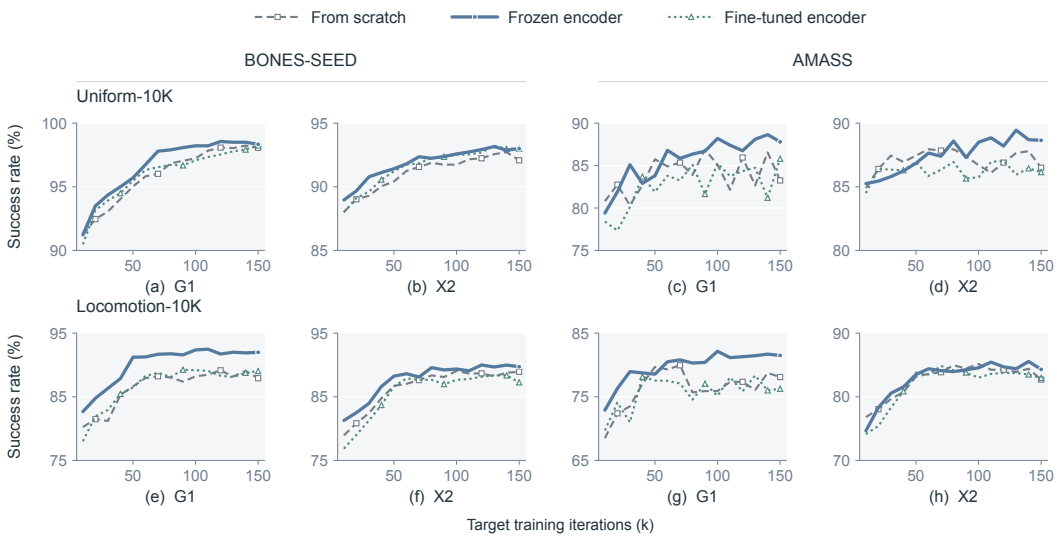

Figure 6: **Success rate during transfer to new robots.** Checkpoint evaluations without smoothing; panels use separate ranges. Iterations exclude pretraining.

5.5 Transfer to Unseen Humanoids

We evaluate whether pretrained motion representations improve tracking on unseen humanoids with limited training data or motion diversity. The encoder is pretrained on 100,000 BONES-SEED motions using T1, T800, PM01, and Kuavo, then transferred to G1 and X2. We compare frozen-encoder transfer, encoder fine-tuning, and training from random initialization, with new control heads trained for 150k iterations.

Both training subsets contain 10,000 motions. *Uniform-10K*, sampled uniformly from the pretraining pool, evaluates whether pretraining on more motions improves tracking under a similar training distribution. *Locomotion-10K*, restricted to walking and jogging, evaluates generalization to categories absent from the new robot's training set, such as dance. All methods use the same subset and the evaluation protocol in Section 5.1.

Table 2: **Transfer to unseen humanoids at 150k iterations.** Each cell reports success rate (%, ↑) / MPKPE (m, ↓). Bold denotes best values, including ties.

	BONES-SEED			AMASS		
Robot	From scratch	Frozen	Fine-tuned	From scratch	Frozen	Fine-tuned
Uniform-10K						
G1	98.06 / 0.0475	**98.34** / 0.0485	98.24 / **0.0471**	83.25 / 0.1056	**87.80** / **0.0902**	85.80 / 0.0984
X2	92.08 / 0.0699	**93.00** / **0.0625**	**93.00** / 0.0626	86.50 / 0.0920	**88.65** / **0.0848**	86.15 / 0.0879
Locomotion-10K						
G1	87.92 / 0.0663	**91.98** / **0.0594**	89.04 / 0.0649	78.10 / 0.1247	**81.50** / **0.1157**	76.25 / 0.1312
X2	88.92 / 0.0683	**89.72** / **0.0673**	87.26 / 0.0733	82.70 / **0.0957**	**84.30** / 0.0965	82.95 / 0.0985

Figure 6 and Table 2 show success rate gains from frozen-encoder transfer over training from scratch on both datasets. On BONES-SEED, gains are 0.28–0.92 percentage points with Uniform-10K and 0.80–4.06 with Locomotion-10K; corresponding AMASS gains are 2.15–4.55 and 1.60–3.40 points. Freezing achieves the highest or tied-highest success rate in all eight comparisons and the lowest MPKPE in six, supporting feature reuse under both data restrictions. Appendix B.3 (Figure 8) provides the MPKPE trajectories.

These outcomes distinguish successful completion from average positional precision: the encoder with the highest success rate need not minimize MPKPE in every setting. Frozen transfer preserves source-learned motion features while fitting target-specific execution. The comparison concerns adaptation after pretraining; it does not establish a reduction in total training cost.

5.6 Real-World Motion Tracking on G1 and K1

We deploy the Human Keypoints interface on Unitree G1 and Booster K1, which simultaneously track the same prerecorded human motion. Figure 7 illustrates their execution across four motion types: locomotion, gestures, boxing, and dance. These deployments demonstrate that a common human-keypoint reference can guide multiple physical humanoids through robot-specific controllers despite differences in body size and joint configuration.

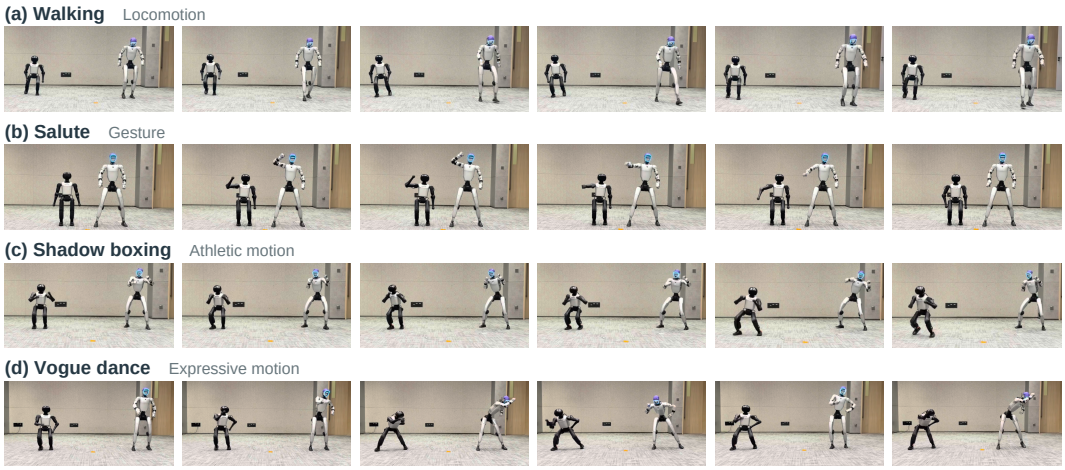

Figure 7: **Same motion reference on two physical humanoids.** K1 (left) and G1 (right) execute four motions using Human Keypoints. Representative frames are shown in chronological order from left to right; each panel retains both robots and ground contacts from one camera frame.

6 Discussion and Limitations

Several limitations bound the scope and interpretation of our results. The study covers humanoid robots only, so how the shared human-keypoint representation transfers to more diverse morphologies, body proportions, and joint layouts remains open. Although the reference encoder is shared across bodies, our controllers still rely on robot-specific control heads and per-robot retargeted supervision. Finally, exploring more principled and unified forms of the reference-motion interface remains a promising direction for future work.

7 Conclusion

We reframe humanoid motion tracking around a common human-motion reference and a factored controller: a shared reference encoder reads a robot-agnostic human-keypoint description of the desired motion, while robot-specific control heads learn how each body executes it. Because the reference is expressed in human terms rather than in any single robot's joint space, the same description drives bodies that differ in size and kinematics, and new reference motions require no per-robot respecification. This factorization lets a single interface serve heterogeneous humanoids, supporting joint learning across source robots and, with the encoder frozen, transfer to embodiments unseen during pretraining. Empirically, human-keypoint conditioning matches average position tracking on the primary benchmark, lowers source-position error near screened reference jumps on G1, and improves target success rate across the evaluated transfer settings. Separating what a motion is from how a given body executes it lets a controller reuse motion understanding as new robots are added, rather than relearning it from scratch for each platform. Together these results indicate that describing motion once, in human terms, and letting each body learn its own execution is a practical route toward humanoid controllers that accumulate, rather than relearn, motor skill as new bodies arrive.

AI USE STATEMENT

Generative AI tools assisted with literature search, manuscript organization, English drafting and editing, figure preparation, and the compilation of experimental data into tables.

REPRODUCIBILITY STATEMENT

Full training settings, including the network architecture, hyperparameters, and reward definitions, are provided in Appendix A. We will publicly release our code upon acceptance to facilitate reproduction and further research.

REFERENCES

Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025. URL https://arxiv.org/abs/2510.02252.

Bones Studio. BONES-SEED: Skeletal everyday embodiment dataset, 2026. URL https://bones.studio/datasets/seed. Dataset; accessed September 22, 2026.

Coline Devin, Abhishek Gupta, Trevor Darrell, Pieter Abbeel, and Sergey Levine. Learning Modular Neural Network Policies for Multi-Task and Multi-Robot Transfer. In *IEEE International Conference on Robotics and Automation*, 2017. URL https://arxiv.org/abs/1609.07088.

Figure AI. Introducing Helix 02: Full-body autonomy. Technical announcement, January 27, 2026. URL https://www.figure.ai/news/helix-02.

Agrim Gupta, Linxi Fan, Surya Ganguli, and Fei-Fei Li. MetaMorph: Learning universal controllers with transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Opmqtk_GvYL.

Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica Hodgins, Linxi "Jim" Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. ASAP: Aligning Simulation and Real-World Physics for Learning Agile Humanoid Whole-Body Skills. *arXiv preprint arXiv:2502.01143*, 2025. URL https://arxiv.org/abs/2502.01143.

Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. ExBody2: Advanced Expressive Humanoid Whole-Body Control. *arXiv preprint arXiv:2412.13196*, 2024. URL https://arxiv.org/abs/2412.13196.

Yitang Li, Zhengyi Luo, Tonghe Zhang, Cunxi Dai, Anssi Kanervisto, Andrea Tirinzoni, Haoyang Weng, Kris Kitani, Mateusz Guzek, Ahmed Touati, Alessandro Lazaric, Matteo Pirotta, and Guanya Shi. BFM-Zero: A Promptable Behavioral Foundation Model for Humanoid Control Using Unsupervised Reinforcement Learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/80b677bb4d17f77220be88e1d1fdebde-Paper-Conference.pdf.

Qiayuan Liao, Takara E. Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C. Karen Liu. BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025. URL https://arxiv.org/abs/2508.08241v4.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. URL https://smpl.is.tue.mpg.de/.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):

eaed4592, 2026. doi: 10.1126/scirobotics.aed4592. URL https://doi.org/10.1126/scirobotics.aed4592.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1904.03278.

NVIDIA. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. URL https://arxiv.org/abs/2503.14734.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1904.05866.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills. *ACM Transactions on Graphics*, 37(4), 2018a. doi: 10.1145/3197517.3201311. URL https://doi.org/10.1145/3197517.3201311.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. In *IEEE International Conference on Robotics and Automation*, 2018b. URL https://arxiv.org/abs/1710.06537.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirotta. Zero-Shot Whole-Body Humanoid Control via Behavioral Foundation Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/37611b0fc2b65cdbb60865af5f6cf453-Paper-Conference.pdf.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109. URL https://doi.org/10.1109/IROS.2012.6386109.

Yufei Xue, YunFeng Lin, Wentao Dong, Yang Tang, Jingbo Wang, Jiangmiao Pang, Ming Zhou, Minghuan Liu, and Weinan Zhang. Scalable and General Whole-Body Control for Cross-Humanoid Locomotion. *arXiv preprint arXiv:2602.05791*, 2026. URL https://arxiv.org/abs/2602.05791v3.

Ming Yang, Tao Yu, Feng Li, and Hua Chen. Any2Any: Efficient Cross-Embodiment Transfer for Humanoid Whole-Body Tracking. *arXiv preprint arXiv:2605.23733*, 2026. URL https://arxiv.org/abs/2605.23733v3.

Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. TWIST: Teleoperated Whole-Body Imitation System. *arXiv preprint arXiv:2505.02833*, 2025. URL https://arxiv.org/abs/2505.02833.

Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C. Karen Liu. TWIST2: Scalable, portable, and holistic humanoid data collection system. In *IEEE International Conference on Robotics and Automation*, 2026. URL https://arxiv.org/abs/2511.02832.

Juntong Zhang, Chun Gu, and Li Zhang. X-WBC: A Cross-Embodiment Foundation Model for Humanoid Whole-Body Control. In *Conference on Robot Learning*, 2026. URL https://arxiv.org/abs/2609.15213v1. Accepted; arXiv:2609.15213.

Qingrui Zhao, Kaiyue Yang, Xiyu Wang, Shiqi Zhao, Yi Lu, Xinfang Zhang, Qiu Shen, Xiao-Xiao Long, and Xun Cao. Make Tracking Easy: Neural Motion Retargeting for Humanoid Whole-body Control. *arXiv preprint arXiv:2603.22201*, 2026. URL https://arxiv.org/abs/2603.22201v3.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1812.07035.

A Implementation Details

A.1 Robot Models and Training Configuration

Table 3 summarizes the controlled degrees of freedom and experimental roles of the seven humanoids. These dimensions refer to the simulated control models.

Table 3: Robot models; all enter the reference-input comparison.

Robot	Controlled DoFs	Role in the main experiments
Unitree G1	29	Joint training; transfer target; hardware
AgiBot X2	31	Joint training; transfer target
Booster T1	21	Joint training; transfer source
Booster K1	22	Hardware demonstration
EngineAI T800	25	Joint training; transfer source
EngineAI PM01	24	Transfer source
Leju Kuavo	29	Transfer source

Table 4 lists implementation defaults; architecture and environment settings vary by experiment.

Table 4: Simulation and policy-training defaults.

Setting	Value
Physics step / control decimation	0.005 s / 4
Parallel environments per robot	4,096
Rollout steps per iteration	24
PPO epochs / minibatches	5 / 4
Discount / GAE parameter	0.99 / 0.95
PPO clipping / entropy coefficient	0.2 / 0.005
Initial learning rate / desired KL	10^{-3} / 0.01
Maximum gradient norm	1.0
Encoder hidden widths / latent width	[512, 256] / 256
Control-head hidden widths	[512, 256]
Gaussian standard-deviation bounds	[0.001, 0.5]

A.2 Rewards and Control

Tracking rewards use exponential kernels for anchor position/orientation, body position/orientation, and body linear/angular velocity. Weights are $(0.5, 0.5, 1, 1, 1, 1)$; scales are $(0.3\,\mathrm{m}, 0.4\,\mathrm{rad}, 0.3\,\mathrm{m}, 0.4\,\mathrm{rad}, 1.0\,\mathrm{m/s}, 3.14\,\mathrm{rad/s})$. Penalties cover action rate, joint limits, and supported self-collisions. Training uses perturbed reference-state initialization and dynamics randomization (Peng et al., 2018b); evaluation disables randomization. Selected collision and friction terms are omitted for T800, PM01, and Kuavo where collision geometry is unavailable.

Actions specify PD targets $q^{(r)}_{\mathrm{target}} = q^{(r)}_{\mathrm{default}} + \alpha_r \odot a^{(r)}_t$, with scales proportional to effort limits divided by proportional gains. Training termination checks anchor height, projected gravity, and end-effector height; evaluation uses the MPKPE criterion in Section 5.1.

A.3 Evaluation and Aggregation

The fixed BONES-SEED and AMASS evaluation sets are disjoint from training and shared across comparisons. Uniform-10K and Locomotion-10K contain equal numbers of training motions but differ in category coverage, as described in Section 5.5.

Robot-level errors weight clips by evaluated length; success rate weights clips equally. Cross-robot summaries weight robots equally. Source-referenced event-window errors (Section 5.3) instead weight clips equally.

B Additional Evaluation Results

We provide per-robot tracking results, learning curves, and a source-referenced analysis of execution near retargeting discontinuities.

B.1 Per-Robot Tracking Results

Table 5 reports the per-robot AMASS results underlying Table 1. Human Keypoints achieves lower MPKPE on T1, K1, T800, and Kuavo; success-rate gains occur on T1 and K1. Robot Motion uses zero head-joint references on K1, T800, PM01, and Kuavo, and each K1 policy uses its training PD gains.

Orientation error is the shortest SO(3) rotation error after the same anchor alignment used for MPKPE. End-effector errors cover the configured feet and upper-limb endpoints; joint errors compare corresponding positions and velocities against the retargeted reference.

Table 5: Reference-input comparison on AMASS at 150k iterations: success rate (%) / MPKPE (m). Values are rounded for display; Table 1 averages the underlying per-robot values with equal weight.

Robot	Human Keypoints	Robot Motion
G1	85.95 / 0.0946	88.05 / 0.0862
X2	86.25 / 0.0898	87.55 / 0.0885
T1	85.50 / 0.0769	84.65 / 0.0837
K1	91.20 / 0.0747	88.40 / 0.0986
T800	79.25 / 0.1159	80.95 / 0.1782
PM01	85.70 / 0.0879	87.30 / 0.0835
Kuavo	81.35 / 0.1064	83.30 / 0.1630

B.2 Cross-Embodiment Learning Results

Table 6 gives the absolute results underlying Figure 5. Cross4 improves both metrics on BONES-SEED, while AMASS outcomes vary by robot. The Solo baselines are the single-embodiment Human Keypoints policies, so their AMASS values coincide with Table 5; Cross4 is trained jointly and evaluated with the same per-motion-mean aggregation.

Table 6: Cross-embodiment (Cross4) versus single-embodiment (Solo) training at 150k iterations: success rate (%) / MPKPE (m). Cross4 trains G1/X2/T1/T800 jointly. Interaction and compute budgets are unmatched, so this comparison does not measure compute efficiency.

Evaluation	Robot	Solo	Cross4
BONES-SEED	G1	98.2 / 0.0497	98.6 / 0.0478
	X2	92.6 / 0.0664	92.9 / 0.0634
	T1	90.3 / 0.0655	90.8 / 0.0629
	T800	86.8 / 0.0902	88.0 / 0.0872
AMASS	G1	85.95 / 0.0946	86.0 / 0.0955
	X2	86.25 / 0.0898	87.5 / 0.0873
	T1	85.50 / 0.0769	85.45 / 0.0765
	T800	79.25 / 0.1159	78.65 / 0.1164

B.3 Position-Error Trajectories

Figure 8 complements Figure 6 with MPKPE during transfer. Table 2 reports the final checkpoint results.

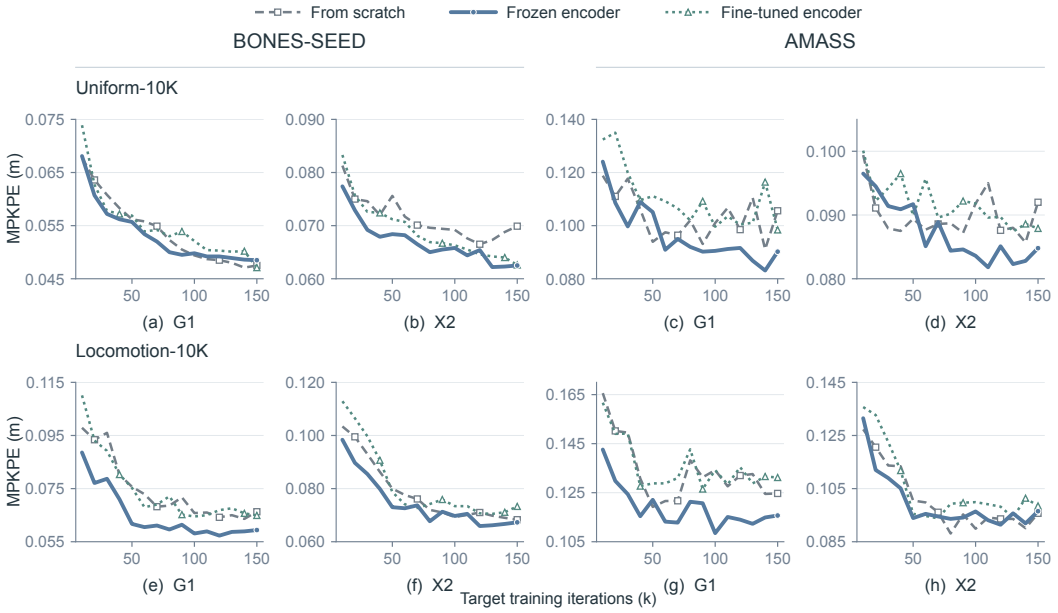

Figure 8: Position error during transfer to new robots, complementing Figure 6. Rows show Uniform-10K and Locomotion-10K training. Lower MPKPE is better; panels use separate ranges. Displayed curves retain all checkpoints through 150k without smoothing or extrapolation.

B.4 BONES-SEED Cross-Embodiment Learning Curves

Figure 9 shows the BONES-SEED learning curves underlying Figure 5 and Table 6.

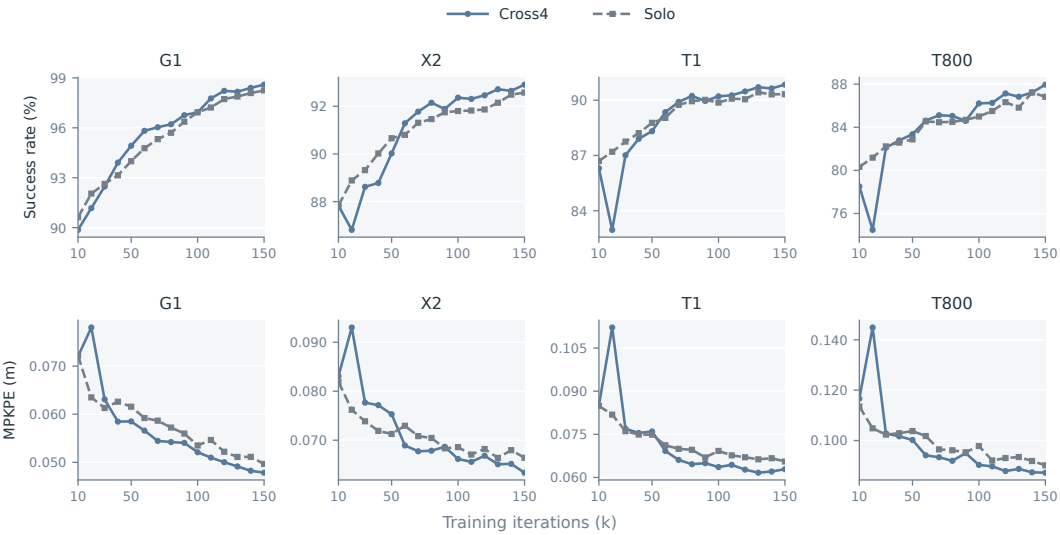

Figure 9: BONES-SEED checkpoint evaluations for cross-embodiment (Cross4) and single-embodiment (Solo) training. Columns show G1, X2, T1, and T800; rows show success rate and MPKPE. All recorded checkpoints from 10k to 150k are retained without smoothing. Panels use separate vertical ranges. Training iterations do not imply matched total robot interactions or compute between Cross4 and Solo.

B.5 BONES-SEED Reference-Input Learning Curves

Figures 10 and 11 supplement Section 5.2 with held-out BONES-SEED evaluations at checkpoints from 10k to 150k training iterations. Each panel compares the two reference inputs on the same robot; the final panel averages the seven robots with equal weight.

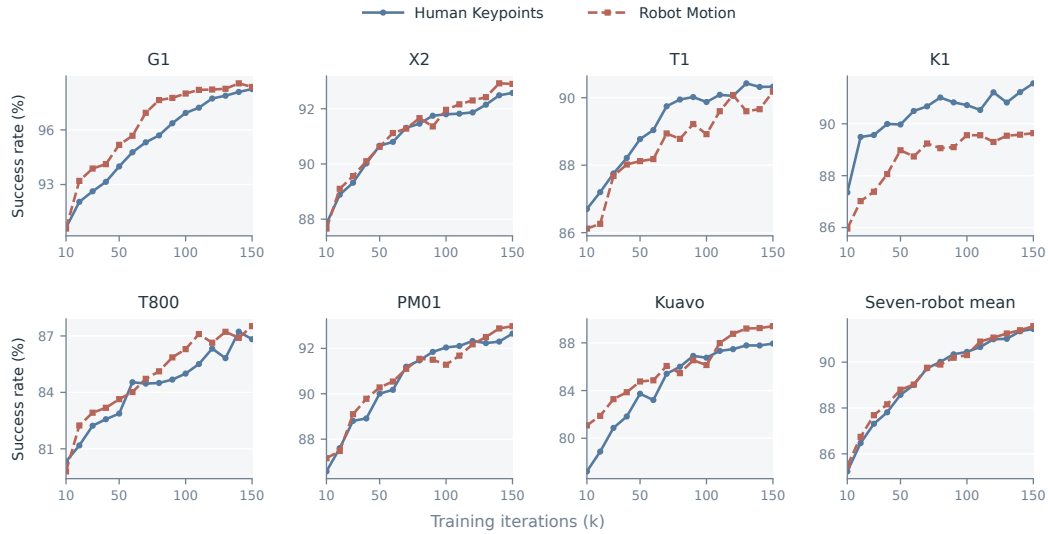

Figure 10: BONES-SEED success rate during reference-input training. Human Keypoints and Robot Motion are evaluated on the fixed 5,000-motion evaluation set. Markers show recorded checkpoints; lines connect them without smoothing. Panels use separate vertical ranges.

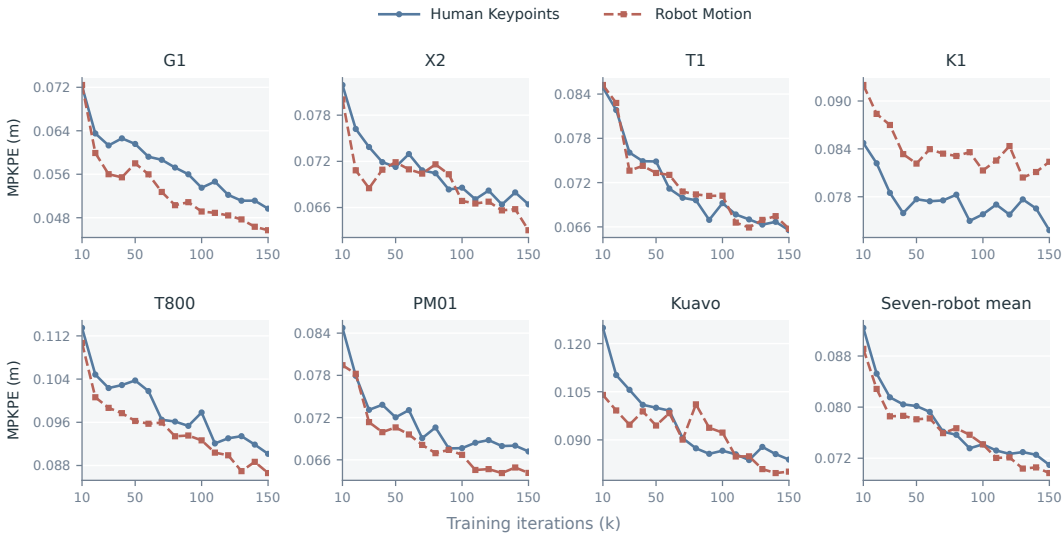

Figure 11: BONES-SEED MPKPE for the reference-input comparison in Figure 10, measured in metres (lower is better). The final seven-robot means correspond to Table 1. Panels use separate vertical ranges.

B.6 Qualitative Execution Examples

Figure 4 compares the source motion, retargeted reference, and recorded G1 execution around a forearm discontinuity. Curves measure the shortest SO(3) rotation between successive torso-relative forearm orientations at 50 Hz. They are invariant to quaternion sign and angular wrapping. This example was selected from outcome-screened recordings and is illustrative rather than a population-level comparison.

Snapshots share a fixed camera within each case and are centered for pose comparison; displayed translations do not alter the recorded states used for the metrics. Human surfaces and standardized hand poses are illustrative. Curves use the recorded rotations without smoothing or temporal realignment.

B.7 Additional Motion Examples

Figures 12–14 illustrate reaching, celebration, and dance. Reference events are selected as pronounced retargeting discontinuities that are absent from the corresponding source motion. These examples are drawn from a separate search that includes outcome-screened recordings; they are not part of the aggregate evaluation in Section 5.3.

The reaching and dance panels measure upper-arm rotation relative to the torso; the celebration panel measures hand rotation relative to the forearm. All streams retain recorded timing. The examples show how local execution responses differ across motions, without implying that either policy is uniformly smoother.

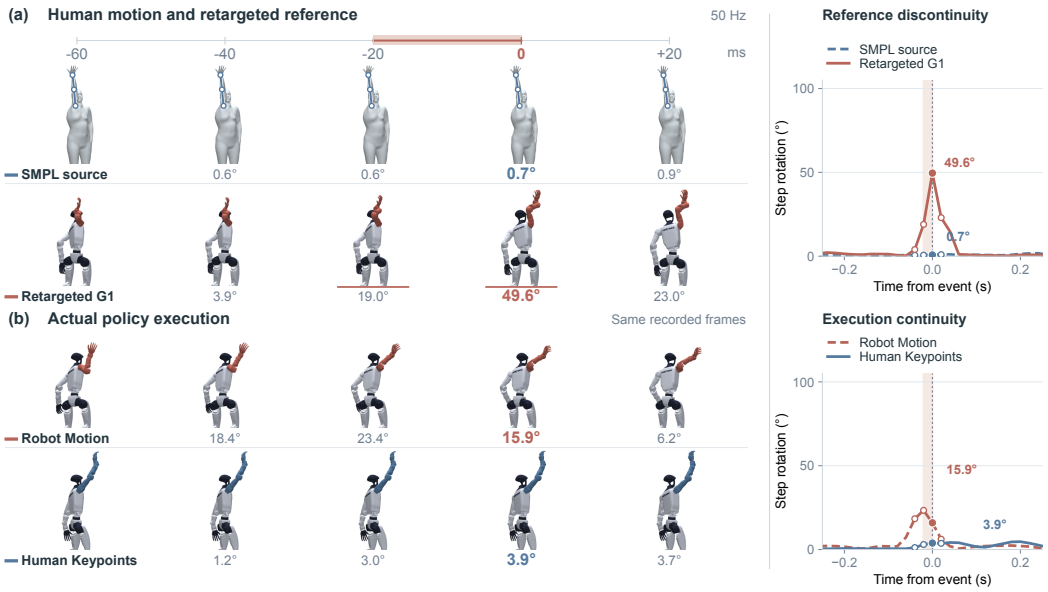

Figure 12: **Upward reaching.** A raised right arm illustrates a reference discontinuity with little corresponding source rotation. Values and curves show right upper-arm rotation relative to the torso over 20 ms. Both policy rows use recorded execution states.

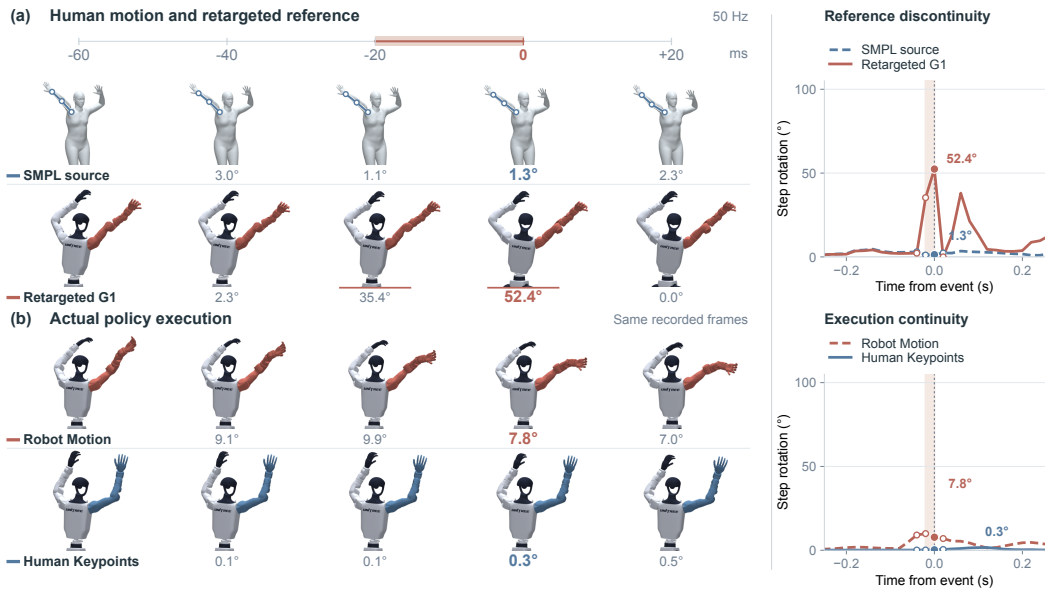

Figure 13: **Two-handed celebration.** With both arms raised, the right hand rotates abruptly in the reference. Values and curves measure hand rotation relative to the forearm, emphasizing wrist orientation rather than whole-arm displacement.

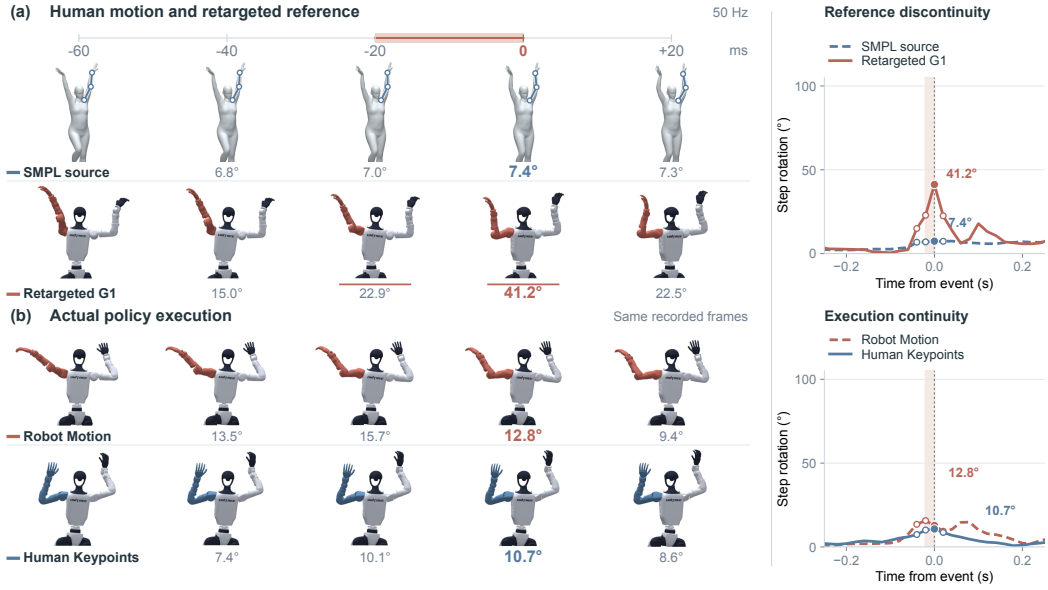

Figure 14: **Latin dance.** Values and curves show left upper-arm rotation relative to the torso. The policies have a smaller event-time contrast here, illustrating variation in execution response across selected reference discontinuities.