HumanTracker: Towards Comprehensive and Human-Aligned Motion Tracking Benchmark

Dairu Liu[1,3,*], Zekun Qi[2,3,*], Jiayu Zeng[3,*], Yu Guan[2,3], Chenghuai Lin[3], Xuchuan Chen[2,3], Xinqiang Yu[3], Wenyao Zhang[3,4], He Wang[3,5] ✉, and Li Yi[2,6] ✉

[1] Nankai University, China
[2] Tsinghua University, China
[3] GALBOT, China
[4] Shanghai Jiao Tong University, China
[5] Peking University, China
[6] Shanghai Qi Zhi Institute, China

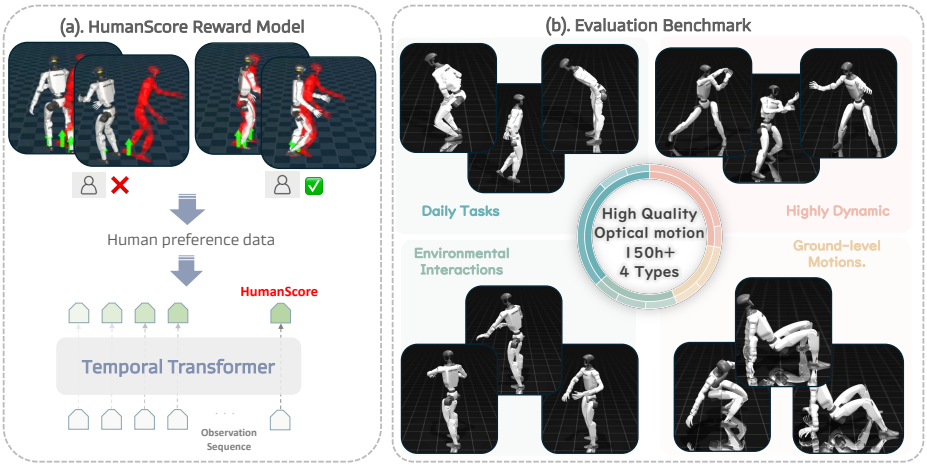


Fig. 1: HumanTracker overview. Left: we train a Human-Aligned Reward Model on pairwise preference data from tracking rollouts, producing a scalar HumanScore via a temporal Transformer. Right: the HumanTracker Benchmark provides 150+ hours of motion capture across four families: daily tasks, highly dynamic motions, environmental interactions, and ground-level movements.

Abstract. Humanoid motion tracking is central to teleoperation and whole-body imitation, yet evaluation often disagrees with what people perceive in videos. Kinematic errors average per-frame pose differences but miss the physical artifacts that matter most, such as unstable support and incorrect contacts (e.g., foot skating and mistimed touchdowns). Meanwhile, widely used test suites are small and lack the diversity needed to stress contact-rich, long-horizon behaviors. We introduce

✉ Corresponding author.
*Equal contribution.

HumanTracker to make humanoid tracking evaluation both perceptually aligned and scalable. HumanTracker contributes 150 hours of newly captured optical motion from 24 professional performers, organized into four motion families with text labels for fine-grained diagnosis. We further propose HumanScore, a preference-aligned metric trained from 12K human-labeled motion pairs on synchronized tracking videos via a trajectory reward model. Across representative state-of-the-art trackers, HumanScore better predicts held-out human preferences and reveals contact and stability failures that kinematic metrics often miss. The project page is available at `https://dairuliu.github.io/humantracker`.

Keywords: Humanoid tracking · Benchmark · Human preference

1 Introduction

Motion tracking is becoming the cornerstone of humanoid robot control [6, 28, 32,33,40,45,51]. This ability drives applications in teleoperation and whole-body imitation [10, 12, 18, 19, 27, 53, 54]. Most systems track a reference motion using a learned feedback policy in physics simulation [6, 32, 33, 51]. Yet, measuring the true quality of this tracking remains surprisingly difficult. Two major issues currently limit progress in this field.

The first major issue is how we measure success. Motion tracking looks easy to measure: one can compare the rollout pose to the reference pose and report kinematic errors such as mean joint error and key point error [6, 33, 39, 51, 60]. However, we find a clear gap between these numbers and what people see in videos. A rollout may achieve low kinematic error but still look bad, especially under contacts [1, 25, 48, 49]. Feet may slide on the ground, and contacts may break and reattach at the wrong time; these artifacts are exactly what contact-aware global motion reconstruction and refinement methods target in video-based settings as well [41, 55, 58]. These artifacts are the key aspect to decide whether the motion looks stable, smooth and human-like.

This gap is not a corner case. It is a consequence of what kinematic metrics measure. They treat tracking as a per-frame pose matching problem. They average errors over joints and time. They do not directly represent contact, support, or balance. They also do not reflect how errors accumulate in closed-loop control. Two rollouts can have similar pose errors but very different stability. Observers reliably prefer the one with clean contacts and steady support, even when MPJPE is similar. Figure 1 highlights this mismatch.

The second major issue is the scope of evaluation data. Despite the availability of large motion repositories such as PHUMA [25] and SONIC [33], the most commonly used evaluation suite for humanoid tracking is still an AMASS test set with only 140 sequences [34]. This small set lacks diversity and underrepresents the long tail of human movement, including challenging contact transitions, asymmetric balancing, and complex recoveries. Moreover, results are often summarized as a single aggregate score, without a detailed breakdown by motion category, making it difficult to pinpoint exactly where and why a tracker fails.

Table 1: Comparison of large-scale humanoid motion datasets. HumanTracker uniquely provides 150 hours of newly captured data explicitly organized into distinct motion categories for comprehensive evaluation.

Dataset	Clips	Hours	Categories	Text Label
AMASS [34]	>11K	>40	No	No
HumanML3D [3]	14.6K	28.6	No	Yes
PHUMA [25]	76K	73	No	No
HumanTracker	24K	**150**	**4**	**Yes**

We introduce HumanTracker to address these problems. To address the metric misalignment, we develop a preference-aligned evaluation: we collect human comparisons on synchronized tracking videos and train a reward model to predict these preferences [9, 36, 46]. We call this metric HumanScore. Unlike joint error, HumanScore explicitly targets human preferences, perceptual stability, and realistic physical contacts. It penalizes the exact physical artifacts that look wrong to human observers. Importantly, we demonstrate that HumanScore captures nuanced perceptual qualities that cannot be simply reduced to a linear combination of non-learned diagnostics such as foot slip or root drift. Extensive experiments and visual analyses demonstrate the accuracy and zero-shot generalization of HumanScore.

To address the lack of diversity in existing test suites, we provide 150 hours of newly captured optical motion data paired with category and text labels, complementing existing large-scale motion sources [25, 29, 34, 57]. All motions are recorded in a controlled studio with a multi-camera optical system by 24 professional performers, including dance teachers and fitness coaches, yielding high-fidelity references for contact-rich tracking evaluation. We explicitly organize the dataset into four distinct families covering daily tasks, highly dynamic movements, environmental interactions, and ground-level motions, enabling per-family metrics and fine-grained diagnosis of failure modes. At this scale and with explicit motion taxonomy and text annotations, HumanTracker substantially exceeds prior mocap datasets used for humanoid tracking (Table 1). We comprehensively evaluate several state-of-the-art trackers, including GMT, TWIST2, SONIC, and Humanoid-GPT, under our benchmark [6, 33, 39, 54].

In summary, our main contributions are threefold. First, we highlight the failure of traditional kinematic metrics and introduce HumanScore to align evaluation with human perception. Second, we provide a massive motion tracking benchmark featuring 150 hours of diverse, categorized optical data with text descriptions. Third, we establish a rigorous and standardized evaluation protocol to ensure future progress in humanoid tracking is both measurable and meaningful.

2 Related Works

2.1 Humanoid motion tracking at scale

Humanoid motion tracking is typically posed as motion imitation in physics simulation, where a learned tracking policy follows reference trajectories under contacts and perturbations [6, 31–33, 51]; early large-scale imitation established robust recovery under curricula/priors [32, 38], and recent work scales toward general and zero-shot whole-body trackers as well as deployment-ready skill spaces [6, 7, 17, 20, 22, 33, 35, 39, 44, 51, 52, 59, 60]. Systems increasingly combine tracking with residual refinement, multi-modal targets, and constrained objectives for physical feasibility [10, 21, 45, 48, 62, 63], while alternative sequence modeling and generative control perspectives (diffusion, token prediction, long-horizon policy learning) broaden the design space [2,8,14,23,24,28,40]. Teleoperation and perception-driven interfaces provide scalable supervision and stress-test interface mismatch [4, 12, 16, 18, 19, 27, 50, 53, 54, 56]. Since reported results are sensitive to simulators, action interfaces, and termination rules, HumanTracker focuses on a standardized simulation evaluation protocol for meaningful comparison.

2.2 Datasets, retargeting, and benchmarking protocols

Scaling also depends on data: mocap repositories and internet-scale reconstructions provide broad coverage [11, 29, 34, 57], complemented by text-annotated resources and mocap-free synthesis [3, 15, 26, 61]. For humanoid tracking, however, usability hinges on physically plausible retargeting and contact consistency, motivating physics-constrained curation and interaction-preserving generation [1, 25, 49], as well as monocular/global 3D motion reconstruction with contact-aware refinement to reduce foot skating and drift [13, 41, 47, 55, 58]. Teleoperation pipelines increasingly collect robot-centric long-horizon behavior data [12, 18, 53, 54], but comparable progress still requires standardized evaluation protocols; HumanTracker provides reproducible evaluation and a unified `qpos` reference format for fair comparison.

2.3 Preference-based evaluation for physical plausibility

Kinematic metrics can miss temporally structured failures (contact timing, slip, drift); pairwise human preferences offer a direct supervision signal for reward modeling when metrics are misaligned with perception [9, 46]. HumanTracker collects preferences over synchronized rollout/reference videos and trains a trajectory reward model (HumanScore) as a preference-aligned metric for stability and contact realism in long-horizon tracking.

3 HumanTracker: Benchmark and Preference-Aligned Evaluation

HumanTracker targets two bottlenecks (Sec. 1). First, existing benchmarks are typically too small and too homogeneous to reveal *where* a humanoid tracker fails in long-horizon, contact-rich regimes. Second, conventional kinematic errors can disagree with what people perceive as stable and physically credible tracking, especially around contacts and support transitions. To address both, we introduce a large, explicitly categorized optical mocap dataset together with a standardized evaluation protocol, and we complement conventional metrics with a learned, preference-aligned trajectory score (*HumanScore*).

3.1 Evaluation protocol

Simulation setup. We evaluate a tracker by rolling it out in a physics simulator while it tracks a reference clip in simulation [6, 32, 33, 51]. The humanoid state is parameterized by generalized coordinates q and velocities $\dot{q}$; each clip provides a time-indexed robot-space reference q_t^{ref} (exported as `qpos`).

Standardization and reproducibility. All methods are evaluated under identical simulator settings, action bounds, initialization, and termination rules. We release the evaluation code and configuration to make the protocol reproducible; additional details and ablations are provided in Sec. 4.

3.2 HumanTracker dataset: 150 hours of high-quality optical mocap

Capture scale and fidelity. HumanTracker contains 150 hours of newly captured *optical* motion capture, collected specifically to benchmark contact-rich humanoid tracking. Optical mocap provides dense and accurate skeletal motion, which is particularly important for fast limb dynamics and subtle contact timing—precisely the regimes where small kinematic discrepancies can trigger large failures. The dataset comprises motions performed by 24 professional performers (e.g., dance teachers, fitness coaches, tennis coaches, and full-time mocap actors) to ensure both technical quality and stylistic breadth.

Quality control for physical plausibility. Raw recordings are not automatically "benchmark-ready" for physics-based tracking: small capture artifacts can translate into implausible contacts or destabilizing reference trajectories. We therefore apply a multi-stage quality control process with manual inspection, removing physically inconsistent segments such as unexplained floating, ground penetration, and contact discontinuities. This curation is crucial for fair evaluation because it prevents penalizing trackers for errors originating from the reference itself.

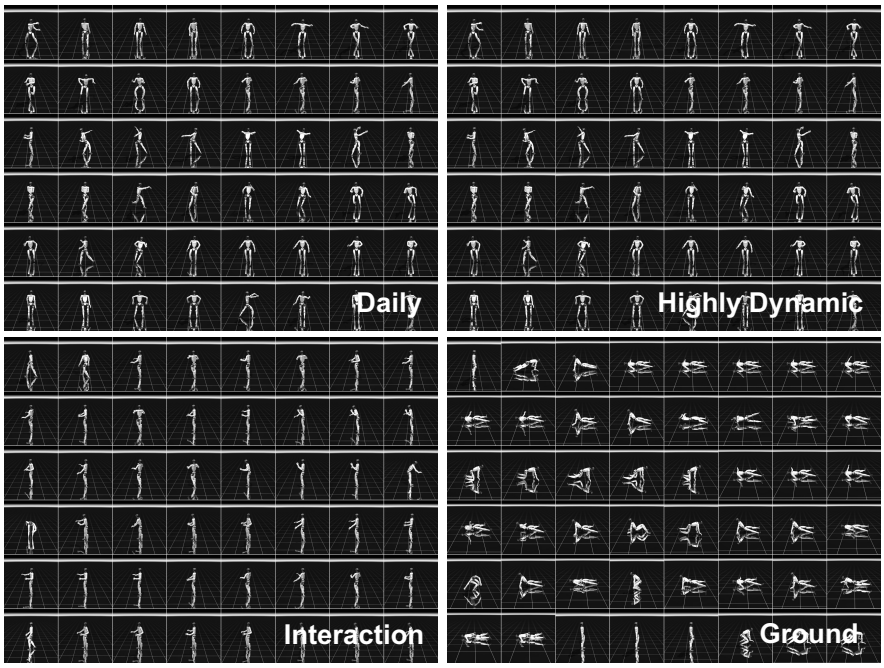


Fig. 2: Motion taxonomy. HumanTracker groups motions into four top-level families based on tracking regime and contact structure, enabling fine-grained diagnosis beyond a single aggregate score.

Annotations and released representations. Each clip is released with (i) a top-level motion family label, (ii) a natural-language text label, and (iii) a retargeted robot-space reference trajectory (`qpos`). We also provide fitted parametric human motion sequences (SMPL) for consistent downstream analysis [30,34,37]. Family labels support stratified evaluation and diagnosis (e.g., isolating failures under impacts versus support switching), while text labels make the benchmark directly queryable and enable fine-grained subset reporting beyond coarse aggregates.

Motion taxonomy for diagnostic evaluation. Trackers rarely fail uniformly: a method can appear strong on steady locomotion yet break under impacts, asymmetric support, or low-posture, multi-contact transitions. To avoid a misleading "average-case" benchmark, we group motions into four top-level families (Fig. 2) that separate tracking regimes by contact pattern, support complexity, and the dominant pathway through which errors amplify over time. Specifically, *Daily* covers routine walking/turning/gestures and primarily probes steady-state stability and residual drift under stable support; *Highly Dynamic* contains jumps, kicks, acrobatics, and fast dance footwork where impacts and rapid support switching amplify phase and timing errors; *Interaction* focuses on human/object

Table 2: Dataset statistics of HumanTracker. The benchmark contains 150 hours and 24K motion clips across four motion families, each representing distinct tracking challenges ranging from steady locomotion to highly dynamic and multi-contact motions.

Family	Hours	#Clips	Typical challenges
Daily	93	10.3k	steady locomotion, mild contacts
Highly Dynamic	10	1.6k	impacts, aerial phases, fast footwork
Interaction	43	11.6k	human/object interaction, hand–body coordination
Ground	4	1.6k	low posture, multi-contact transitions
Total	150	24000	diverse

interactions (e.g., pick & place, door opening, greetings) that require precise end-effector timing together with whole-body stabilization; and *Ground* emphasizes kneeling/sitting/rolling/recovery with low center-of-mass and complex multi-contact transitions where contact geometry and friction consistency are critical. This taxonomy makes evaluation *diagnostic*: per-family results reveal not only which method is better, but *where* it improves and which failure modes remain.

Train/test split. We provide an 8:2 clip-level split stratified by motion family, covering diverse regimes in both partitions while avoiding duplicate leakage. We evaluate trackers both in-domain (trained or fine-tuned on HumanTracker train) and zero-shot (without using HumanTracker), as detailed in Sec. 4.

3.3 Preference collection and HumanScore

From rollouts to preference supervision. To learn an evaluation signal aligned with human perception, we construct a preference dataset over tracking rollouts. We train a collection of trackers on the HumanTracker training split while varying training choices that are known to affect contact realism and long-horizon stability (e.g., reward parameterization, domain randomization strength, and the availability of privileged information during training). For each reference clip, we then generate rollouts using the standardized protocol and render synchronized videos.

Pair construction and sampling. Each query compares two rollouts that track the same reference segment and differ only in the tracker that generated them. We balance queries across motion families and over-sample "hard" pairs where conventional metrics disagree, since such comparisons are most informative for separating perceptual quality beyond frame-wise kinematics.

Annotation interface, labels, and quality control. Figure 3 shows the interface: each query samples a 5s window and presents two synchronized rollout videos for the same reference segment (same camera and playback). Annotators choose one of four labels—*left better*, *right better*, *both good*, or *both bad*—which

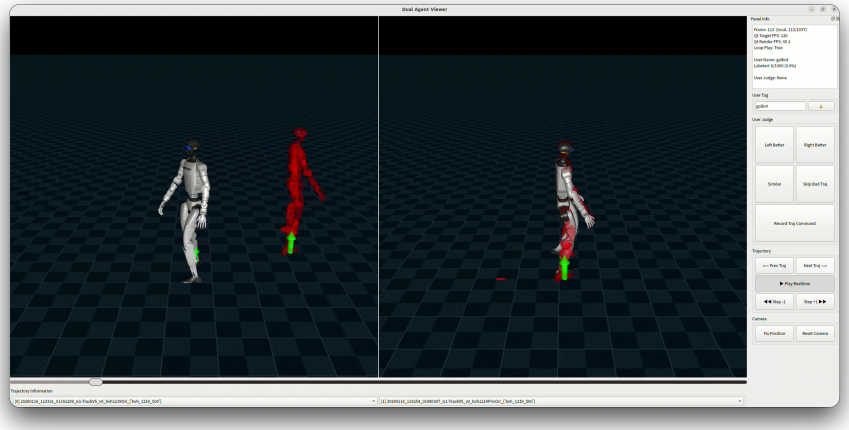


Fig. 3: Preference collection interface (in-house tool). Each query randomly samples a 5s segment and presents two synchronized rollout videos for the same reference, together with four options: *Left better*, *Right better*, *Both good*, and *Both bad*.

captures both strict preferences and common cases where both candidates are acceptable or neither is satisfactory. Annotators are instructed to judge (i) fidelity to the shown reference and (ii) physical credibility (contacts, balance, and smooth support transitions), rather than motion "style". We apply basic quality control (e.g., repeated queries and sanity checks) and filter inconsistent annotations before training the preference model.

Dataset size and split. We use 12K annotated motion pairs and an 8:2 split for preference-model training/testing. All quantitative validation of HumanScore is reported on the held-out preference test split (Sec. 4).

HumanScore model We train a trajectory reward model $r_\theta(\tau)$ that predicts human preferences and assigns a scalar HumanScore to a rollout trajectory τ. Although supervision is collected from videos, the model operates on the underlying time series to avoid visual confounds: each frame t is encoded as a token x_t that concatenates rollout information (including privileged signals available during evaluation, such as full-body state and contact-related quantities) with time-aligned reference features from q_t^{ref}. We use a bidirectional Transformer encoder [43] to capture long-horizon phenomena (e.g., slow drift, unstable support switching, and contact inconsistency), mean-pool the encoded temporal tokens, and map the pooled representation through an MLP head to obtain $r_\theta(\tau)$.

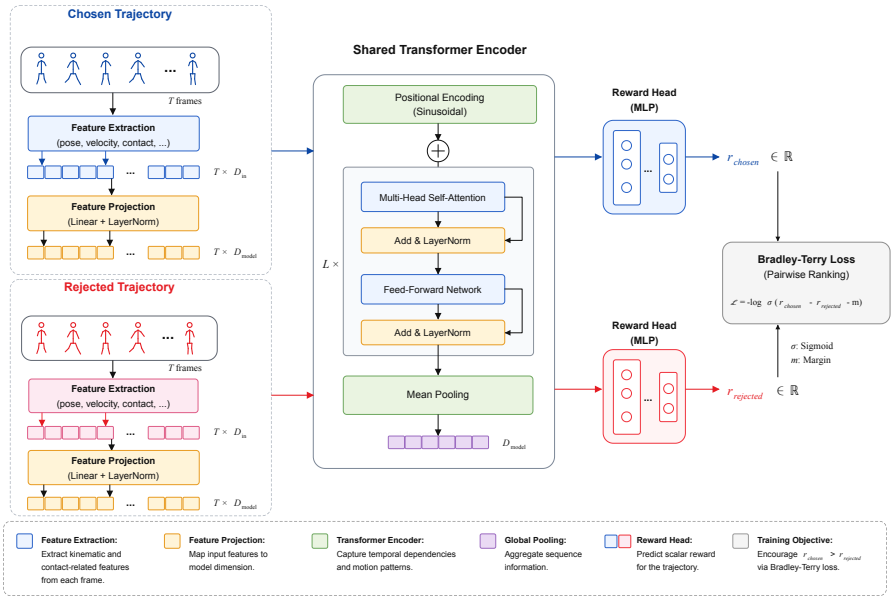


Fig. 4: Trajectory reward model (HumanScore). A bidirectional temporal Transformer encodes per-frame tokens, each containing rollout privileged states and reference information, and aggregates them via mean pooling to produce a global representation, which is mapped to a scalar score by an MLP head.

Training objective from four-way labels. For strict preferences (*left better* or *right better*), we use a Bradley–Terry likelihood [5, 9, 42]:

$$P(\tau_i \succ \tau_j) = \sigma\big(r_\theta(\tau_i) - r_\theta(\tau_j)\big),$$

and minimize the negative log-likelihood. For ties (*both good* and *both bad*), we enforce score consistency by penalizing large score differences between the two rollouts, and we use the tie type to calibrate score scale (good ties should receive higher absolute scores than bad ties). Implementation details (feature set, loss weights, and ablations) are provided in Sec. 4.

3.4 Metrics and evaluation with HumanScore

Reported metrics. We report three primary metrics: *Succ*, *MPJPE*, and *HumanScore*. *Succ* is the fraction of rollouts that finish a clip without termination under the standardized protocol. *MPJPE* is the mean per-joint pose error in radians (rad). *HumanScore* is computed from the raw reward-model output $\text{Score}^{\text{raw}}_{\text{RM}} = r_\theta(\tau)$, with higher values indicating a higher human preference. For benchmark tables, we use:

$$\text{HumanScore}_b = 100 \cdot \sigma\left(\frac{\text{Score}^{\text{raw}}_{\text{RM}} - \text{median}_b}{\text{scale}_b}\right),$$

where b denotes either the zero-shot setting or the in-domain training setting, and median_b and scale_b are computed once from the complete rollout sample pool of setting b. The factor 100 reports HumanScore on a percentage-style 0–100 scale. In our implementation, $\text{scale}_b = \max((q_{75,b} - q_{25,b})/(2\log 3), \epsilon)$, where $q_{25,b}$ and $q_{75,b}$ are the pooled first and third quartiles and $\epsilon = 10^{-8}$.

Preference alignment of metrics. To quantify alignment with human judgments, we evaluate each metric on the held-out preference test split. We report pairwise preference accuracy (predicting which rollout is preferred) and rank correlation across methods, together with the per-family breakdowns enabled by our motion taxonomy.

4 Experiments

HumanTracker makes three recurring patterns explicit by evaluating diverse motions in physics simulation under a standardized rollout protocol, consistent with how modern trackers are trained and evaluated [6, 32, 33, 51]. First, category composition matters: *High Dynamic* and *Ground* amplify contact-timing and recovery failures and therefore separate methods more clearly than comparatively easy locomotion. Second, metric mismatch is systematic: low MPJPE does not necessarily imply stable contacts or long-horizon consistency, especially in contact-rich families where retargeting and contact timing artifacts are visually salient [25, 49]. Third, HumanScore is more predictive of held-out human preferences than Succ and MPJPE in our setting, enabling more targeted diagnosis of the failures that conventional kinematic errors under-penalize.

4.1 Setup

We evaluate all methods under the same simulator configuration and episode protocol. We fix the control frequency, action bounds, and episode termination rules, and we use a shared action interface so that results reflect tracking quality

Table 3: In-domain preference test comparison between the reward model and traditional metrics. Align Rate denotes the probability of correctly predicting the preferred trajectory in a human-labeled pair.

Metric	**Align Rate**
Reward Model (HumanScore)	**0.9291**
MPJPE (rad)	0.8218
MPJVE (rad/s)	0.4770
KPT Position MAE (m)	0.5594
Foot Contact Accuracy	0.5307
Avg Joint Accel (rad/s^2)	0.6686
Avg Joint Jerk (rad/s^3)	0.7088

Table 4: HumanTracker benchmark (In-Domain Training). For each family we report Succ (%) ↑, MPJPE (rad) ↓, and HumanScore ↑ on a 0–100 scale.

Method	Daily			High Dynamic			Interaction			Ground		
	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score
TWIST2 [54]	88.9	0.149	18.9	77.6	0.174	21.4	95.3	0.125	32.2	0.0	**0.248**	**38.2**
Humanoid-GPT [39]	**97.3**	**0.057**	**71.0**	**95.1**	**0.085**	**63.1**	**97.5**	**0.053**	**69.5**	**17.4**	0.301	34.6

Table 5: HumanTracker benchmark (Zero-Shot). HumanScore is reported on a 0–100 scale.

Method	Daily			High Dynamic			Interaction			Ground		
	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score
Humanoid-GPT [39]	**97.2**	**0.060**	**72.0**	89.9	**0.102**	**64.0**	**97.6**	**0.051**	**70.9**	**19.2**	0.608	23.3
SONIC [33]	93.6	0.103	52.5	**92.3**	0.118	43.1	96.6	0.080	65.9	0.0	**0.260**	**35.3**
TWIST2 [54]	80.0	0.106	24.3	62.3	0.164	21.8	93.5	0.075	41.2	0.0	0.320	13.3
GMT [6]	21.1	0.244	12.7	30.7	0.250	16.2	76.7	0.151	40.2	3.0	0.436	15.7

rather than actuator modeling differences. We use the same episode termination rules as defined in Sec. 3.1. HumanTracker provides one train split and one test split with an 8:2 ratio. We report all benchmark numbers on the test split. We report an in-domain training setting and a zero-shot setting. In-domain training trains or fine-tunes a tracker on HumanTracker train. Zero-shot evaluates a tracker without using HumanTracker train.

4.2 Evaluated methods

We evaluate four representative trackers. GMT [6] is a general motion tracking policy. TWIST2 [54] is a scalable humanoid teleoperation and data collection system; we evaluate its tracking policy under the HumanTracker protocol. SONIC [33] scales motion tracking and reports strong zero-shot transfer. Humanoid-GPT [39] scales up large-scale motion tracking and uses a modern causal Transformer for better capacity. For each method, we follow the official evaluation protocol as closely as possible and adapt its control outputs to the HumanTracker shared evaluation interface.

4.3 Benchmark results on HumanTracker

Tab. 4 and Tab. 5 summarize performance across the four motion families. The per-family breakdown exposes different failure regimes: steady locomotion and mild contacts in *Daily*, fast impacts and phase errors in *High Dynamic*, precise end-effector timing under asymmetric support in *Interaction*, and low posture with complex multi-contact transitions in *Ground*.

In-domain training. With access to HumanTracker train (Tab. 4), the main separation occurs in contact-heavy families. *Daily*, *High Dynamic*, and *Interaction* are dominated by Humanoid-GPT, which combines high success, low MPJPE, and high HumanScore after in-domain training. *Ground* remains difficult even with in-domain training: Humanoid-GPT completes more rollouts, while TWIST2 receives a slightly higher HumanScore and lower MPJPE, reflecting the different contact geometry and support transitions required at low center-of-mass.

Zero-shot generalization. Without using HumanTracker for training (Tab. 5), performance gaps widen substantially. Generalization remains most challenging on *Ground*, where performance varies widely and most methods still fail to complete rollouts, and it remains difficult in *High Dynamic* for weaker trackers, reflecting sensitivity to action interfaces and unseen contact regimes. Humanoid-GPT shows the strongest zero-shot performance on *Daily* and *Interaction*, while SONIC remains competitive and obtains the highest HumanScore on *Ground*. Overall, *Ground* remains largely unsolved under zero-shot transfer despite these differences.

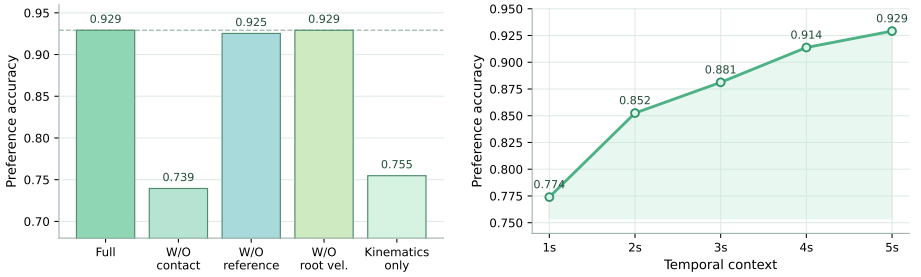
4.4 Preference-aligned evaluation with HumanScore

HumanScore vs. conventional diagnostics. We quantify alignment with human judgments on the held-out preference test split. As shown in Tab. 3, HumanScore achieves the highest alignment rate (0.9291), outperforming kinematic errors such as MPJPE as well as individual diagnostic signals such as foot-contact accuracy. This supports our motivation: perceptual tracking quality depends on *temporally structured* physical artifacts (contact timing, slip, drift, and support transitions) that are not faithfully summarized by frame-wise pose errors or any single handcrafted diagnostic.

Annotator reliability. To ensure the preference signal is stable, we measure inter-annotator agreement on 500 overlapping pairs. Annotators achieve a Cohen's Kappa of $\kappa = 0.76$ (95% CI: 0.71–0.81), indicating substantial agreement and validating the preference supervision used to train HumanScore. We further analyze model design choices via ablations in Sec. 4.5 (Fig. 5).

4.5 Ablation studies

We conduct ablations to understand which information HumanScore uses to predict human preferences. All ablations use the same 12K human-labeled motion pairs as the reward-model training protocol, without collecting additional labels or generating new rollouts. We use the same random 80/20 train/evaluation split with seed 42 and report pairwise preference accuracy on the held-out human preference split.



(a) Inference-time feature masking. **(b)** Temporal context length.

Fig. 5: HumanScore ablations on held-out human preference pairs. The full model is trained with all features; W/O denotes zeroing a feature group only at evaluation time.

Input ablation. For feature sensitivity, we train a full HumanScore model once and zero out selected feature groups only at evaluation time. This isolates which signals the trained model actually relies on. As shown in Fig. 5(a), removing contact-related features reduces Align Rate from 0.9291 to 0.7395, while keeping only kinematic pose features drops it to 0.7548. In contrast, removing reference features changes the score only slightly to 0.9253, and removing root-velocity features leaves the result unchanged at 0.9291. This indicates that HumanScore is not merely learning a kinematic pose-matching proxy; it relies strongly on physical and contact-related cues such as foot contact, foot motion, and contact stability.

Temporal ablation. Temporal context is also important. We truncate each 5s preference clip to the first 1s, 2s, 3s, 4s, or the full 5s window. As shown in Fig. 5(b), Align Rate improves monotonically from 0.7739 at 1s to 0.8525 at 2s, 0.8812 at 3s, 0.9138 at 4s, and 0.9291 at 5s. This supports scoring complete trajectory windows rather than isolated frames, since perceptual failures such as foot skating, jitter, instability, and near-fall events are temporally structured.

4.6 Optimization utility.

While HumanScore is primarily designed as an evaluation metric, a natural question is whether it can be used directly as a reward signal in reinforcement learning. In preliminary experiments, we incorporated HumanScore as a dense reward component during PPO training. We observed that the policy successfully learned to minimize visible artifacts like foot skating and jitter, leading to perceptually smoother motions. However, we also noted instances of reward hacking, where the policy adopted overly stiff postures to artificially inflate the score. This suggests that while HumanScore provides a valuable learning signal, it must be carefully balanced with task-specific rewards or regularized to prevent exploitation. We leave the full exploration of preference-based RL for humanoid tracking to future work.

Table 6: Reward model training hyperparameters.

Hyperparameter	Value
Layers	5
Attention heads	2
FFN dimension	1024
Pooling	Mean
Batch size	128
Learning rate	1×10^{-5}
Epochs	4
LR schedule	Cosine
Dropout	0.05
Weight decay	1×10^{-5}

4.7 Implementation details and dataset statistics.

We include additional implementation details from the supplementary material to make the camera-ready paper self-contained. HumanScore is evaluated on 5s windows: each rollout is divided into non-overlapping segments, and the reward model scores each segment independently. We first average the raw segment scores for the rollout, then apply the global sigmoid calibration from Sec. 3.4 so that the reported sequence-level HumanScore is on a 0–100 scale. This aggregation reduces temporal variance while preserving failures that persist over multiple seconds. Formally, for a rollout τ partitioned into N non-overlapping segments $\{s_i\}_{i=1}^{N}$, we compute

$$\mathrm{Score}_{\mathrm{RM}}^{\mathrm{raw}}(\tau) = \frac{1}{N} \sum_{i=1}^{N} r_\theta(s_i),$$

before applying the global setting-level normalization. The reward model itself is lightweight, using a five-layer Transformer encoder with two attention heads, mean pooling, and a feed-forward dimension of 1024. The final checkpoint used for all HumanScore results is fine-tuned for four epochs with batch size 128, learning rate 1×10^{-5}, dropout 0.05, weight decay 1×10^{-5}, and a cosine learning-rate schedule.

Table 7: Detailed HumanTracker dataset statistics by split and motion family. Durations are reported in hours before rounding in the main benchmark summaries.

Split statistics

Split	Clips	Frames	Hours
Train	19,716	52.16M	121.65
Test	4,926	12.71M	29.65
Overall	24,642	64.87M	151.31

Motion-family statistics

Family	Clips	Frames	Hours
Daily	10,304	40.49M	93.73
High Dynamic	1,630	4.31M	9.98
Interaction	11,068	18.58M	43.01
Ground	1,640	1.49M	4.59

5 Conclusions

This paper addressed two obstacles that currently limit progress in humanoid motion tracking: the misalignment between kinematic errors and perceived tracking quality under contacts, and the lack of reproducible, comparable evaluation protocols across methods. We introduced HumanTracker, a unified benchmark that evaluates tracking as control under a common and reproducible simulator setup, enabling fair cross-method comparison. HumanTracker also contributes 150 hours of newly captured optical motion data from 24 professional performers, including dance teachers and fitness coaches, curated into four motion families, complete with multimodal annotations including category labels, text labels, fitted SMPL sequences, and retargeted reference `qpos` trajectories.

To better reflect what humans judge as "tracks well," we used 12K synchronized rollout/reference motion pairs and trained a trajectory reward model whose output we report as HumanScore. Across representative trackers and settings, our results highlight systematic cases where low kinematic error does not imply stable contacts, and show that HumanScore is more predictive of held-out preferences than conventional metrics.

We hope HumanTracker will enable fairer comparison, clearer diagnosis, and ultimately tracking methods that are more robust and credible on real humanoids.

References

1. Araujo, J.P., Ze, Y., Xu, P., Wu, J., Liu, C.K.: Retargeting matters: General motion retargeting for humanoid motion tracking. ArXiv preprint **abs/2510.02252** (2025), `https://arxiv.org/abs/2510.02252`
2. Belkhale, S., Ding, T., Xiao, T., Sermanet, P., Vuong, Q., Tompson, J., Chebotar, Y., Dwibedi, D., Sadigh, D.: RT-H: action hierarchies using language. ArXiv preprint **abs/2403.01823** (2024), `https://arxiv.org/abs/2403.01823`
3. Bensabath, L., Petrovich, M., Varol, G.: A cross-dataset study for text-based 3d human motion retrieval. vol. abs/2405.16909 (2024), `https://arxiv.org/abs/2405.16909`
4. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: A vision-language-action flow model for general robot control. ArXiv preprint **abs/2410.24164** (2024), `https://arxiv.org/abs/2410.24164`
5. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
6. Chen, Z., Ji, M., Cheng, X., Peng, X., Peng, X.B., Wang, X.: Gmt: General motion tracking for humanoid whole-body control. ArXiv preprint **abs/2506.14770** (2025), `https://arxiv.org/abs/2506.14770`
7. Cheng, X., Ji, Y., Chen, J., Yang, R., Yang, G., Wang, X.: Expressive whole-body control for humanoid robots. ArXiv preprint **abs/2402.16796** (2024), `https://arxiv.org/abs/2402.16796`
8. Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. In: Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023 (2023)

9. Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 4299–4307 (2017), https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html
10. Dugar, P., Shrestha, A., Yu, F., van Marum, B., Fern, A.: Learning multi-modal whole-body control for real-world humanoid robots. In: Proceedings of the AAAI Symposium Series. vol. 7, pp. 650–657 (2025)
11. Fan, K., Lu, S., Dai, M., Yu, R., Xiao, L., Dou, Z., Dong, J., Ma, L., Wang, J.: Go to zero: Towards zero-shot motion generation with million-scale data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13336–13348 (2025)
12. Fu, Z., Zhao, Q., Wu, Q., Wetzstein, G., Finn, C.: Humanplus: Humanoid shadowing and imitation from humans **270**, 2828–2844 (2024)
13. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 14737–14748. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01358, https://doi.org/10.1109/ICCV51070.2023.01358
14. Gu, J., Kirmani, S., Wohlhart, P., Lu, Y., Arenas, M.G., Rao, K., Yu, W., Fu, C., Gopalakrishnan, K., Xu, Z., Sundaresan, P., Xu, P., Su, H., Hausman, K., Finn, C., Vuong, Q., Xiao, T.: Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=F1TKzG8LJO
15. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion inbetweening. ACM Transactions on Graphics (TOG) **39**(4), 60–1 (2020)
16. He, J., Li, D., Yu, X., Qi, Z., Zhang, W., Chen, J., Zhang, Z., Zhang, Z., Yi, L., Wang, H.: Dexvlg: Dexterous vision-language-grasp model at scale. ArXiv preprint **abs/2507.02747** (2025), https://arxiv.org/abs/2507.02747
17. He, T., Gao, J., Xiao, W., Zhang, Y., Wang, Z., Wang, J., Luo, Z., He, G., Sobanbab, N., Pan, C., et al.: Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. ArXiv preprint **abs/2502.01143** (2025), https://arxiv.org/abs/2502.01143
18. He, T., Luo, Z., He, X., Xiao, W., Zhang, C., Zhang, W., Kitani, K., Liu, C., Shi, G.: Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. ArXiv preprint **abs/2406.08858** (2024), https://arxiv.org/abs/2406.08858
19. He, T., Luo, Z., Xiao, W., Zhang, C., Kitani, K., Liu, C., Shi, G.: Learning human-to-humanoid real-time whole-body teleoperation. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8944–8951. IEEE (2024)
20. He, X., Dong, R., Chen, Z., Gupta, S.: Learning getting-up policies for real-world humanoid robots. ArXiv preprint **abs/2502.12152** (2025), https://arxiv.org/abs/2502.12152
21. Huang, T., Wang, H., Ren, J., Yin, K., Wang, Z., Chen, X., Jia, F., Zhang, W., Long, J., Wang, J., et al.: Towards adaptable humanoid control via adaptive motion tracking. ArXiv preprint **abs/2510.14454** (2025), https://arxiv.org/abs/2510.14454

22. Ji, M., Peng, X., Liu, F., Li, J., Yang, G., Cheng, X., Wang, X.: Exbody2: Advanced expressive humanoid whole-body control. ArXiv preprint **abs/2412.13196** (2024), https://arxiv.org/abs/2412.13196
23. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., Fagan, P.D., Hejna, J., Itkina, M., Lepert, M., Ma, Y.J., Miller, P.T., Wu, J., Belkhale, S., Dass, S., Ha, H., Jain, A., Lee, A., Lee, Y., Memmel, M., Park, S., Radosavovic, I., Wang, K., Zhan, A., Black, K., Chi, C., Hatch, K.B., Lin, S., Lu, J., Mercat, J., Rehman, A., Sanketi, P.R., Sharma, A., Simpson, C., Vuong, Q., Walke, H.R., Wulfe, B., Xiao, T., Yang, J.H., Yavary, A., Zhao, T.Z., Agia, C., Baijal, R., Castro, M.G., Chen, D., Chen, Q., Chung, T., Drake, J., Foster, E.P., et al.: DROID: A large-scale in-the-wild robot manipulation dataset. ArXiv preprint **abs/2403.12945** (2024), https://arxiv.org/abs/2403.12945
24. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: Openvla: An open-source vision-language-action model. ArXiv preprint **abs/2406.09246** (2024), https://arxiv.org/abs/2406.09246
25. Lee, K., Kim, S., Park, M., Kim, H., Hwang, D., Lee, H., Choo, J.: Phuma: Physically-grounded humanoid locomotion dataset. ArXiv preprint **abs/2510.26236** (2025), https://arxiv.org/abs/2510.26236
26. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023)
27. Li, Y., Lin, Y., Cui, J., Liu, T., Liang, W., Zhu, Y., Huang, S.: Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. ArXiv preprint **abs/2506.08931** (2025), https://arxiv.org/abs/2506.08931
28. Liao, Q., Truong, T.E., Huang, X., Tevet, G., Sreenath, K., Liu, C.K.: Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. ArXiv preprint **abs/2508.08241** (2025), https://arxiv.org/abs/2508.08241
29. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/4f8e27f6036c1d8b4a66b5b3a947dd7b-Abstract-Datasets_and_Benchmarks.html
30. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
31. Luo, Z., Cao, J., Merel, J., Winkler, A., Huang, J., Kitani, K.M., Xu, W.: Universal humanoid motion representations for physics-based control. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=OrOd8PxOO2
32. Luo, Z., Cao, J., Winkler, A., Kitani, K., Xu, W.: Perpetual humanoid control for real-time simulated avatars. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 10861–10870. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01000, https://doi.org/10.1109/ICCV51070.2023.01000

33. Luo, Z., Yuan, Y., Wang, T., Li, C., Chen, S., Castañeda, F., Cao, Z.A., Li, J., Minor, D., Ben, Q., et al.: Sonic: Supersizing motion tracking for natural humanoid whole-body control. ArXiv preprint **abs/2511.07820** (2025), https://arxiv.org/abs/2511.07820
34. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: archive of motion capture as surface shapes. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 5441–5450. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00554, https://doi.org/10.1109/ICCV.2019.00554
35. Mao, J., Zhao, S., Song, S., Shi, T., Ye, J., Zhang, M., Geng, H., Malik, J., Guizilini, V., Wang, Y.: Learning from massive human videos for universal humanoid pose control. ArXiv preprint **abs/2412.14172** (2024), https://arxiv.org/abs/2412.14172
36. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
37. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 10975–10985. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.01123, http://openaccess.thecvf.com/content_CVPR_2019/html/Pavlakos_Expressive_Body_Capture_3D_Hands_Face_and_Body_From_a_CVPR_2019_paper.html
38. Peng, X.B., Abbeel, P., Levine, S., Van de Panne, M.: Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions On Graphics (TOG) **37**(4), 1–14 (2018)
39. Qi, Z., Chen, X., Liu, D., Lin, C., Lian, Y., Liang, S., Zhang, Z., Guan, Y., Wang, J., Zhang, W., Yu, X., Wang, H., Yi, L.: Humanoid-gpt: Scaling data and structure for zero-shot motion tracking. ArXiv preprint **abs/2606.03985** (2026), https://arxiv.org/abs/2606.03985
40. Radosavovic, I., Zhang, B., Shi, B., Rajasegaran, J., Kamat, S., Darrell, T., Sreenath, K., Malik, J.: Humanoid locomotion as next token prediction. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024), http://papers.nips.cc/paper_files/paper/2024/hash/90afd20dc776bc8849c31d61a0763a0b-Abstract-Conference.html
41. Shin, S., Kim, J., Halilaj, E., Black, M.J.: WHAM: reconstructing world-grounded humans with accurate 3d motion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 2070–2080. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.00202, https://doi.org/10.1109/CVPR52733.2024.00202

42. Thurstone, L.L.: A law of comparative judgment. In: Scaling, pp. 81–92. Routledge (2017)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
44. Wang, Y., Yang, M., Zeng, W., Zhang, Y., Xu, X., Jiang, H., Ding, Z., Lu, Z.: From experts to a generalist: Toward general whole-body control for humanoid robots. ArXiv preprint **abs/2506.12779** (2025), https://arxiv.org/abs/2506.12779
45. Xie, W., Han, J., Zheng, J., Li, H., Liu, X., Shi, J., Zhang, W., Bai, C., Li, X.: Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills. ArXiv preprint **abs/2506.12851** (2025), https://arxiv.org/abs/2506.12851
46. Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., Zhang, T.: Iterative preference learning from human feedback: Bridging theory and practice for RLHF under kl-constraint. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net (2024), https://openreview.net/forum?id=c1AKcA6ry1
47. Xue, L., Guo, C., Zheng, C., Wang, F., Jiang, T., Ho, H.I., Kaufmann, M., Song, J., Hilliges, O.: Hsr: holistic 3d human-scene reconstruction from monocular videos. In: European Conference on Computer Vision. pp. 429–448. Springer (2024)
48. Yan, Y., Mascaro, E.V., Egle, T., Lee, D.: I-ctrl: Imitation to control humanoid robots through constrained reinforcement learning. ArXiv preprint **abs/2405.08726** (2024), https://arxiv.org/abs/2405.08726
49. Yang, L., Huang, X., Wu, Z., Kanazawa, A., Abbeel, P., Sferrazza, C., Liu, C.K., Duan, R., Shi, G.: Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. ArXiv preprint **abs/2509.26633** (2025), https://arxiv.org/abs/2509.26633
50. Yang, S., Du, Y., Ghasemipour, S.K.S., Tompson, J., Kaelbling, L.P., Schuurmans, D., Abbeel, P.: Learning interactive real-world simulators. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=sFyTZEqmUY
51. Yin, K., Zeng, W., Fan, K., Dai, M., Wang, Z., Zhang, Q., Tian, Z., Wang, J., Pang, J., Zhang, W.: Unitracker: Learning universal whole-body motion tracker for humanoid robots. ArXiv preprint **abs/2507.07356** (2025), https://arxiv.org/abs/2507.07356
52. Yin, S., Ze, Y., Yu, H.X., Liu, C.K., Wu, J.: Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. ArXiv preprint **abs/2509.20322** (2025), https://arxiv.org/abs/2509.20322
53. Ze, Y., Chen, Z., AraÃšjo, J.P., Cao, Z.a., Peng, X.B., Wu, J., Liu, C.K.: Twist: Teleoperated whole-body imitation system. ArXiv preprint **abs/2505.02833** (2025), https://arxiv.org/abs/2505.02833
54. Ze, Y., Zhao, S., Wang, W., Kanazawa, A., Duan, R., Abbeel, P., Shi, G., Wu, J., Liu, C.K.: TWIST2: Scalable, portable, and holistic humanoid data collection system. ArXiv preprint **abs/2511.02832** (2025), https://arxiv.org/abs/2511.02832

55. Zhang, S., Bhatnagar, B.L., Xu, Y., Winkler, A., Kadlecek, P., Tang, S., Bogo, F.: Rohm: Robust human motion reconstruction via diffusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 14606–14617. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01384, https://doi.org/10.1109/CVPR52733.2024.01384
56. Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Wang, H., Zhang, Z., Yi, L., Zeng, W., Jin, X.: Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. ArXiv preprint **abs/2507.04447** (2025), https://arxiv.org/abs/2507.04447
57. Zhang, Y., Lin, J., Zeng, A., Wu, G., Lu, S., Fu, Y., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. ArXiv preprint **abs/2501.05098** (2025), https://arxiv.org/abs/2501.05098
58. Zhang, Y., Zhang, H., Hu, L., Zhang, J., Yi, H., Zhang, S., Liu, Y.: Proxycap: Real-time monocular full-body capture in world space via human-centric proxy-to-motion learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 1954–1964. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.00191, https://doi.org/10.1109/CVPR52733.2024.00191
59. Zhang, Z., Chen, C., Xue, H., Wang, J., Liang, S., Liu, Y., Zhang, Z., Wang, H., Yi, L.: Unleashing humanoid reaching potential via real-world-ready skill space. ArXiv preprint **abs/2505.10918** (2025), https://arxiv.org/abs/2505.10918
60. Zhang, Z., Guo, J., Chen, C., Wang, J., Lin, C., Lian, Y., Xue, H., Wang, Z., Liu, M., Lyu, J., et al.: Track any motions under any disturbances. ArXiv preprint **abs/2509.13833** (2025), https://arxiv.org/abs/2509.13833
61. Zhang, Z., Li, Y., Huang, H., Lin, M., Yi, L.: Freemotion: Mocap-free human motion synthesis with multimodal large language models. In: European Conference on Computer Vision. pp. 403–421. Springer (2024)
62. Zhao, S., Ze, Y., Wang, Y., Liu, C.K., Abbeel, P., Shi, G., Duan, R.: Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. ArXiv preprint **abs/2510.05070** (2025), https://arxiv.org/abs/2510.05070
63. Zhao, W., Zhao, Y., Pajarinen, J., Muehlebach, M.: Bi-level motion imitation for humanoid robots. vol. abs/2410.01968 (2024), https://arxiv.org/abs/2410.01968