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,†], **Li Yi**[2,6,†]

[1]Nankai University [2]Tsinghua University [3]Galbot [4]Shanghai Jiao Tong University [5]Peking University
[6]Shanghai Qi Zhi Institute
[*]Equal Contribution, [†]Corresponding author

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, particularly unstable support and incorrect contacts such as foot skating and mistimed touch-downs. Meanwhile, widely used test suites are small and lack the diversity needed to stress contact-rich, long-horizon behaviors. We introduce HumanTracker to make humanoid tracking evaluation both perceptually aligned and scalable. HumanTracker contributes 150 hours of newly captured optical motion from multiple professional performers, organized into four motion families with text labels for fine-grained diagnosis. We further propose HumanScore, a preference-aligned metric trained from 6K human-labeled comparisons spanning 12K synchronized tracker trajectories 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.

Date: August 3, 2026
Project Page: https://dairuliu.github.io/humantracker

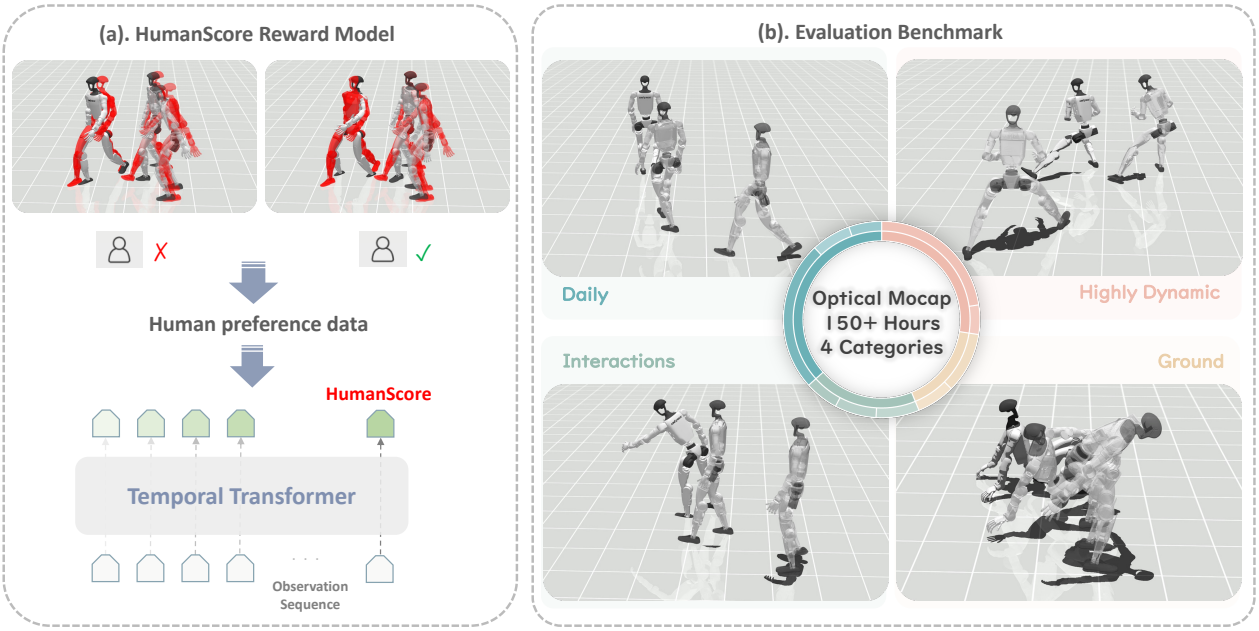


Figure 1 HumanTracker overview. Left: we train a Human-Aligned Reward Model on pairwise preference data from tracking rollouts, producing a scalar HumanScore via a temporal Transformer. Right: the HumanTracker Benchmark provides 150+ hours of motion capture across four families: daily tasks, highly dynamic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

1 Introduction

Motion tracking is becoming the cornerstone of humanoid robot control [15, 3, 17, 30, 12, 23, 26]. This ability drives applications in teleoperation and whole-body imitation [31, 32, 7, 6, 11, 8, 5]. Most systems track a reference motion using a learned feedback policy in physics simulation [15, 3, 17, 30]. 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 [3, 17, 30, 37, 22]. 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 [9, 29, 1, 28]. 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 [24, 35, 33]. These artifacts are the key aspect to decide whether the motion looks stable, smooth and human-like.

This gap is not a corner case but a consequence of what kinematic metrics measure. By treating tracking as a per-frame pose matching problem and averaging errors over joints and time, these metrics fail to capture contact, support, balance, and the accumulation of errors in closed-loop control. As a result, two rollouts can have similar pose errors yet differ substantially in stability. Observers reliably prefer the rollout 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 [9] and SONIC [17], the most commonly used evaluation suite for humanoid tracking is still an AMASS test set with only 140 sequences [18]. This small set lacks diversity and under-represents 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 compares HumanTracker with representative large-scale motion datasets in scale, categorization and text annotation.

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 [4, 19, 27]. 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.

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 [18]	>11K	>40	No	No
HumanML3D [2]	14.6K	28.6	No	Yes
PHUMA [9]	76K	73	No	No
HumanTracker	24K	**150**	**4**	**Yes**

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 [18, 13, 34, 9]. 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. We comprehensively evaluate several state-of-the-art trackers, including GMT, TWIST2, SONIC, and Humanoid-GPT, under our benchmark [3, 32, 17, 22].

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 Work

2.1 Humanoid motion tracking

Humanoid motion tracking is commonly formulated as reference-conditioned control in physics simulation, where a policy must reproduce a time-varying motion while satisfying the dynamics and contact constraints of the robot [21, 15, 16, 3, 17, 30]. Recent systems improve robustness and motion coverage through larger training corpora, stronger sequence models, residual correction, privileged training signals, and more expressive action interfaces [3, 17, 22, 30, 38, 5, 28, 37]. The same capability also underpins whole-body teleoperation and imitation systems, in which human motion is mapped to a deployable humanoid policy [31, 32, 7, 6, 11, 8]. These advances have produced increasingly capable trackers, but their reported performance is difficult to compare because results depend not only on the learned policy, but also on the reference set, simulator, action representation, initialization, and termination rule. HumanTracker therefore treats evaluation as a controlled experiment: tracker-specific policy interfaces are preserved, whereas the reference representation, rollout accounting, and reported metrics are standardized.

2.2 Motion data and contact-aware evaluation

Large motion repositories such as AMASS, Motion-X and Motion-X++ have substantially expanded the diversity of human motion available for learning [18, 13, 34], while text-annotated datasets support semantic retrieval and conditional generation [2, 36, 10]. For humanoid tracking, however, dataset size alone does not define a useful benchmark. Human motion must be retargeted to the robot morphology, remain physically plausible around contacts, and contain sufficiently difficult regimes to expose controller failures. Physics-aware retargeting and reconstruction methods consequently devote particular attention to foot skating, global drift, penetration and inconsistent support [9, 29, 1, 24, 35, 33]. Existing tracking evaluations nevertheless remain concentrated on comparatively small test sets and often collapse heterogeneous motions into one aggregate value. HumanTracker complements prior motion sources with a large, explicitly categorized test bed whose four families separate steady locomotion, rapid impact-rich motion, environmental interaction and ground-level multi-contact transitions.

2.3 Preference-based evaluation

Pairwise preference learning is useful when the quality of a structured output is easier to recognize than to specify as a fixed analytic objective [4, 19, 27]. This distinction is particularly relevant to humanoid tracking: per-frame joint errors describe reference deviation, but observers also respond to temporally accumulated failures such as jitter, delayed contact, foot sliding and loss of support. HumanTracker applies preference learning at the trajectory level. Annotators compare synchronized rollouts of the same reference segment, and a temporal reward model learns a scalar score from the corresponding reference and simulated-state sequences. Unlike a video-quality model, HumanScore is evaluated directly on simulator trajectories; unlike an individual handcrafted diagnostic, it can integrate pose, contact, force and velocity evidence across the complete motion window.

3 HumanTracker: Benchmark and Preference-Aligned Evaluation

HumanTracker is designed around the two limitations identified in the Introduction. The benchmark supplies enough motion diversity to expose failures that are hidden by a small aggregate test set, while HumanScore supplies a trajectory-level measurement that is trained to reproduce human judgments of tracking quality. The two components share the same unit of analysis: a reference motion is retargeted to the humanoid, executed by a tracking policy in simulation, and evaluated from the resulting rollout. This section describes the benchmark data and protocol first, and then the construction and training of HumanScore.

3.1 The HumanTracker benchmark

Motion collection and processing. HumanTracker contains 150 hours of newly captured optical motion from 24 professional performers, including dance teachers, fitness coaches, tennis coaches and full-time motion-capture actors. The performers and recording plan were chosen to cover both routine movement and motions that place substantially different demands on a humanoid controller. Because a visually valid human recording is not automatically a valid robot reference, we inspect the retargeted sequences and remove segments with capture or processing artifacts such as unexplained floating, ground penetration and discontinuous contacts. Each released clip contains a top-level motion-family label, a natural-language

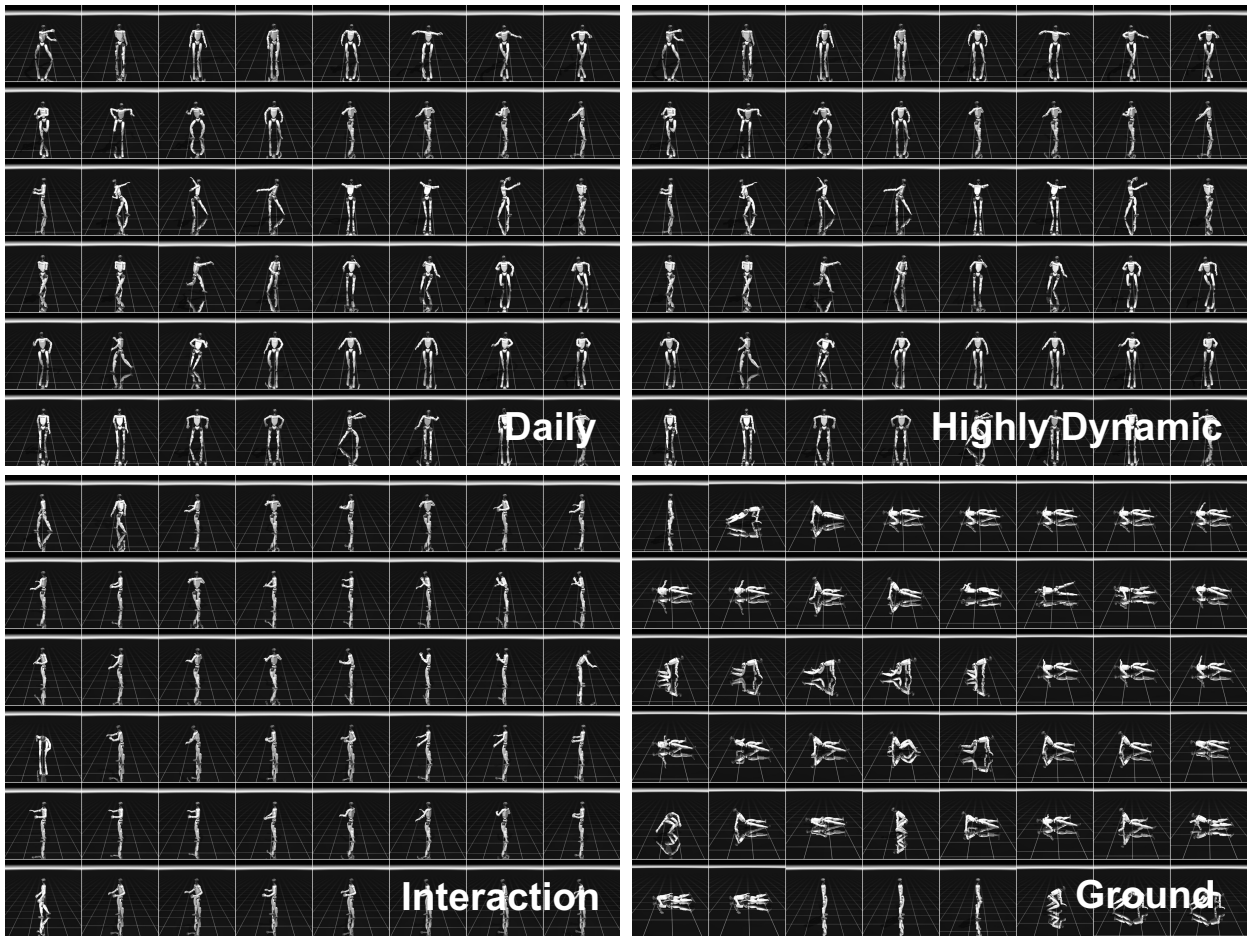


Figure 2 Motion taxonomy. HumanTracker groups motions into four families according to their dominant tracking and contact regimes. The taxonomy supports category-level diagnosis rather than only a single aggregate score.

description, a fitted SMPL sequence and a robot-space reference trajectory in `qpos` format [14, 20, 18]. These representations support semantic subset selection while keeping the input to every evaluated tracker identical. Figure 2 illustrates the resulting motion coverage, and Table 2 summarizes the family-level scale of the benchmark.

A taxonomy for diagnostic evaluation. The four motion families are defined by the failure regimes that they expose. *Daily* contains walking, turning and routine gestures, and therefore measures steady-state stability and residual drift under comparatively regular support. *Highly Dynamic* contains jumps, kicks, acrobatics and fast dance footwork, for which impacts and rapid support switching amplify phase and timing errors. *Interaction* contains motions involving objects or other people, which require end-effector timing together with whole-body stabilization. *Ground* covers kneeling, sitting, rolling and recovery, where a low centre of mass and multiple simultaneous contacts make the controller sensitive to contact geometry and friction. The distribution intentionally reflects the frequency of the captured activities rather than forcing equal family sizes; all benchmark results are therefore reported by family as well as in aggregate. The complete dataset is split 8:2 into disjoint training and test partitions, with the family distribution preserved and duplicate motions kept within one partition.

3.2 Standardized tracker evaluation

For every method, we convert the reference to the same 29-DoF humanoid `qpos` representation and execute the policy through a common MuJoCo evaluation entry point. Each tracker retains its native policy observations and action decoder, because replacing these components would change the learned controller; the evaluator instead standardizes the motion list, robot model, reference indexing, rollout accounting and metric implementation. The control trajectory is recorded at 50 Hz. At every step, the evaluator stores the simulated generalized position

Table 2 HumanTracker dataset statistics. The benchmark contains approximately 150 hours and 24K clips across four complementary tracking regimes.

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, coordination between hands and body
Ground	4	1.6k	low posture, multi-contact transitions
Total	150	24K	diverse

and velocity, policy action and motor target, foot contacts and contact forces, foot and pelvis velocities, and 14 keypoint poses and spatial velocities. The same state history is used for conventional diagnostics and HumanScore, which prevents differences in post-processing from being mistaken for differences between trackers.

An episode is successful only if the tracker reaches the final reference frame without triggering the shared SONIC termination criterion. We report *Succ* as the fraction of completed episodes and *MPJPE* as the mean absolute error over the 29 actuated joint angles, in radians, over the executed portion of each rollout. Additional diagnostics include joint-velocity error, keypoint-position error, foot-contact agreement and finite-difference joint acceleration and jerk. We evaluate two settings. In-domain trackers are trained or fine-tuned on the HumanTracker training split; zero-shot trackers are evaluated without using that split. All reported benchmark values use the held-out HumanTracker test split.

3.3 Preference data construction

Rollout segmentation for paired comparison. The preference pool is generated exclusively from motions in the HumanTracker training split, so no benchmark test motion is used to train HumanScore. For each source motion, GMT, Humanoid-GPT, SONIC and TWIST2 produce aligned rollouts of the same robot-space reference [3, 22, 17, 32]. At 50 Hz, every rollout is divided into consecutive 250-frame windows, each spanning 5 s. If a rollout ends with fewer than 250 frames, the final short window is retained rather than discarded. Windows are admitted only when all four trackers provide an aligned candidate for the same source-frame interval; this ensures that a preference reflects tracker behaviour rather than a change in reference content.

Uniform and tracker-balanced pair sampling. We concatenate all aligned windows in a deterministic order by motion family, source motion and temporal position, and assign each window a global index. From this catalogue we select a uniformly spaced set of unique indices over the full ordered list. This construction samples the complete HumanTracker training distribution without manually favouring visually difficult examples or any particular family. Each selected window yields exactly one comparison between two of the four trackers. The six unordered tracker combinations are allocated equally, so every pairing contributes the same number of comparisons and each tracker appears equally often. Candidate order is alternated within each combination and the final task order is deterministically shuffled, so no tracker is favoured by display position. The resulting preference pool is partitioned into training and held-out sets by source motion.

Human annotation and split. Using the interface shown in Figure 3, annotators first decide whether either rollout fails to complete the motion or loses balance; when both remain viable, they compare jitter, foot sliding, locomotion consistency and whole-body naturalness in that order. They choose *Similar* when neither candidate is meaningfully better and *Cannot compare* when both are unusable or the evidence is insufficient. The interface randomizes display order and records the underlying candidate indices, eliminating a fixed association between display side and tracker identity. The collected labels comprise strict preferences, similar pairs and cannot-compare pairs. Cannot-compare pairs are excluded from reward-model optimization. We construct an 80/20 training/test split with seed 42 by grouping records by the original `motion_id`; all clips from one source motion therefore remain in one partition. The split is balanced jointly over family, tracker pairing, label type, annotator and full-versus-short clip status, and the remaining records define the training and held-out preference sets.

3.4 HumanScore

Trajectory representation. Figure 4 summarizes how HumanScore maps a rollout segment τ to a scalar $r_\theta(\tau)$. Although annotations are collected from rendered videos, the model consumes simulator trajectories so that its prediction is not affected by camera

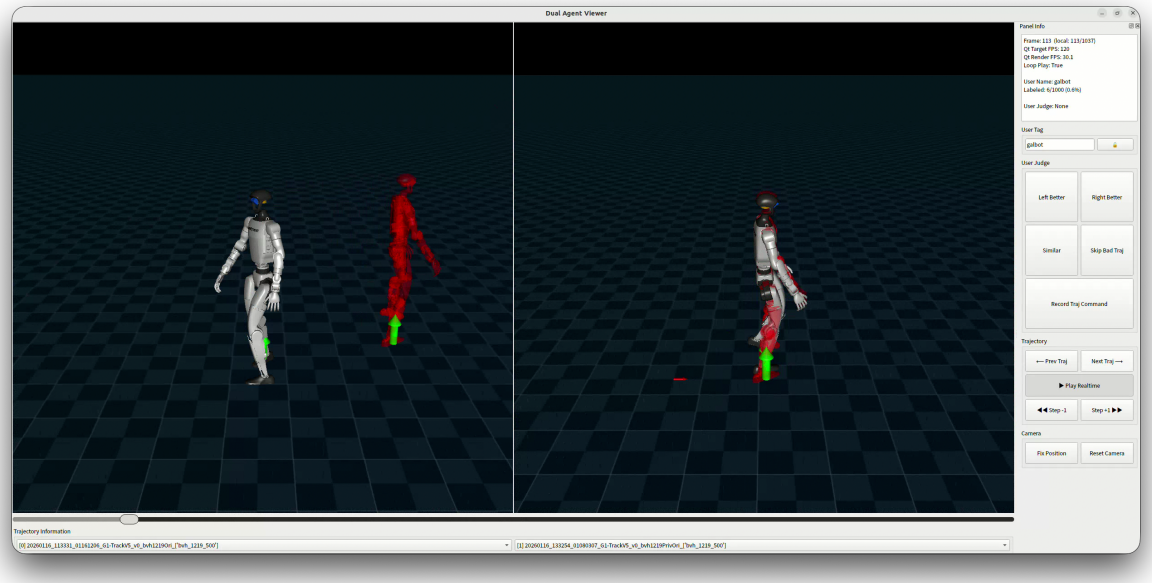


Figure 3 Preference collection interface. Annotators view two synchronized 5 s rollouts of the same reference segment. Display order is randomized and the available responses are *Left better*, *Right better*, *Similar* and *Cannot compare*.

or rendering choices. Each frame is represented by an 850-dimensional vector. Seventy reference dimensions encode the root pose and navigation velocity, 29 joint positions and velocities, and two reference foot-contact states. The remaining 780 dimensions encode the simulated root and IMU poses, action and motor target, joint state, foot contact, force, velocity and acceleration, pelvis and navigation velocities, 14 keypoint poses and spatial velocities in the gravity-aligned frame, and the residual between the next reference and the current simulated state. The exact feature decomposition is given in the supplementary material.

The per-frame vector is linearly projected, normalized and combined with sinusoidal positional encoding before being processed by a bidirectional Transformer encoder [25]. A padding mask is passed to every self-attention layer, and the encoded valid tokens are averaged with the same mask before an MLP produces the scalar score. This detail is essential for the retained tail clips: a segment of $L < 250$ frames is right-padded with zeros to 250 frames, while a Boolean mask marks only positions $1, \dots, L$ as valid. Padding can therefore neither participate in attention nor change the denominator of temporal pooling.

Preference objective. For a strict comparison in which trajectory τ_a is preferred to τ_b, we define the pairwise preference probability as

$$P(\tau_a \succ \tau_b) = \sigma\left(\frac{r_\theta(\tau_a) - r_\theta(\tau_b)}{T}\right),$$

where T is a temperature. We optimize binary cross-entropy on the score difference, using target 1 for a strict preference and target $1/2$ for a *Similar* pair; an optional tie weight controls the contribution of the latter. Preferred candidates are placed consistently in the first input position during training, whereas similar pairs retain their recorded order. Model selection uses preference accuracy on a motion-disjoint validation subset of the training partition, with validation loss as a tie-breaker; the held-out preference test set is evaluated only after selecting the checkpoint.

3.5 From segment scores to benchmark metrics

At evaluation time, a rollout is divided into consecutive 250-frame windows using the same keep-tail rule as the annotation pipeline. Each window is scored independently; a short final window is right-padded and masked exactly as during training. For a rollout τ with windows $\{s_i\}_{i=1}^{N}$, the raw sequence score is

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

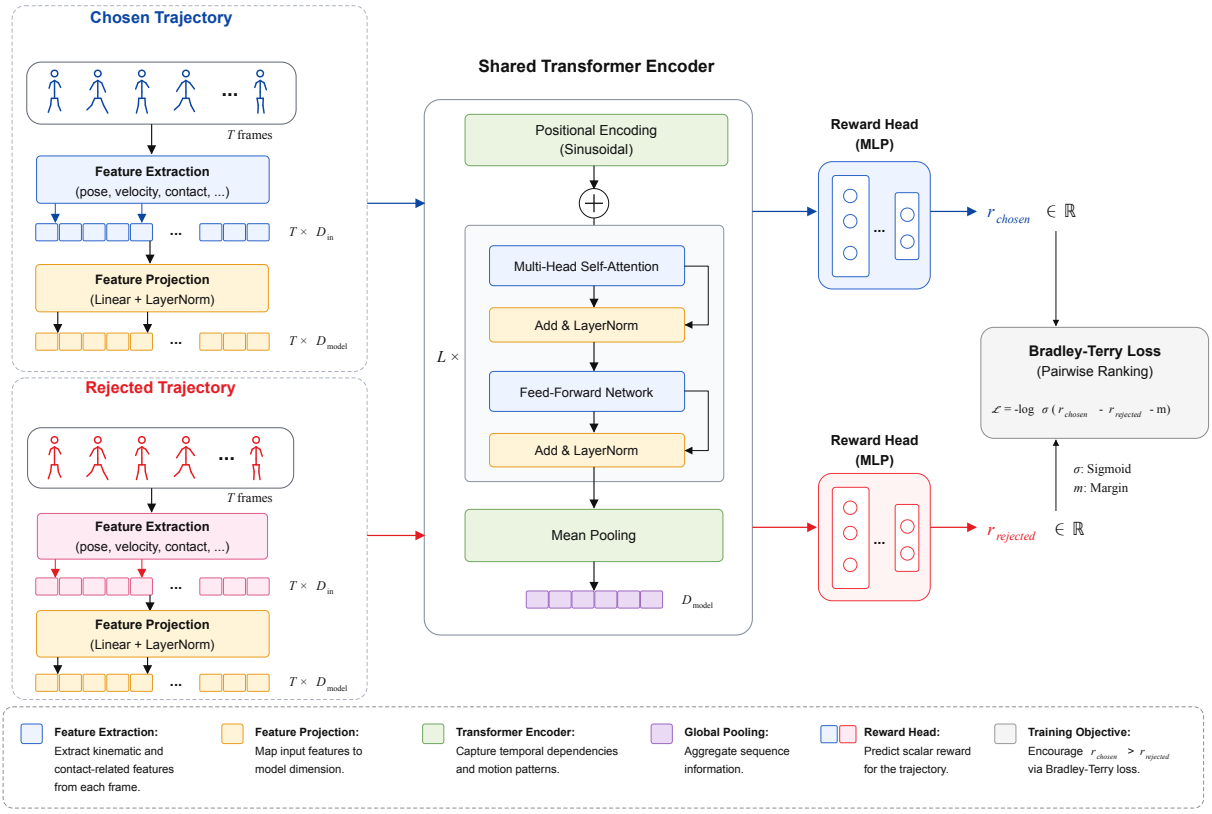


Figure 4 HumanScore trajectory model. Reference and simulated-state features form one token per frame. A bidirectional temporal Transformer processes the valid tokens, mask-aware mean pooling forms a trajectory representation, and an MLP maps it to a scalar score.

We convert raw scores to a readable scale from 0 to 100 without changing their ordering:

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

$$\mathrm{scale}_b = \max\left(\frac{q_{75,b} - q_{25,b}}{2\log 3}, 10^{-8}\right).$$

Here b denotes the in-domain or zero-shot benchmark setting, and the median and quartiles are computed once from the pooled rollouts of all methods and families in that setting. Calibration is never performed separately for a tracker or family. To measure alignment with people, a metric predicts the preferred candidate by comparing its value on the two trajectories of each strict held-out pair; error metrics are sign-inverted so that the preferred direction is consistent.

4 Experiments

Our experiments address three questions that follow directly from the Introduction. First, does a large category-aware benchmark expose tracker behaviour that a single aggregate test set would obscure? Second, does HumanScore predict held-out human choices more accurately than conventional kinematic and physical diagnostics? Third, which information and temporal context account for that alignment? We answer these questions using the four HumanTracker motion families, two tracker-training regimes and a motion-disjoint preference test set.

4.1 Experimental setup

Trackers and evaluation settings. We evaluate GMT, TWIST2, SONIC and Humanoid-GPT through the standardized protocol of Sec. 3.2 [3, 32, 17, 22]. GMT and SONIC are general-purpose tracking policies with an emphasis on broad motion coverage; TWIST2 supplies the tracking component of a scalable whole-body teleoperation system; and Humanoid-GPT uses a causal Transformer to scale motion-conditioned control. Each policy retains its native observation and action-processing stack, while all methods receive the same retargeted reference motions and are measured by the same evaluator. We use the SONIC

Table 3 HumanTracker benchmark (In-Domain Training). For each family we report Succ (%) ↑, MPJPE (rad) ↓, and HumanScore ↑ on a 0–100 scale.

Method	Daily			Highly Dynamic			Interaction			Ground		
	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score
TWIST2 [32]	30.1	0.150	32.6	38.7	0.177	26.3	70.8	0.128	30.0	**14.3**	**0.243**	**31.1**
Humanoid-GPT [22]	**74.3**	**0.071**	**76.8**	**57.7**	**0.130**	**54.8**	**93.1**	**0.055**	**75.0**	4.0	0.334	14.9

termination criterion for every tracker. The zero-shot comparison includes all four trackers without training on HumanTracker. The in-domain comparison includes TWIST2 and Humanoid-GPT, for which we train policies from scratch on the HumanTracker dataset. Every benchmark number is computed on the HumanTracker test split and is reported separately for *Daily*, *Highly Dynamic*, *Interaction* and *Ground*.

Metrics. We use Succ to denote completion rate and report it alongside joint-space MPJPE and HumanScore as complementary measurements. Succ captures catastrophic loss of tracking but cannot distinguish the quality of two completed rollouts. MPJPE measures local reference fidelity but averages over time and joints. HumanScore measures trajectory-level preference and is designed to respond to contact, stability and smoothness as well as pose. We evaluate HumanScore in consecutive 5-second windows and retain a final shorter window using right-zero padding and a validity mask. For the preference study, we additionally compare joint-velocity error, keypoint-position error, foot-contact agreement, and mean joint acceleration and jerk. A lower error predicts the preferred trajectory when its value is smaller; for foot-contact accuracy and HumanScore, a larger value predicts the preference.

4.2 Results

The category breakdown reveals a consistent asymmetry: improvements on common upright motion do not transfer uniformly to impact-rich and ground-level behaviour. Tables 3 and 4 report all three metrics for every motion family. The results should therefore be read across metrics within a family, rather than by using any one column as a universal ranking.

As shown in Table 3, Humanoid-GPT outperforms TWIST2 on *Daily*, *Highly Dynamic* and *Interaction* across Succ, MPJPE and HumanScore. *Ground* reverses this ordering across all three metrics: TWIST2 reaches 14.3% Succ, 0.243 MPJPE and 31.1 HumanScore, compared with 4.0%, 0.334 and 14.9 for Humanoid-GPT. This divergence is informative rather than contradictory: success measures whether a rollout reaches the end, while MPJPE and HumanScore describe the frames that were actually generated. Reporting the three together prevents a low error on short failed trajectories from being interpreted as robust tracking.

Table 4 shows a stronger dependence on motion family in the zero-shot setting. Humanoid-GPT leads all three metrics on *Daily*. On *Highly Dynamic*, SONIC obtains the highest Succ and HumanScore, whereas Humanoid-GPT has the lowest MPJPE. Humanoid-GPT leads all three metrics on *Interaction*. *Ground* remains difficult: Humanoid-GPT reaches the highest Succ at 33.8% and the lowest MPJPE at 0.183, but SONIC receives the highest HumanScore. A single aggregate score would dilute these differences because *Ground* is the smallest family. The family-wise view instead identifies low-posture, multi-contact transitions as a concrete target for future tracker design.

4.3 Does HumanScore align with people?

We evaluate each metric on strict preferences from the motion-disjoint HumanScore test partition. As reported in Table 5, HumanScore predicts 90.97% of held-out choices. The strongest individual diagnostic, keypoint-position MAE, reaches 84.76%, followed by joint-velocity error at 84.66%. Contact accuracy also aligns substantially with preference at 82.78%, but remains below HumanScore. These results show that no individual frame-aggregated diagnostic captures the joint contribution of reference fidelity, contact timing, stability and temporal accumulation used by annotators. The 6.21-point gap over the strongest conventional diagnostic supports the metric-misalignment motivation in the Introduction.

Because the test split is grouped by source motion, this result measures generalization to unseen HumanTracker training motions rather than memorization of adjacent clips from the same sequence. It also evaluates tracker combinations that are balanced by construction: each of the six pairings contributes approximately one sixth of the split. HumanScore is therefore not rewarded merely for learn-

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

Method	Daily			Highly Dynamic			Interaction			Ground		
	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score	Succ (%)	MPJPE (rad)	Human Score
GMT [3]	19.1	0.243	6.4	31.3	0.251	7.6	79.3	0.150	25.4	0.3	0.447	4.2
TWIST2 [32]	58.1	0.106	26.6	50.0	0.174	18.5	90.1	0.076	43.8	0.0	0.330	4.0
SONIC [17]	92.1	0.103	66.6	**83.1**	0.121	**51.3**	96.4	0.082	60.4	21.0	0.233	**41.8**
Humanoid-GPT [22]	**93.8**	**0.045**	**73.0**	75.8	**0.080**	50.2	**97.1**	**0.045**	**63.5**	**33.8**	**0.183**	26.6

Table 5 Alignment with held-out strict human preferences. Align Rate is the fraction of pairs for which a metric assigns the better value to the human-preferred trajectory. Acceleration and jerk exclude six pairs shorter than their finite-difference stencil.

Metric	Align Rate	Pairs
HumanScore	**0.9097**	1916
MPJPE (rad)	0.8173	1916
MPJVE (rad/s)	0.8466	1916
KPT Position MAE (m)	0.8476	1916
Foot Contact Accuracy	0.8278	1916
Avg Joint Accel (rad/s^2)	0.6901	1910
Avg Joint Jerk (rad/s^3)	0.7152	1910

ing the frequency of a dominant tracker matchup. A stronger claim about a new tracker family would require collecting rollouts from a tracker excluded from preference-model training, which we leave as a separate generalization test.

4.4 Ablation Study

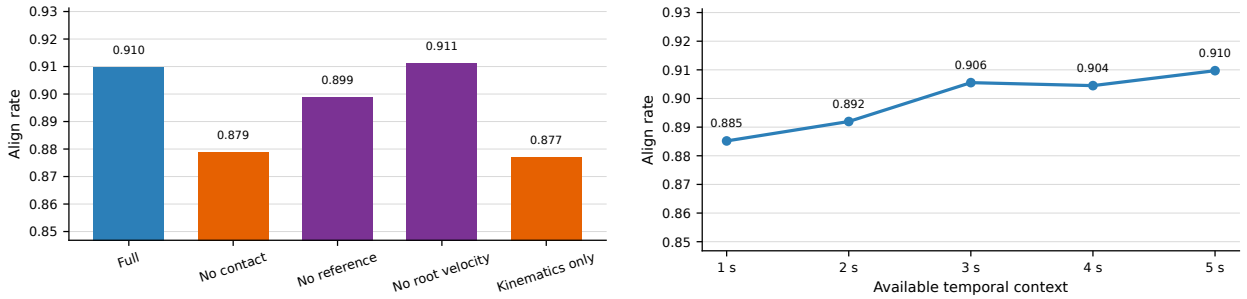
Figure 5 summarizes inference-time feature and temporal sensitivity analyses of the trained model. Zeroing all contact-related inputs reduces alignment from 0.9097 to 0.8789, and retaining only pose-kinematic inputs yields 0.8768. Removing reference features reduces alignment to 0.8987. Masking root-velocity features changes the score by only 0.16 points, from 0.9097 to 0.9113. These interventions do not retrain the model and should not be interpreted as architecture ablations; they identify which evidence the fitted score function depends on. The contact-feature decrease, together with the 0.8278 alignment of foot-contact accuracy alone, indicates that HumanScore uses contact information jointly with the rest of the trajectory rather than treating contact agreement as a sufficient scalar proxy.

Temporal context supplies complementary evidence. With the first 1, 2, 3, 4 and 5 seconds of a preference window, alignment is 0.8852, 0.8920, 0.9055, 0.9045 and 0.9097, respectively. The trend improves overall, with a small 0.10-point fluctuation between three and four seconds; the final second improves alignment by 0.52 points. This is consistent with the target phenomena: a single frame can reveal pose error, whereas foot sliding, repeated jitter, progressive drift and near-fall recovery require an interval over which their evolution can be observed.

5 Discussion and Limitations

The benchmark and preference results point to the same conclusion from different directions. Category-level evaluation shows that contact regime changes the relative behaviour of trackers: a method that is reliable on upright daily motion can still fail almost completely during ground-level transitions. The preference experiment explains why this distinction is not fully represented by joint error. HumanScore gains most of its advantage when it can observe contact-related state over several seconds, suggesting that the perceptual unit of failure is often an event such as a slide, impact, support switch or recovery, rather than an isolated pose. HumanScore should therefore be read alongside success and kinematic error, not as a replacement for all analytic diagnostics: the former summarizes perceived trajectory quality, whereas the latter remain valuable for locating a specific source of error.

Several boundaries remain. HumanScore is trained from HumanTracker training motions and rollouts produced by four trackers, so its motion-disjoint test measures generalization to unseen motions, not to an entirely unseen robot, simulator or controller family. Its 850-dimensional input also includes privileged simulator state and contact quantities that are available for benchmarking but may not be observable on hardware; applying the metric to real-world trajectories will require an observable feature set or a separately validated state estimator. The preference pool assigns one primary judgment to each pair, which provides broad coverage but does not quantify uncertainty through repeated independent labels. Finally, the



(a) Inference-time feature masking.

(b) Available temporal context.

Figure 5 HumanScore sensitivity on held-out preference pairs. The full model is trained once with all features. Panel a masks feature groups only at inference; panel b restricts the available temporal prefix.

present metric comparison covers representative individual diagnostics; broader tests against fitted linear and nonlinear diagnostic composites would further delimit what must be learned from trajectory-level preference data.

HumanTracker also inherits the limits of its data distribution. The four families are deliberately diagnostic but imbalanced, with far fewer Ground clips than Daily or Interaction clips, and the evaluation uses one 29-DoF humanoid embodiment in MuJoCo. Results should not be interpreted as a population estimate over all human activity or as evidence of hardware robustness. Extending the benchmark to additional embodiments, real-robot rollouts and rare contact regimes is therefore a natural next step. We also restrict HumanScore to evaluation in this work. Directly optimizing a learned score can create behaviours that exploit model imperfections; using it as a reinforcement-learning reward will require explicit regularization and an independent human evaluation rather than assuming that benchmark alignment transfers automatically to policy optimization.

6 Conclusion

HumanTracker addresses the two obstacles identified in the Introduction: conventional tracking errors are misaligned with perceived physical quality, and existing evaluation sets are too limited to diagnose the long tail of humanoid motion. The benchmark contributes 150 hours of curated motion from 24 professional performers, organized into four families and evaluated through a common rollout and metric pipeline. Its category-level results reveal that progress on daily and interaction motion does not imply robust tracking of highly dynamic or ground-level contacts; in particular, Ground remains difficult even for the strongest evaluated trackers.

HumanScore complements this benchmark with a trajectory metric learned from synchronized human comparisons. Its preference pool is sampled from the HumanTracker training split using aligned 5-second rollouts and balanced comparisons among four trackers, while motion-disjoint data splits and padding masks preserve both generalization testing and short tail clips. On held-out preferences, HumanScore reaches 0.9097 alignment, exceeding the strongest conventional diagnostic by 6.21 percentage points. Feature and temporal sensitivity analyses further show that this advantage depends on contact-related evidence accumulated across the motion window. Together, the dataset, protocol and human-aligned metric make tracker comparison more reproducible and failure analysis more specific, while leaving cross-embodiment, hardware and reward-optimization validation as important directions for future work.

References

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

[2] Léore Bensabath, Mathis Petrovich, and Gül Varol. A cross-dataset study for text-based 3d human motion retrieval. volume abs/2405.16909, 2024. https://arxiv.org/abs/2405.16909.

[3] Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. Gmt: General motion tracking for humanoid whole-body control. *ArXiv preprint*, abs/2506.14770, 2025. https://arxiv.org/abs/2506.14770.

[4] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg,

Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307, 2017. https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

[5] Pranay Dugar, Aayam Shrestha, Fangzhou Yu, Bart van Marum, and Alan Fern. Learning multi-modal whole-body control for real-world humanoid robots. In *Proceedings of the AAAI Symposium Series*, volume 7, pages 650–657, 2025.

[6] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. 270:2828–2844, 2024.

[7] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *ArXiv preprint*, abs/2406.08858, 2024. https://arxiv.org/abs/2406.08858.

[8] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8944–8951. IEEE, 2024.

[9] Kyungmin Lee, Sibeen Kim, Minho Park, Hyunseung Kim, Dongyoon Hwang, Hojoon Lee, and Jaegul Choo. Phuma: Physically-grounded humanoid locomotion dataset. *ArXiv preprint*, abs/2510.26236, 2025. https://arxiv.org/abs/2510.26236.

[10] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

[11] Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. *ArXiv preprint*, abs/2506.08931, 2025. https://arxiv.org/abs/2506.08931.

[12] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *ArXiv preprint*, abs/2508.08241, 2025. https://arxiv.org/abs/2508.08241.

[13] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *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.

[14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[15] Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 10861–10870. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01000. https://doi.org/10.1109/ICCV51070.2023.01000.

[16] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. 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.

[17] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *ArXiv preprint*, abs/2511.07820, 2025. https://arxiv.org/abs/2511.07820.

[18] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 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*, pages 5441–5450. IEEE, 2019. doi: 10.1109/ICCV.2019.00554. https://doi.org/10.1109/ICCV.2019.00554.

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *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.

[20] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. doi: 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.

[21] 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 (TOG)*, 37(4):1–14, 2018.

[22] Zekun Qi, Xuchuan Chen, Dairu Liu, Chenghuai Lin, Yunrui Lian, Sikai Liang, Zhikai Zhang, Yu Guan, Jilong Wang, Wenyao Zhang, Xinqiang Yu, He Wang, and Li Yi. Humanoid-gpt: Scaling data and structure for zero-shot motion tracking. *ArXiv preprint*, abs/2606.03985, 2026. https://arxiv.org/abs/2606.03985.

[23] Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. Humanoid locomotion as next token prediction. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *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.

[24] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. 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*, pages 2070–2080. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00202. https://doi.org/10.1109/CVPR52733.2024.00202.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[26] Weiji Xie, Jinrui Han, Jiakun Zheng, Huanyu Li, Xinzhe Liu, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills. *ArXiv preprint*, abs/2506.12851, 2025. https://arxiv.org/abs/2506.12851.

[27] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 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.

[28] Yashuai Yan, Esteve Valls Mascaro, Tobias Egle, and Dongheui Lee. I-ctrl: Imitation to control humanoid robots through constrained reinforcement learning. *ArXiv preprint*, abs/2405.08726, 2024. https://arxiv.org/abs/2405.08726.

[29] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omnire-target: 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.

[30] Kangning Yin, Weishuai Zeng, Ke Fan, Minyue Dai, Zirui Wang, Qiang Zhang, Zheng Tian, Jingbo Wang, Jiangmiao Pang, and Weinan Zhang. Unitracker: Learning universal whole-body motion tracker for humanoid robots. *ArXiv preprint*, abs/2507.07356, 2025. https://arxiv.org/abs/2507.07356.

[31] Yanjie Ze, Zixuan Chen, JoÃĢo Pedro AraÃšjo, Ziang Cao, Xue Bin Peng, Jiajun Wu, and C Karen Liu. Twist: Teleoperated whole-body imitation system. *ArXiv preprint*, abs/2505.02833, 2025. https://arxiv.org/abs/2505.02833.

[32] 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. *ArXiv preprint*, abs/2511.02832, 2025. https://arxiv.org/abs/2511.02832.

[33] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. 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*, pages

14606–14617. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01384. https://doi.org/10.1109/CVPR52733.2024.01384.

[34] Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. *ArXiv preprint*, abs/2501.05098, 2025. https://arxiv.org/abs/2501.05098.

[35] Yuxiang Zhang, Hongwen Zhang, Liangxiao Hu, Jiajun Zhang, Hongwei Yi, Shengping Zhang, and Yebin Liu. 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*, pages 1954–1964. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00191. https://doi.org/10.1109/CVPR52733.2024.00191.

[36] Zhikai Zhang, Yitang Li, Haofeng Huang, Mingxian Lin, and Li Yi. Freemotion: Mocap-free human motion synthesis with multimodal large language models. In *European Conference on Computer Vision*, pages 403–421. Springer, 2024.

[37] Zhikai Zhang, Jun Guo, Chao Chen, Jilong Wang, Chenghuai Lin, Yunrui Lian, Han Xue, Zhenrong Wang, Maoqi Liu, Jiangran Lyu, et al. Track any motions under any disturbances. *ArXiv preprint*, abs/2509.13833, 2025. https://arxiv.org/abs/2509.13833.

[38] Siheng Zhao, Yanjie Ze, Yue Wang, C Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. 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.

A HumanScore implementation details

A.1 Segment construction and padding

All preference and evaluation rollouts are represented at 50 Hz. We split a sequence into consecutive 250-frame windows and retain the final window when fewer than 250 frames remain. A full window is therefore 5 s long. For a retained tail of $L < 250$ frames, the feature tensor is right-padded with zeros to shape 250×850, and a Boolean validity mask contains ones for the first L frames and zeros elsewhere. After any configured temporal downsampling, the corresponding downsampled mask is passed to the Transformer as its key-padding mask and to mean pooling as its validity mask. If h_t is the encoded token at time t, the pooled representation is

$$\bar{h} = \frac{\sum_{t=1}^{250} m_t h_t}{\max(\sum_{t=1}^{250} m_t, 1)},$$

where $m_t \in \{0, 1\}$. Thus, retaining a short segment increases data coverage without introducing artificial zero frames into either temporal attention or the trajectory representation.

At benchmark time, the same rule is applied to a complete rollout. If its N windows are $\{s_i\}_{i=1}^{N}$, including a possible masked tail, we compute the raw rollout score as

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

The robust sigmoid calibration in Sec. 3.5 is then applied once at the evaluation-setting level. Because both averaging and calibration are monotonic with respect to a uniform increase in segment scores, the reported scale from 0 to 100 is interpretable across methods within the same setting while preserving the ordering induced by the reward model.

A.2 Per-frame features

Table 6 reports the complete feature decomposition. Reference features occupy 70 dimensions and rollout features occupy 780 dimensions. All rollout quantities are computed from the simulated robot rather than copied from the reference. Foot contact and force are obtained from contacts between the robot and floor; foot acceleration is the temporal derivative of the measured foot velocity; and keypoint features use 14 bodies expressed in a gravity-aligned navigation frame. The next-reference residuals encode the difference between the next desired state and the current simulated state, allowing the model to distinguish a transient deviation that is being corrected from one that is growing.

Table 6 HumanScore feature decomposition for each frame.

Group	Contents	Dim.
Reference	root pose/navigation velocity; joint position/velocity; foot contact	70
Robot state and action	root and IMU pose; action; motor target; joint position/velocity	126
Contact and local motion	foot contact/force/velocity/acceleration; pelvis and root velocities	35
Current keypoints	14 gravity-frame 4×4 poses and 6D spatial velocities	308
Next-reference residual	root velocity and 14 keypoint pose/velocity residuals	311
Total	concatenated per-frame token	850

A.3 Reward-model optimization

The reported model linearly projects each 850-dimensional token, applies layer normalization and sinusoidal positional encoding, and processes the sequence with a pre-normalized bidirectional Transformer. A mask-aware temporal mean is followed by a three-layer MLP score head. Table 7 lists the configuration used for the reported HumanScore checkpoint. We optimize strict preferences and similar pairs jointly: after ordering a strict pair as preferred/non-preferred, binary cross-entropy is applied to the temperature-scaled score difference with target 1; a similar pair uses target 0.5. Cannot-compare labels are excluded. AdamW and a cosine learning-rate schedule are used for optimization, and checkpoint selection is based on validation preference accuracy with validation loss as a tie-breaker.

B Preference data statistics

The pair catalogue is built from non-flipped motions in the HumanTracker training split. All aligned 5-second and retained tail clips are concatenated in the fixed family order *Highly Dynamic*, *Daily*, *Ground* and *Interaction*, followed by source-motion and temporal order. The sampler selects 6,000 midpoint-uniform positions from this catalogue. Each of the six tracker combinations contributes exactly 1,000 pairs, and each tracker therefore occurs in 3,000 pairs; the pool contains 12,000 tracker trajectories in total. The 6,000 comparisons contain 4,808 strict prefer-

Table 7 Hyperparameters of the reported HumanScore checkpoint.

Hyperparameter	Value
Transformer layers	5
Attention heads	2
FFN dimension	1024
Pooling	masked mean
Maximum sequence length	250 frames
Batch size	128
Learning rate	1×10^{-5}
Epochs	4
LR schedule	cosine
Dropout	0.05
Weight decay	1×10^{-5}
Padding	right zero + validity mask

ences, 949 similar judgments and 243 cannot-compare judgments. After excluding the last group, 4,609 comparisons are available in the training partition and 1,148 in the test partition. The 80/20 split groups by source `motion_id` and uses seed 42; its balancing objective includes motion family, tracker pair, label type, annotator, clip length and the number of sampled clips from the same source motion.

C HumanTracker dataset statistics

HumanTracker source recordings have native rates of 120 Hz for the *Daily*, *Highly Dynamic* and *Interaction* families and 90 Hz for *Ground*; the evaluation representation is converted to the common 50 Hz robot trajectory described above. The family sizes reflect the captured activity distribution, with more *Daily* and *Interaction* motion and fewer *Highly Dynamic* and *Ground* sequences. Tables 8 and 9 give the unrounded counts and durations used to form the summaries in the main paper.

Table 8 HumanTracker statistics by benchmark split.

Split	Clips	Frames	Duration (h)
Train	19,716	52,157,880	121.65
Test	4,926	12,711,300	29.65
Overall	24,642	64,869,180	151.31

Table 9 HumanTracker statistics by motion family.

Family	Clips	Frames	Duration (h)
Daily	10,304	40,489,540	93.73
Highly Dynamic	1,630	4,313,190	9.98
Ground	1,640	1,486,558	4.59
Interaction	11,068	18,579,892	43.01