Towards Motion Turing Test: Evaluating Human-Likeness in Humanoid Robots
Supplementary Material

	Source	#Clips	Raw Data (s)	Sanitatized Data (s)	Remarks
Humanoid (Real)	WRC, WAIC, WHRG 2025	257	9,600	1,285	11 robot models
Humanoid (Simulation)	LAFAN1-Retargeting Dataset	243	62,198	1,215	1 robot model (Unitree G1). 7 actions
Human (Captured)	10 Volunteers	365 **(50 imitated, 315 normal)**	5,003	1,305	Indoor&outdoor recordings
Human (Online)	YouTube Videos	135	1,200	1,195	Diverse demographics
Total	**--**	**1000**	**78,001**	**5,000**	**15 motion categories**

Figure S1. **HHMotion dataset.** The left table summarizes the full data composition, including all collection sources spanning real-world humanoid robots, simulated environments, and human motion recordings. The right lists every humanoid robot included in HHMotion.

Thanks for reading the supplementary. In this supplementary file, we provide additional technical details for the Evaluating Human-Likeness in Humanoid Robots framework. Moreover, a demo video is included in the supplementary to illustrate the overall pipeline and showcase representative evaluation examples.

A. More Details on HHMotion Dataset

A.1. Dataset Composition and Sources

A.1.1. HHMotion dataset sources

The HHMotion dataset integrates motion data from a wide spectrum of human and humanoid behaviors to construct a unified benchmark for evaluating human-likeness. It comprises five complementary motion sources, each contributing distinct characteristics and distribution patterns:

1) Real-World Humanoid Robots. Videos were collected from major public events and online releases of state-of-the-art humanoid robots in 2025. The public international events include the World Artificial Intelligence Conference (WAIC) [1], World Robot Conference (WRC) [3], World Humanoid Robot Games (WHRG) [2], which represent the current state of the art in humanoid robotics. These sequences capture genuine hardware constraints such as mechanical latency, limited joint ranges, ground interaction noise, and stability control behaviors.

2) Simulated Humanoid Robots. To expand the diversity of robot motions, we include high-quality simulated sequences. These clips often exhibit smoother trajectories and reduced noise compared to real robots, providing a valuable contrast for analyzing comprehensive discrepancies in human-likeness score.

3) Human Motions from Volunteers. Ten subjects with varied movement habits performed the same set of action categories as the robots. These sequences serve as the reference for natural human movement, capturing subtle cues such as inertia continuity and posture fluidity.

Figure S2. Representative examples of human and humanoid robot motions across four action categories. Full visualizations of all 15 categories are provided on the last page of the supplementary.

4) Humans Imitating Humanoid Robots subset. To introduce ambiguous and challenging cases for human-likeness evaluation, the volunteers also performed robot motions. Participants mimic the stiffness of humanoid robots. These sequences blur the boundary between human and robotic motion and increase the challenge of evaluation scenarios.

5) Additional Human motions from YouTube. To further enhance diversity, we curate additional human action videos from YouTube, covering a range of performance styles and outdoor scenes. These samples enrich motion variety and

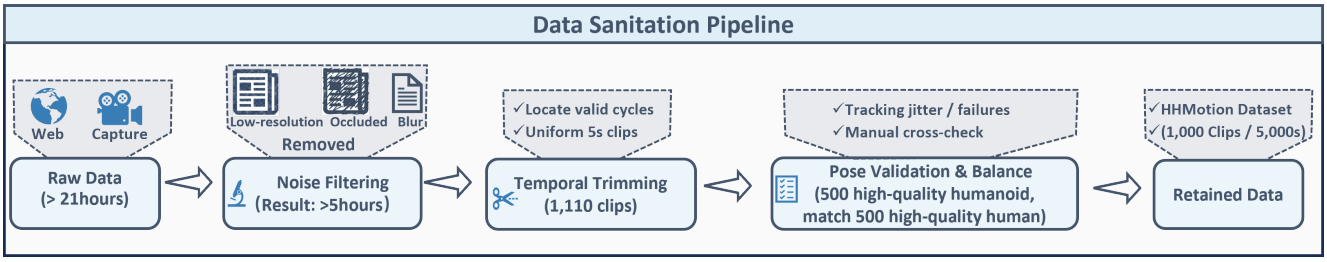

Figure S3. **The Data Sanitation Pipeline of HHMotion.** We illustrate the rigorous multi-stage pipeline used to construct the dataset from over 21 hours of raw footage.

help prevent overfitting to controlled laboratory environments.

Across these five sources, HHMotion incorporates 11 distinct humanoid robot models, as shown in Fig. S1, offering a comprehensive foundation for evaluating human-likeness in motion dataset.

A.1.2. Motion Categories in HHMotion

The HHMotion dataset spans 15 diverse action categories covering full-body, lower-limb, and upper-limb motion patterns essential for assessing human-likeness. In this section, we present only four representative categories to illustrate the contrast between human and humanoid motion in Fig. S2. The complete visualization of all 15 action categories for both humans and humanoid robots is provided in Fig. S8 and Fig. S9.

These 15 motion categories jointly cover slow, fast, cyclic, and fine-grained limb actions, ensuring that the dataset captures a broad spectrum of kinematic characteristics essential for measuring human-likeness across both humans and humanoid robots.

A.2. Data Sanitation Pipeline

To ensure the reliability of the benchmark and address the domain gap challenge when generalizing HPE methods to humanoid robots, we implemented a rigorous, multi-stage data sanitation pipeline. Starting with over 21 hours of raw footage, we processed the data through the steps as shown in Fig. S3.

1) Visual Quality Filtering and Standardization. We first performed strict **noise filtering** over the entire 21-hour raw corpus. Videos exhibiting low resolution, severe occlusion, incomplete subject cropping, strong compression artifacts, or motion blur were removed. This stage eliminated a substantial portion of low-quality content, resulting in approximately **5 hours** of visually reliable videos. We then conducted manual **temporal trimming** on this 5-hour subset to extract semantically complete action cycles. Each identified motion segment was normalized to a uniform 5-second duration, producing a pool of **1,110 candidate clips** across both human and humanoid sources.

2) Rigorous Pose Validation via Manual Inspection. A major challenge lies in ensuring that human-oriented HPE models (e.g., GVHMR) generalize reliably to humanoid robots. To mitigate domain-shift artifacts, we applied a strict **pose validation** stage to all 1,110 candidate clips. We reconstructed SMPL-X poses for every clip and performed frame-by-frame manual inspection to detect common failure cases, including tracking jitter, mesh distortions, and physiologically implausible poses. Following this quality screening, we strictly enforced **class balance** to prevent evaluation bias. We retained the top **500 high-quality humanoid clips** that passed inspection and matched them with an equal volume of verified high-quality human clips. This process ensures a consistent action distribution across both domains while guaranteeing that all included sequences demonstrate stable kinematic reconstruction. Please refer to details regarding the HPE methods in Sec. A.3.

After this comprehensive sanitation pipeline, we obtained approximately 5,000 seconds of clean, high-quality motion data (comprising 1,000 clips), establishing a robust foundation for the Motion Turing Test.

A.3. Data Processing and Pose Estimation

To obtain reliable 3D body parameters for all clips in HHMotion, we establish a unified data processing pipeline centered on SMPL-X reconstruction. Multiple state-of-the-art (SOTA) Human Pose Estimation (HPE) models are evaluated, and only high-quality reconstructions are retained through human cross-validation.

A.3.1. Manual Bounding Box Adjustment

Although modern HPE pipelines include automated person detection, raw detections often suffer from loose boxes, missing extremities, or frame-to-frame jitter, which can significantly degrade downstream SMPL-X fitting quality. To address these issues, we incorporate a refinement stage using the CVAT [15] annotation platform for efficient bounding-box correction. Annotators adjust boxes only on frames where automated detection clearly fails, such as rapid limb motion causing partial cropping, humanoid robot structures confusing the detector, or multi-person scenes

Figure S4. **Manual Adjustment of Bounding Boxes.** Illustration of the effect of manual correction on automatically detected bounding boxes. The upper shows raw detector outputs, which may include loose framing. The lower shows manually refined bounding boxes that ensure tighter alignment with the humanoid.

Figure S5. **Annotated humanoid key joints for SMPL-X evaluation.** Representative examples of the 2D skeletal annotations used to assess SMPL-X reconstruction quality on humanoid robots.

leading to identity switches. This targeted correction ensures stable and consistent tracking across all clips while keeping annotation costs manageable. The tracking effect before and after optimizing the bounding box is shown in Fig. S4.

A.3.2. Comparison results of the HPE methods

To systematically evaluate the robustness of existing SMPL-X reconstruction methods on humanoid robots, we curated 50 representative robot videos in the HHMotion dataset. For each video, annotators manually labeled **14 key joints** using CVAT [15], including the **nose, neck, shoulders, elbows, wrists, hips, knees, and ankles,** to provide a ground-truth 2D skeletal reference for evaluating pose accuracy in humanoid. Fig. S5 shows the representative examples of the labeled humanoid's 14 key joints.

Table S1. **Comparison of state-of-the-art HPE methods on our annotated robot motion set.** GVHMR is chosen as the SMPL-X reconstruction method for HHMotion since it achieves the lowest MPJPE and PAMPJPE (Unit: *pixel*).

Method	MPJPE ↓	PAMPJPE ↓	PCK@0.3 ↑
TRACE [13]	64.30	39.47	0.9384
WHAM [10]	66.13	28.42	0.9283
PromptHMR [16]	57.73	25.38	0.9150
GVHMR [9]	**28.23**	**25.33**	**0.9925**

Using these keypoint annotations, we tested several SOTA SMPL-X recovery methods and computed standard metrics such as MPJPE and PAMPJPE. The experimental results are shown in Tab. S1. The comparison shows that most HPE methods exhibit varying degrees of degradation on humanoid robots due to non-human limb proportions, rigid body articulation, or reflective materials. Among all evaluated methods, GVHMR [9] demonstrates the most stable performance, achieving the lowest MPJPE and the most consistent reconstruction across different robot forms. This method is therefore adopted as the primary SMPL-X reconstruction backend for the HHMotion dataset.

A.4. Annotation Quality and Analysis

In this work, we invited 30 volunteers to annotate the motion human-likeness score for each clip. To ensure the reliability of human-likeness annotations in HHMotion, we conducted a rigorous Inter-Annotator Consistency (IAC) analysis across all 30 annotators, following best practices widely adopted in **psychometrics**, **subjective quality assessment**, and **human rating studies** [6, 7, 11, 12, 14]. These communities routinely rely on error-to-consensus deviation (e.g., MSE) combined with rank-based agreement metrics (e.g., Spearman ρ) to quantify rating stability, especially for subjective perceptual judgments where absolute values and relative ordering are both meaningful.

In our setup, each annotator rated 1,000 clips. As shown in Alg. 1, we first computed the consensus trajectory by taking the clip-wise mean score, which is a standard and statistically robust aggregation strategy when the number of annotators is sufficiently large. We then evaluated each annotator using two complementary reliability indicators:

- **Absolute Deviation (MSE):** We compute the Mean Squared Error (MSE) to measure the numerical precision of the annotator, which is commonly used in video quality assessment (VQA) benchmarks to identify systematic bias or overly noisy raters [8]. This metric effectively

Table S2. **Summary of annotator reliability based on Inter-Annotator Consistency (IAC).** High-risk annotators (MSE $\geq$ 1.41 and Spearman $\rho \leq 0.50$) were removed from the final rating pool.

Annotator Group	Count	Mean MSE	Mean Spearman ρ
Retained Annotators	25	0.9764	0.5457
Removed Annotators	5	2.2081	0.4002
Total	30	1.1817	0.5215

detects evaluators who have a biased offset (e.g., consistently rating too high or too low) or high variance compared to the group.

- **Ranking Consistency (Spearman's ρ):** We employ Spearman's rank correlation coefficient to evaluate whether the annotator correctly orders the relative human-likeness of different clips, which is widely used in sports judging [17], and pairwise preference modeling [14].

Following established annotation filtering practice in VQA and human-rating literature [8, 12], we determined the filtering thresholds using the 75th percentile of MSE and the 25th percentile of Spearman correlation. Annotators exceeding both thresholds simultaneously were labeled as **high-risk**, as this indicates they diverge from consensus both in absolute terms and in relative ranking. Medium-risk annotators violate only one of the two criteria and were retained but flagged for analysis.

The analysis identified 5 annotators as high-risk (MSE $\geq$ 1.41 and $\rho \leq 0.50$). As summarized in Tab. S2, removing these outliers significantly improved the dataset's internal consistency, with the mean MSE dropping from 1.1817 to 0.9764 and the mean Spearman correlation increasing from 0.5215 to 0.5457. This validation step confirms that our final annotations represent a stable and reliable consensus on human-likeness.

Algorithm 1: Inter-Annotator Consistency (IAC) Filtering Algorithm.

Input: Score matrix $S \in \mathbb{R}^{M \times N}$ where s_{ij} is annotator j's score on clip i.
Output: Retained annotator set $\mathcal{A}_{\text{retain}}$ and removed set $\mathcal{A}_{\text{remove}}$.
Compute clip-wise mean scores
for $i \leftarrow 1$ **to** M **do**
  $\bar{s}_i \leftarrow \frac{1}{N} \sum_{j=1}^{N} s_{ij}$
Compute annotator MSE and Spearman correlation
for $j \leftarrow 1$ **to** N **do**
  $\text{MSE}_j \leftarrow \frac{1}{M} \sum_{i=1}^{M} (s_{ij} - \bar{s}_i)^2$
  $\rho_j \leftarrow \text{Spearman}(\{s_{1j}, \ldots, s_{Mj}\}, \{\bar{s}_1, \ldots, \bar{s}_M\})$
Compute filtering thresholds
$\tau_{\text{MSE}} \leftarrow \text{quantile}_{0.75}(\{\text{MSE}_j\}_{j=1}^{N})$
$\tau_{\rho} \leftarrow \text{quantile}_{0.25}(\{\rho_j\}_{j=1}^{N})$
Risk scoring and filtering
for $j \leftarrow 1$ **to** N **do**
  $R_j \leftarrow \mathbb{1}(\text{MSE}_j \geq \tau_{\text{MSE}}) + \mathbb{1}(\rho_j \leq \tau_{\rho})$
  if $R_j = 2$ **then**
    mark annotator j as **High-risk** (Remove)
  else if $R_j = 1$ **then**
    mark annotator j as **Medium-risk** (Retain, flagged)
  else
    mark annotator j as **Low-risk** (Retain)
return $\mathcal{A}_{\text{retain}} = \{j : R_j \leq 1\}$, $\mathcal{A}_{\text{remove}} = \{j : R_j = 2\}$

A.5. Data Analysis on Simulated Environment

We further analyzed the subset of simulated humanoid motions, which consists of 243 samples spanning seven distinct action categories. Tab. S3 reports the average human-likeness scores for each category, comparing human subjects and simulated humanoid motions. Overall, the simulated motions achieve higher scores than real humanoid robots, indicating that simulation allows for more precise or idealized motion execution. However, they still fall short of actual human performance in every category, highlighting the inherent gap between synthetic and natural human motion.

The analysis reveals that the discrepancy between simulated humanoids and real humans varies across actions. For example, the largest gap is observed in the *jump* category (2.46 points), whereas *waving* shows the smallest difference (0.35 points). This suggests that while simulations can effectively capture certain types of motions, they are less capable of replicating more complex or expressive movements that humans perform naturally. These findings emphasize that, although simulated data can serve as a useful proxy for evaluating motion algorithms, they cannot fully replace real human motion in benchmarks aiming to assess human-likeness.

B. More Details on Motion Turing Test Task

B.1. PTR-Net Model Details

The Pose-Temporal Regression Network (PTR-Net) is designed as a lightweight yet effective architecture for the Motion Turing Test task, mapping a 3D pose sequence to a continuous human-likeness score. PTR-Net consists of three main components: a temporal encoder, a spatial–temporal graph convolution module, and an attention-based regression head. The complete pipeline is illustrated in Fig. 7 of

Table S3. **Human vs. Humanoid (simulated) motion quality comparison.** Simulated humanoids consistently score lower than humans, with the largest gaps appearing in high-dynamic actions such as *jump* and *boxing*.

Category	Human	Humanoid (sim)	Score Difference
jump	4.43	1.97	2.46
boxing	3.76	2.41	1.35
run	3.73	2.45	1.28
kicking ball	3.93	2.70	1.23
kungfu	3.60	2.47	1.13
dance	3.47	2.50	0.97
waving	3.70	3.35	0.35

the main paper.

Temporal Encoder. Given an input pose sequence $X = \{x_t\}_{t=1}^{T}$, a two-layer bidirectional LSTM is employed to capture long-range temporal dependencies across frames. The encoder outputs temporally enriched features $H_t \in \mathbb{R}^{2h}$ that retain both forward and backward contextual information, which is essential for modeling motion continuity and temporal smoothness.

Spatial–Temporal Graph Convolution (ST-GCN). The encoded sequence is reshaped into a graph representation where each node corresponds to a body joint, and edges represent human skeletal connections. A stack of ST-GCN blocks alternates between spatial graph convolutions and temporal convolutions to extract joint coordination patterns. Unlike conventional skeleton-based GCNs, PTR-Net adopts a parameter-free adjacency matrix, avoiding the need to learn fixed graph structures and enabling more flexible feature aggregation across joints and frames.

Attention Pooling and Regression Head. To emphasize motion segments that are most informative for judging human-likeness, a temporal attention mechanism is applied over the ST-GCN outputs. The weighted features are then passed into a lightweight MLP regression head that outputs a scalar human-likeness score. Formally, PTR-Net learns a mapping

$$s = f_\theta(X), \tag{1}$$

where f_θ denotes the network parameters and $s \in [0, 5]$ is the predicted score.

Training Objective. PTR-Net is trained with an L2 regression loss combined with a temporal smoothness regularization term:

$$\mathcal{L} = \|\hat{s} - s^*\|_2^2 + \lambda \mathcal{L}_{\text{reg}}, \tag{2}$$

where s^* is the ground-truth human-likeness score and $\hat{s}$ is the prediction. The regularization term $\mathcal{L}_{\text{reg}}$ penalizes excessive fluctuations across consecutive predictions, encouraging stability and consistency in the output scores.

Additional implementation details, hyperparameters, and network configurations are provided in the following sections of the supplementary material.

B.2. Evaluation Metrics Definitions

To systematically quantify the alignment between model predictions and human evaluations, we adopt three complementary evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Spearman's Rank Correlation (ρ). Their formal definitions and usage descriptions are provided below.

Mean Absolute Error (MAE). MAE measures the absolute difference between the predicted scores and the human preference annotations, defined as:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{s}_i - s_i^*|, \tag{3}$$

where $\hat{s}_i$ and s_i^* denote the predicted and human scores for the i-th motion sequence, respectively. A lower MAE indicates closer numerical alignment to human judgments.

Root Mean Squared Error (RMSE). RMSE provides a more penalized metric for large deviations, given by:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{s}_i - s_i^*)^2}. \tag{4}$$

Compared with MAE, RMSE emphasizes stability and robustness by amplifying large prediction errors.

Spearman's Rank Correlation (ρ). Spearman's ρ measures the monotonic consistency between model-predicted rankings and human-provided rankings. We also compute Spearman's rank correlation to evaluate monotonic consistency between model predictions and human annotations:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{N(N^2 - 1)}, \tag{5}$$

where d_i is the rank difference between predicted and annotated scores. A higher ρ indicates that the model better preserves human-like relative ordering across motion samples.

B.3. Training Details

For the Motion Turing Test task, we train our proposed PTR-Net baseline exclusively on the HHMotion dataset. All the SMPL-X sequences are zero-padded to a maximum fixed length and processed into joint coordinate representations. A corresponding binary mask is generated for each sequence to ensure that the network ignores the padded regions during computation. The human-likeness scores (ranging from 0 to 5) are normalized $[0, 1]$ for training stability. We partition the dataset into training and testing sets

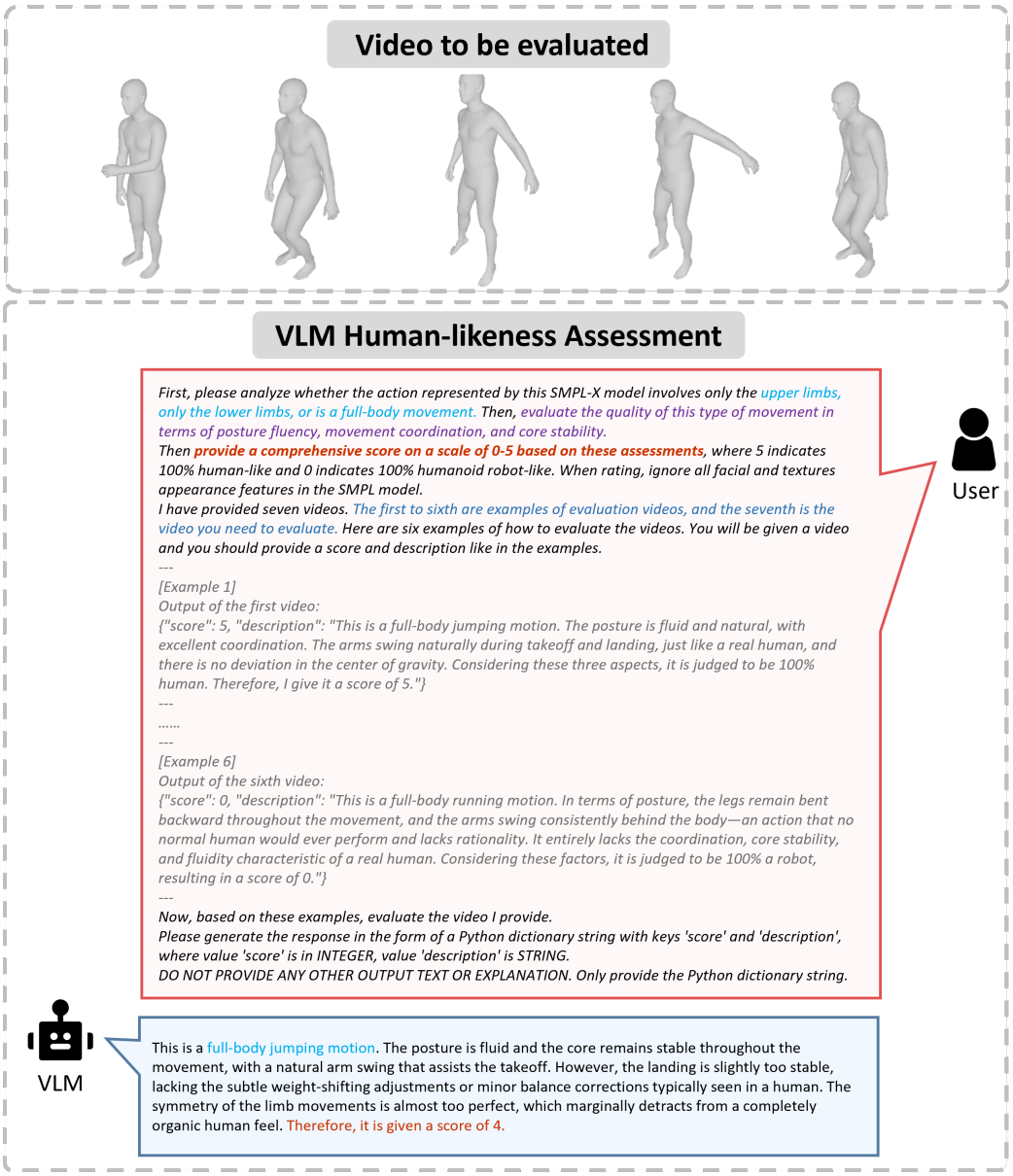

Figure S6. **Prompt used for VLM-based human-likeness evaluation.**

with a ratio of 80% and 20%, respectively. The model is trained for 500 epochs using the AdamW optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. To make the model more robust to diversity in motion, data augmentation techniques like temporal jittering and joint-level Gaussian noise are used.

For comparison, we additionally evaluate two representative LLMs, Gemini-2.5 Pro [5] and Qwen3-VL-Plus [4]. Both models process each SMPL-X motion clip as a video sequence input and output a human-likeness score following the same evaluation protocol as our baseline.

C. Prompts for VLM Evaluation

C.1. Evaluation Prompts

To ensure a fair and reproducible evaluation of motion human-likeness, we design a structured prompt for the VLM. The prompt explicitly instructs the model to assess the human-likeness of a motion sequence represented by an SMPL-X model, based solely on kinematic information without any texture, facial details, or appearance cues.

Specifically, the VLM is first required to determine whether the motion belongs to upper-lower limb movement or full-body movement. Then instructed to evaluate the

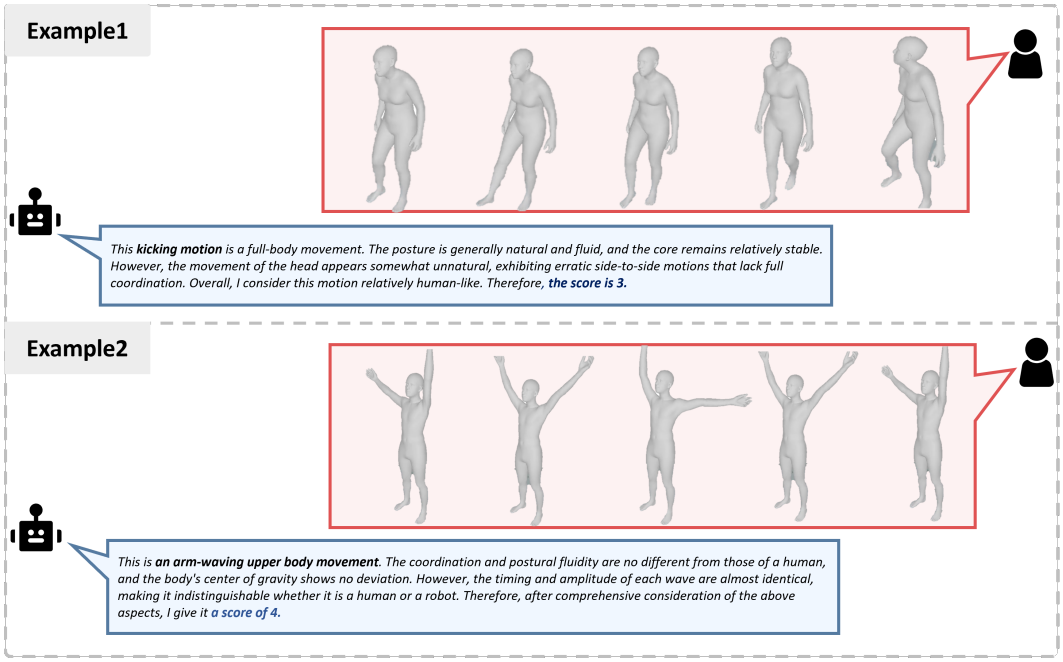

Figure S7. **VLM evaluation results on two unseen test videos.**

motion quality from three aspects: (1) posture fluency, (2) movement coordination, and (3) core stability. Based on these criteria, the model provides a comprehensive human-likeness score on a 0-5 scale, where 5 indicates indistinguishable from real human motion and 0 indicates completely robotic motion.

To reduce ambiguity and improve consistency, we provide six example videos in the prompt, each paired with a reference output including both a score and a detailed explanation. The VLM is instructed to follow the same format when evaluating a new input video. Fig. S6 shows the complete prompt used in our experiments.

C.2. Cases of VLM Evaluation

We present two representative test cases evaluated by the VLM PA-CoT instructions in Fig. S7. In the first case, although the overall posture is natural and the core remains stable, slightly rigid head movements reduce coordination, leading to a score of 3. In the second case, the posture and balance appear highly human-like, but overly uniform timing and amplitude introduce subtle robotic characteristics, resulting in a score of 4.

D. Limitations and Future Work

While the proposed HHMotion dataset and the Motion Turing Test benchmark provide a rigorous foundation for evaluating the human-likeness of humanoid robots, we acknowledge several limitations in our current work that point towards promising directions for future research.

Demographic Diversity of Annotators. To ensure annotation consistency and quality control, our evaluation was primarily conducted in a laboratory setting with 30 university students and researchers (aged 20–30). While this provides a stable baseline for the Motion Turing Test, we acknowledge that this demographic may not fully represent the perceptual thresholds of broader populations, such as children or the elderly. Future iterations could employ crowdsourcing to verify the generalization of our human-likeness scores across diverse age groups and backgrounds.

Kinematic Structure Mismatch. Although we employed a rigorous manual cross-validation pipeline to filter reconstruction errors, mapping diverse humanoid robots to the human SMPL-X topology inevitably introduces minor failures that influence the evaluation. Future work could incorporate multi-view reconstruction and IMU-based validation to further strengthen reliability.

E. Data Privacy and IRB

This study, which involves the collection and processing of human motion data, was reviewed and approved by the Institutional Review Board (IRB). The online video data utilized in the HHMotion dataset were sourced from publicly accessible online platforms (e.g., YouTube) and are used strictly for non-commercial research purposes. We do not claim any copyright over the original video materials; all rights are retained by their respective creators.

To rigorously protect privacy, all individuals in the videos are anonymized. This process involves the full

masking of faces and any other personally identifiable information (PII). The dataset to be publicly released consists solely of these anonymized video clips and their corresponding SMPL-X representations, ensuring that individual identities are safeguarded.

For videos involving volunteers, informed consent was obtained, authorizing the use of their data for research purposes. All annotations were conducted exclusively for academic research, in compliance with prevailing ethical standards in the field of human motion analysis.

References

[1] The 2025 World AI Conference. https://www.worldaic.com.cn/, 2025. 1

[2] The 2025 World Humanoid Robot Games. https://www.whrgoc.com/, 2025. 1

[3] The 2025 World Robot Conference. https://www.worldrobotconference.com/, 2025. 1

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 6

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6

[6] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010. 3

[7] B Series. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT*, 500(13), 2012. 3

[8] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing*, 19(6):1427–1441, 2010. 3, 4

[9] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[10] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 3

[11] Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. 3

[12] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263, 2008. 3, 4

[13] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8866, 2023. 3

[14] Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017. 3, 4

[15] Computer Vision Annotation Tool. Cvat. https://www.cvat.ai/, 2025. 2, 3

[16] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1148–1159, 2025. 3

[17] Kanglei Zhou, Ruizhi Cai, Liyuan Wang, Hubert PH Shum, and Xiaohui Liang. A comprehensive survey of action quality assessment: Method and benchmark. *arXiv preprint arXiv:2412.11149*, 2024. 4

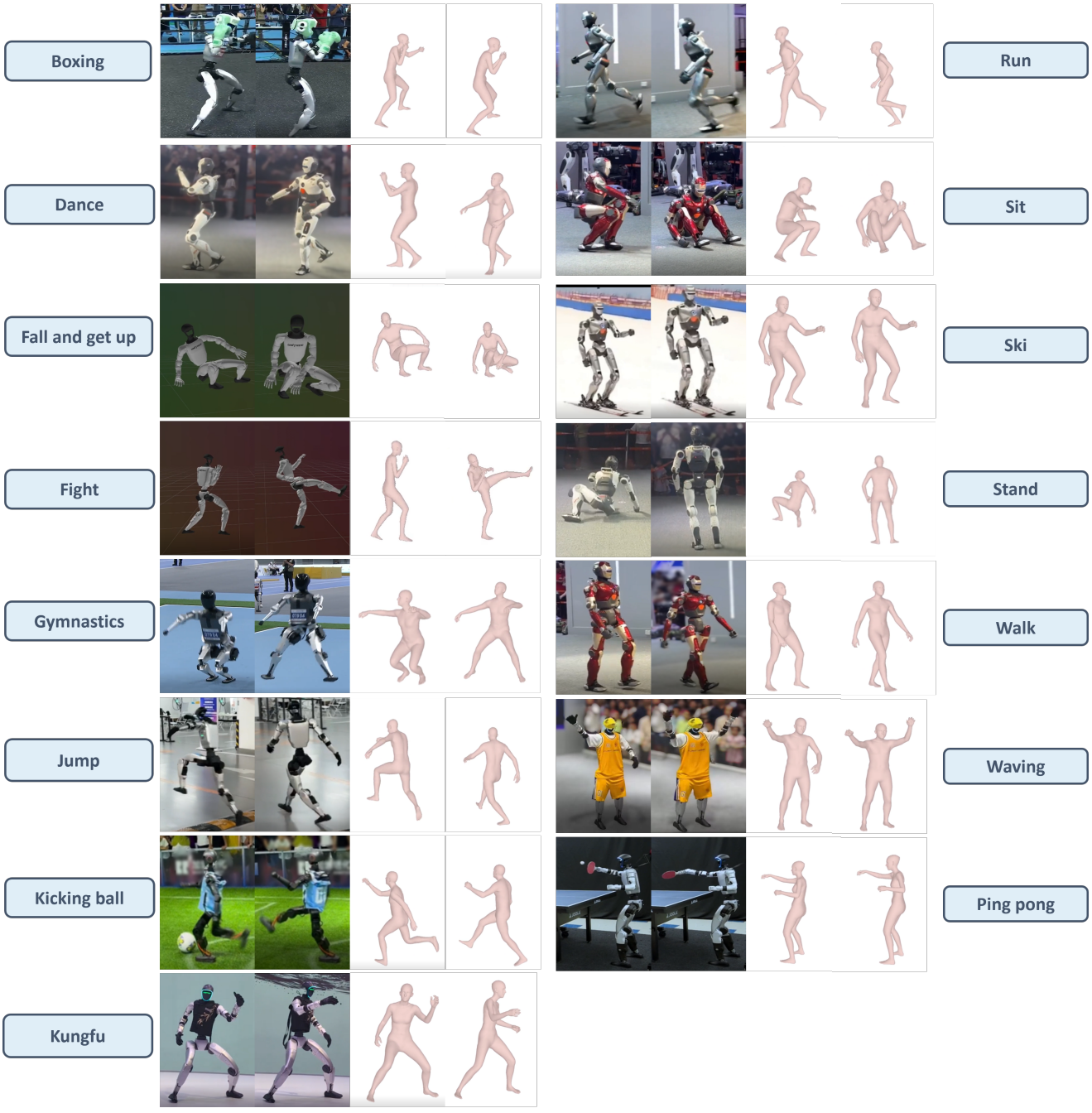

Figure S8. **Fifteen humanoid motion categories.** This figure illustrates the full set of 15 action categories captured from real-world and simulated humanoid robots in HHMotion, covering both sport and daily motions.

Figure S9. **Fifteen human motion categories.** The corresponding 15 action categories were collected from human subjects, providing a comprehensive reference of natural human motion patterns for evaluating human-likeness.