I-Robot: Identifying Robotic and Human Motion in Humanoids

Taehoon Kim [1] **Jongwook Choi** [1] **Haeun Noh** [1] **Junyeup Hwang** [2] **Jongwon Choi** [1 2 3]

Abstract

Commercial humanoid robots now closely match human body proportions and movement, and ordinary clothing is often enough to make them visually indistinguishable from a person. This visual similarity introduces new risks: a humanoid can be mistaken for a human in surveillance footage, enabling the fabrication of false alibis. To our knowledge, no existing benchmark directly studies human-versus-humanoid identification under an appearance-free motion-only setting. We address this problem by determining whether a motion sequence corresponds to a human or a humanoid, using only temporal pose information without relying on visual appearance cues. We construct HvH (HumanVsHumanoid), a dataset spanning 38 humanoid platforms and 11 shared action classes, and propose I-Robot, a dual-branch model that processes raw pose sequences and their temporal differences in parallel and fuses them via learned per-channel attention. I-Robot consistently outperforms standard sequence and skeleton classifiers on HvH.

1. Introduction

Distinguishing a humanoid robot from a human is becoming visually hard. In just a few years, commercial humanoid platforms have reached human-scale body proportions, human-like walking, and a joint range close to that of a person (Peng et al., 2018; Rudin et al., 2022). Putting ordinary clothes on such a robot is now often enough to make it look like a human, and recent public demos show that actions like walking, turning, and picking up objects already sit within the natural range of human movement.

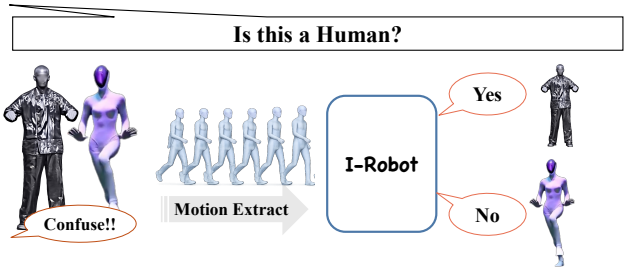

Figure 1. Human-versus-humanoid identification.

This visual similarity opens a new space for misuse, while identifying humanoids (Abdulatif et al., 2018) has received little attention. A humanoid walking past a surveillance camera could be logged as a person, fabricating an alibi at a location the real person never visited. The same pattern has played out twice before with face spoofing and deepfakes (Van Den Oord et al., 2016; Liu et al., 2018; Karras et al., 2019; Kim et al., 2025): generation and attack tools reached practical quality before reliable detectors existed (Rossler et al., 2019; Jeong et al., 2022; Shiohara & Yamasaki, 2022; Jung et al., 2025; Chen et al., 2025), and detection matured only in response to documented cases of misuse. Addressing humanoid identification before a comparable trigger occurs motivates this work.

We therefore address human-versus-humanoid identification from motion. To support this task, we construct the HvH (HumanVsHumanoid) dataset, which consists of 574 videos collected from public YouTube footage, covering a broad range of real-world humanoid motions across both recent and earlier platforms. HvH covers 38 humanoid platforms and 11 human action classes, with both sides sharing the same action taxonomy to reduce context bias. Each clip is processed by off-the-shelf MHR (Ferguson et al., 2025) and SMPL (Loper et al., 2023) estimators, such as SAM3D (Yang et al., 2026) and ScoreHMR (Stathopoulos et al., 2024), to produce pose sequences, and the dataset provides evaluation protocols that hold out unseen platforms and actions to measure generalization beyond the training distribution.

As a simple but effective baseline, we propose I-Robot, a dual-branch model that processes raw pose sequences and temporal difference sequences in parallel to distinguish

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul, South Korea [2]Department of Metaverse Convergence, Chung-Ang University, Seoul, South Korea [3]Department of Advanced Imaging, GSAIM, Chung-Ang University, Seoul, South Korea. Correspondence to: Jongwon Choi <choijw@cau.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

humans from humanoids, visualized in Figure 1. I-Robot consistently outperforms standard sequence classification baselines, showing that temporal pose differences provide a highly discriminative cue for separating humans from humanoids. Beyond detection, I-Robot can serve as a general-purpose probe that quantifies how distinguishable a platform's motion is from natural human movement, providing a useful signal for human-robot interaction and motion generation as humanoid motion continues to close the gap with human movement.

Our contributions are as follows:

- We address human-vs-humanoid identification from motion, motivated by the misuse risk of humanoids whose visual appearance already matches humans.
- We construct the HvH dataset, covering 38 humanoid platforms and 11 action classes collected from public YouTube footage, with a held-out split of advanced platforms for measuring generalization to unseen systems.
- We propose I-Robot, which processes raw pose and temporal difference sequences in parallel, and show that it consistently outperforms standard sequence classifiers, identifying frame-to-frame pose differentials as the most discriminative cue.

2. Related Work

2.1. Human Motion Imitation for Robots

Humanoid robotics has aimed to create artificial agents that resemble humans in both form and behavior, motivating extensive research on human motion and behavior imitation. Early motion imitation studies formulated human-to-robot transfer as constrained optimization or inverse-kinematics problems, allowing captured human motions to be reproduced under robot-specific physical limitations (Nakazawa et al., 2002; Safonova et al., 2003; Dariush et al., 2009). Physics-based reinforcement learning later extended this direction by training simulated humanoid agents to imitate diverse reference motions with physically plausible control (Heess et al., 2017; Peng et al., 2017; 2018; 2021; 2022). More recent motion retargeting and whole-body teleoperation systems have further bridged human–robot embodiment gaps, allowing real humanoid robots to reproduce increasingly natural human-like behaviors (He et al., 2024; Ze et al., 2025; Li et al., 2025). Motivated by these advances, we study a new question raised by them: whether humanoid motion, as it grows ever closer to human motion, can still be reliably distinguished.

2.2. Pose Estimation

We use pose estimation methods to obtain meaningful body representations from visual observations. 2D pose estimation is limited in capturing spatially consistent body configurations because they encode body joints only in the image plane, making it sensitive to viewpoint changes, depth ambiguity, and occlusion (Toshev & Szegedy, 2014; Newell et al., 2016; Cao et al., 2017; Sun et al., 2019). In contrast, 3D pose estimation provides explicit structural and kinematic information, such as joint configuration, body orientation, and pose geometry, which is more suitable for comparing human and humanoid behaviors beyond appearance-level cues (Kanazawa et al., 2018; Kolotouros et al., 2019; Kocabas et al., 2020; 2021; Zhu et al., 2023). Recent 3D human recovery methods, such as Score-Guided Human Mesh Recovery (Stathopoulos et al., 2024) and SAM 3D Body (Yang et al., 2026), further improve pose representation by estimating parametric body models from images. In this work, we leverage these advances to extract MHR (Ferguson et al., 2025) and SMPL (Loper et al., 2023) representations as effective 3D body-level features.

2.3. Human Identification

A related line of research recognizes humans while reducing reliance on appearance cues such as clothing and outfit. Cloth-changing person re-identification exploits clothing-irrelevant cues such as body shape, contour, and motion (Qian et al., 2020; Gu et al., 2022; Jin et al., 2022), while gait recognition models body dynamics from RGB, skeleton, or pose sequences under variations in viewpoint, clothing, and carrying condition (Chao et al., 2019; Fan et al., 2023; Teepe et al., 2021; Liao et al., 2020). These studies are relevant because humanoids may be clothed or covered with human-like shells, making appearance unreliable. However, they mainly assume human subjects and focus on identity recognition, whereas our goal is category-level human-versus-humanoid identification across diverse actions and platforms.

2.4. Human–Robot Comparison

Prior work (Abdulatif et al., 2018) classifies humans and robots using radar micro-Doppler signatures, but requires specialized radar hardware confined to constrained industrial settings. Concurrent with our work, (Feng et al., 2026) quantifies biomechanical divergence between human and humanoid gaits from joint-level measurements to provide benchmarks for humanoid robot dynamics, yet it is limited to walking and demands direct access to the robot's internal state. (Li et al., 2026) predicts scalar human-likeness scores on pose sequences via the Motion Turing Test, but its formulation is inherently a regression over subjective human ratings and does not yield a detection signal for distinguish-

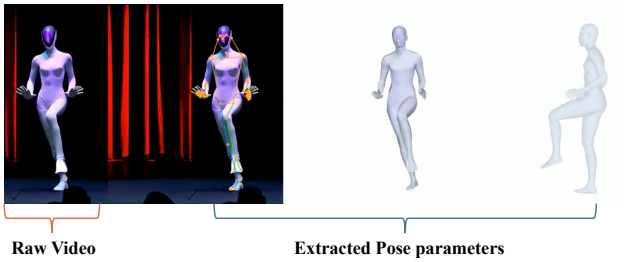

Figure 2. Raw humanoid video and the corresponding extracted pose parameter.

Table 1. HvH dataset overview.

Metric	Value
Total clips (scenes)	574
Humanoid clips	399
Human clips	175
Humanoid platforms	38
Action classes	11
Total duration	2.8 h
Avg clip length – Humanoid	20.7 s
Avg clip length – Human	10.3 s
Median clip length – Humanoid	9.9 s
Median clip length – Human	6.7 s

ing human from humanoid motion. Our work also operates on pose motion but reframes the problem as detection across diverse humanoid motion types, overcoming the sensing, scope, and task-formulation limitations of prior efforts.

3. HumanVsHumanoid Dataset (HvH Dataset)

HvH is a new dataset for human and humanoid identification from pose motion, collected entirely from publicly available YouTube videos. Table 1 summarizes the dataset. It contains 574 clips in total, split into 399 humanoid clips across 38 commercial and research platforms and 175 human clips, covering 11 shared action classes with a total duration of approximately 2.8 hours.

3.1. Data Collection and Processing

We collected videos from YouTube and processed each clip into MHR (Ferguson et al., 2025) and SMPL (Loper et al., 2023) parameter sequences to extract motion. We visualize in Figure 2 the raw video alongside the corresponding extracted MHR parameter.

Videos were searched and downloaded from YouTube using platform names and action keywords as queries. For humanoid clips, queries were formed by combining each of the 38 platform names with action terms such as walking, picking, and dancing. For human clips, the same action terms were used to collect clips of people performing the same movement categories.

Every clip was manually reviewed before pose estimation was applied. Reviewers checked that the subject remained visible without heavy occlusion throughout the clip, and that the bounding box of the robot was clearly defined for humanoid clips. Clips that failed either check were discarded, and only the 574 clips that passed both conditions formed the final dataset.

Pose estimation was applied to each passing clip using two independent pipelines, one per representation. MHR parameters were extracted directly using SAM 3D Body (Yang et al., 2026), which recovers full-body mesh parameters from monocular video without a separate detection stage. In our experiments, we use the SMPL parameters that were extracted using a two-stage pipeline: D-FINE (Peng et al., 2024) first detected the subject bounding box at every frame, and the ScoreHMR (Stathopoulos et al., 2024) was then applied inside the detected region to extract SMPL parameters. Both representations are stored at a fixed frame rate and used as input to all experiments, allowing us to compare their effectiveness as motion descriptors for the identification task.

3.2. Dataset Statistics

Figure 3 shows the distribution of clips across action classes and humanoid platforms. Walking, exercise, and picking are the three largest action classes, while stairs and jumping are the smallest. Humanoid clips are on average longer than human clips, with a mean of 20.7 s versus 10.3 s, because some humanoid demonstration videos are unusually long. The 37 clips labeled as unknown are excluded from action-level evaluation. The dataset spans 38 humanoid platforms and follows a long-tail distribution. A few platforms, such as *Boost Robotics T1*, *Phoenix*, and *Figure 02*, account for a large share of clips, while most platforms have fewer than ten samples. This reflects the uneven availability of humanoid footage in the wild.

3.3. Evaluation Protocols

The validation split is designed to test generalization across both action categories and humanoid platforms. For human clips, an 8:2 split is applied per action class, resulting in 141 training clips and 34 validation clips across 10 action categories. For humanoid clips, an 8:2 split is applied per platform at the clip level, resulting in 314 training clips and 108 validation clips across 38 platforms. Three platforms are assigned entirely to the validation set and never appear in training, including Unitree H1, Unitree H2, and Xpeng IRON. These platforms were not seen during training, providing a stricter evaluation of generalization to unseen platforms. Among them, Unitree H1, Unitree H2, and IRON

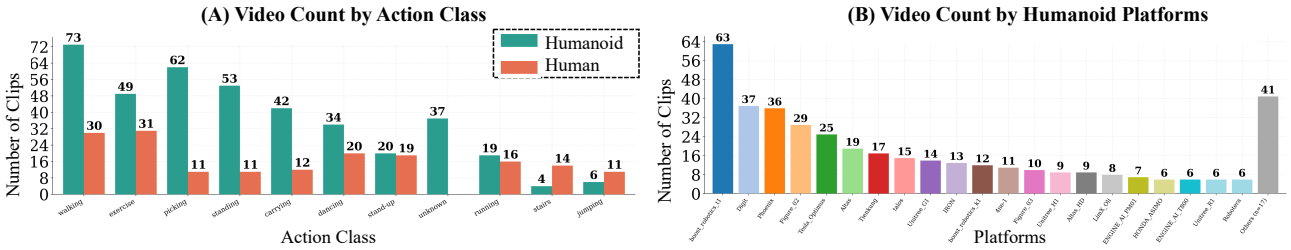

Figure 3. Statistics of the HvH dataset. (A) Distribution of videos across action classes for both humanoid robots and humans. (B) Distribution of videos across different humanoid platforms.

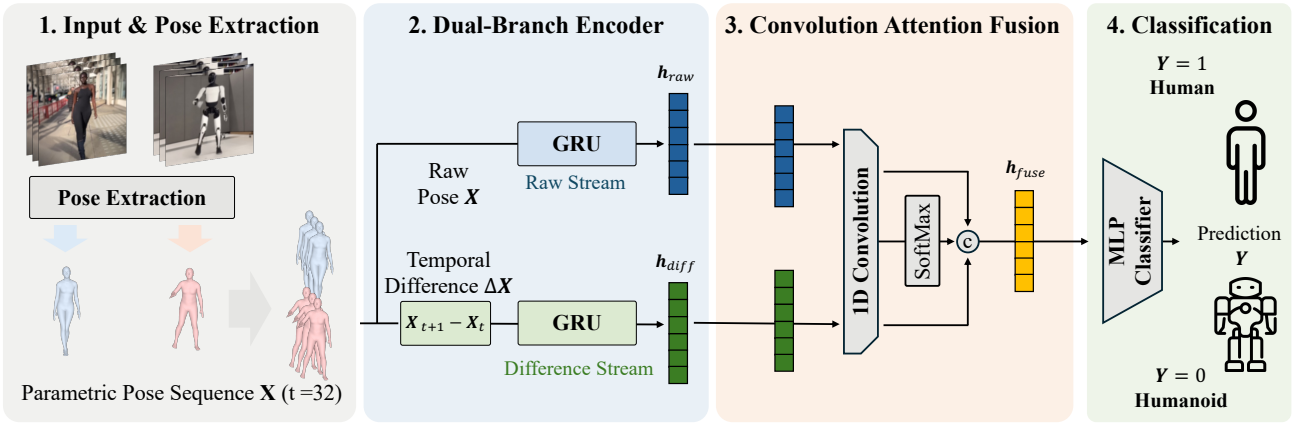

Figure 4. Visualization of our proposed framework, I-Robot.

represent particularly advanced humanoid systems whose motion characteristics are closest to natural human movement, making them the hardest cases in the validation set.

4. Proposed Method

We propose I-Robot, a pose motion-based framework for distinguishing humans from humanoid robots. I-Robot operates on parametric pose sequences alone, processes the raw pose trajectory and its temporal difference in two parallel branches, and fuses them through a learned convolutional attention so that the contribution of postural and velocity-level cues is reweighted on a per-sample basis. We visualized the proposed framework in Figure 4.

4.1. Problem Formulation

We focus on a setting where a humanoid robot already looks like a human, perhaps wearing clothes, a mask, or a realistic outer shell, so visual cues alone can no longer tell the two apart. Appearance features such as texture, color, and body shape can be copied from a human on demand, but motion cannot: it is produced by the robot's own physics and controllers, and is hard to fake from one frame to the next. We therefore identify humans and humanoids from how they move, using pose parameter sequences extracted from monocular video instead of raw image frames.

Formally, given a sequence of pose parameters $\mathbf{X} = \{x_t\}_{t=1}^{T}$ extracted from a monocular video clip, the goal is to predict a binary label $y \in \{0, 1\}$ indicating whether the subject is a human or a humanoid robot. Each frame x_t is a flattened pose vector, and the sequence is sampled to a fixed window of $T = 32$ frames. The input dimension depends on the pose representation: $x_t \in \mathbb{R}^{207}$ for SMPL parameters and $x_t \in \mathbb{R}^{381}$ for MHR joint coordinates. In both cases, the model receives no image or appearance information, only the motion trajectory encoded in the pose sequence.

4.2. I-Robot

We propose I-Robot, a dual-branch model for human vs. humanoid identification from pose sequences. The model encodes the raw pose sequence and its frame-wise difference in parallel and fuses the two branches via a learned per-channel attention.

Difference stream. Given $\mathbf{X} \in \mathbb{R}^{T \times D}$, we define the frame-wise difference as

$$\Delta x_t = x_t - x_{t-1}, \quad t = 2, \ldots, T, \tag{1}$$

with $\Delta x_1 = \mathbf{0}$. The difference signal exposes subtle frame-to-frame irregularities that are largely flattened in the raw

Table 2. Comparison results on validation set on MHR parameters.

Model	F1	AUC	Accuracy	Precision	Recall
1D-CNN	0.7827	0.9181	0.8863	0.7544	0.8264
GRU	0.7896	0.9145	0.8896	0.7602	<u>0.8354</u>
LSTM	0.7740	0.8814	0.8885	0.7562	0.7967
TCN	<u>0.8197</u>	**0.9368**	<u>0.9224</u>	**0.8410**	0.8020
CTR-GCN	0.5877	0.8996	0.8571	0.4606	0.8117
ST-GCN	0.5547	0.8753	0.8645	0.4719	0.6728
I-Robot (Ours)	**0.8479**	<u>0.9304</u>	**0.9286**	<u>0.8388</u>	**0.8579**

pose, providing a complementary cue for detection.

Dual-branch encoder. The raw sequence $\mathbf{X}$ and the difference sequence $\Delta\mathbf{X}$ are encoded by two independent two-layer bidirectional GRUs with layer normalization, then mean-pooled along time to produce summary vectors $\mathbf{h}_{\text{raw}}, \mathbf{h}_{\text{diff}} \in \mathbb{R}^d$. Independent encoders prevent the two streams from collapsing into a shared representation, so each branch can specialize in its own signal.

Convolutional attention fusion. A 1D convolution over the two summaries produces per-channel attention weights $(\alpha_{\text{raw}}, \alpha_{\text{diff}}) \in \mathbb{R}^d \times \mathbb{R}^d$, normalized across branches by softmax. The fused representation is

$$\mathbf{h}_{\text{fuse}} = \alpha_{\text{raw}} \odot \mathbf{h}_{\text{raw}} + \alpha_{\text{diff}} \odot \mathbf{h}_{\text{diff}} + \tfrac{1}{2}(\mathbf{h}_{\text{raw}} + \mathbf{h}_{\text{diff}}), \quad (2)$$

where $\odot$ denotes element-wise multiplication. Replacing the standard concatenation with a 1D-convolutional gate lets the relative weight of the two streams adapt per sample and per channel, rather than being fixed by the linear projection that follows a concat. For the final decision, $\mathbf{h}_{\text{fuse}}$ is passed to a two-layer MLP to predict $\hat{y} \in \{0, 1\}$.

5. Experiments

5.1. Implementation Details

All models are trained on pose sequences of length $T = 32$ with a stride of 32 with non-overlapping chunks. All models use a hidden dimension $d = 256$, a two-layer classification head with GELU (Hendrycks & Gimpel, 2016) and dropout 0.2, and are trained from scratch with AdamW (Loshchilov & Hutter, 2019) (learning rate 10^{-3}, weight decay 10^{-4}) for 100 epochs with a cosine annealing schedule (Loshchilov & Hutter, 2017) and gradient clipping at norm 1.0. Batch size is 32, and we use binary cross-entropy loss with a positive-class weight computed as $N_{\text{neg}}/N_{\text{pos}}$ on the training split to balance the human-versus-humanoid ratio. All experiments run on a single NVIDIA RTX A6000 GPU. We report five metrics on the held-out split: Macro F1, AUC, Accuracy, Precision, and Recall.

5.2. Baselines

We compare I-Robot against six baselines spanning families of sequence and skeleton classifiers, all operating on the same fixed-length pose sequence of $T = 32$ frames. Each model ends with a two-layer MLP head to produce the final binary prediction. Performance is reported with the best result highlighted in **bold** and the second-best result <u>underlined</u>.

1D-CNN (Tang et al., 2020). A three-layer 1D convolutional network processes the pose sequence along the temporal axis with kernel sizes of 5, 3, and 3. Each layer is followed by batch normalization and a GELU activation, and the output is summarized by global average pooling over time.

LSTM and GRU (Hochreiter & Schmidhuber, 1997; Cho et al., 2014). Two recurrent baselines process the input with a two-layer bidirectional LSTM or GRU, capturing temporal dependencies in both directions. The hidden states across all time steps are aggregated by mean pooling along the temporal axis before the classifier head.

TCN (Lea et al., 2017). A temporal convolutional network with four dilated residual blocks of kernel size 3 and exponentially increasing dilation $\{1, 2, 4, 8\}$, yielding a receptive field that covers the full $T = 32$ frame window. Each block uses two dilated convolutions with batch normalization, GELU, and a residual connection, and the sequence is mean-pooled before classification.

ST-GCN (Yan et al., 2018). A spatio-temporal graph convolutional network that reshapes the flat pose vector into a joint graph and alternates spatial graph convolutions with temporal convolutions.

CTR-GCN(Chen et al., 2021). An extension of ST-GCN in which each block refines a shared static adjacency with a channel-wise dynamic adjacency learned per sample.

5.3. Main Results

Table 2 reports results using MHR joint coordinates as the input representation. I-Robot achieves the highest F1 and Accuracy, and is second-best on the remaining three metrics, indicating balanced performance across the full precision-recall trade-off. Among the baselines, TCN is the strongest by F1 and AUC, while graph-based models (ST-GCN, CTR-GCN) fall behind despite their explicit joint-level inductive bias, suggesting that for the human-vs-humanoid decision, the dominant signal lies in the temporal structure of motion rather than in static inter-joint relations. CTR-GCN achieves the highest Recall but with markedly lower Precision, a behavior consistent with a model that is biased

Table 3. Per-Action F1 performance on MHR parameters

Action	I-Robot	GRU	1D-CNN	TCN	ST-GCN	CTR-GCN	LSTM
stairs	**1.000**	0.897	0.966	0.966	0.933	0.966	0.966
standing	**0.966**	0.810	0.914	0.946	0.881	0.821	0.766
stand up	0.856	0.851	0.893	**0.920**	0.770	0.889	0.889
exercise	0.861	**0.881**	0.874	0.874	0.757	0.802	0.826
running	**0.985**	0.943	0.675	0.653	0.957	0.664	0.749
picking	**0.879**	0.767	0.769	0.766	0.635	0.793	0.771
walking	0.767	0.714	0.665	0.766	**0.769**	0.740	0.719
carrying	0.493	0.614	0.731	**0.782**	0.530	0.601	0.608
dancing	0.653	0.575	**0.687**	0.629	0.548	0.527	0.601
jumping	0.486	**0.500**	0.486	0.438	0.486	0.471	0.486
unknown	0.484	0.485	0.440	**0.497**	0.447	0.457	0.484

toward predicting the humanoid class.

5.4. Analysis

Table 3 reveals several results that differ from our initial expectations. Actions with dynamic motion, such as jumping and dancing, were expected to be easy to classify since humanoid controllers are known to struggle with rapid and coordinated movements. However, all models perform relatively lower on these actions compared to other categories, suggesting that even imperfect humanoid motion is hard to distinguish from human movement once projected onto parametric pose parameters. One possible explanation is that the pose estimation process itself introduces noise for fast and irregular motions, which may obscure the discriminative signal that would otherwise separate human and humanoid movement in these action categories.

Figure 5 shows the accuracy of I-Robot by platform, separating humans, seen humanoids, and unseen humanoids. The single orange bar on the left reports accuracy on human clips. The green bars report accuracy on humanoid platforms seen during training, and the blue bars on the right report accuracy on three platforms held out from training (IRON, Unitree H1, Unitree H2). Numbers above each bar are per-platform accuracy (%), and the N below each bar is the number of clips (non-overlap) contributed by that platform.

Accuracy drops on more recent and more human-like platforms. Unitree H1 is classified almost perfectly, but its successor H2 is much harder despite coming from the same product family. IRON, one of the newest humanoids in the dataset, is also harder to identify. All three platforms are unseen during training, but the gap between H1 and H2 inside the same family shows that the drop is not just about generalization; it reflects the hardware and control improvements between generations. The same trend shows up among seen platforms: Unitree R1 has one of the lowest accuracies, even though it is included in training, which means that more advanced humanoids remain hard to detect even after the model has seen them. Together, these observations support our broader claim that the detection task becomes strictly harder as humanoid motion approaches human motion.

Table 4. Ablation study on I-Robot components in MHR parameters.

Variant	Raw	Diff	Conv. Fusion	F1	AUC	Accuracy	Precision	Recall
Raw only	✓	×	×	0.7959	0.9189	0.9014	0.7811	0.8137
Diff. only	×	✓	×	0.7906	0.8977	0.9025	0.7850	0.7965
Concat.	✓	✓	×	0.8396	**0.9483**	0.9209	0.8174	**0.8677**
I-Robot (Ours)	✓	✓	✓	**0.8479**	0.9304	**0.9286**	**0.8388**	0.8579

Table 5. I-Robot performance comparison between MHR and SMPL representations.

Representation	F1	AUC	Accuracy	Precision	Recall
SMPL ($D = 207$)	0.7317	0.8870	0.8519	0.6991	0.8084
MHR ($D = 381$)	**0.8479**	**0.9304**	**0.9286**	**0.8388**	**0.8579**

Figure 6 illustrates both successful and failure cases for *walking* and *dancing* action classes. Notably, in the fail cases, even human evaluators find it difficult to distinguish the render results. The generated motions exhibit smooth and highly human-like dynamics, which become more evident when observing temporally connected frames.

5.5. Ablation Study

Table 4 validates the two core design choices in I-Robot: the difference branch and the convolutional attention fusion. Adding the difference branch alone already yields a substantial gain over the raw-only baseline, with F1 increasing. This confirms that frame-wise motion changes carry discriminative information beyond what the raw pose provides. Replacing the simple concatenation fusion with convolutional attention fusion further improves F1 and Accuracy, showing that per-channel adaptive weighting between the two branches is more effective than treating them equally.

Table 5 compares I-Robot under two pose representations. MHR outperforms SMPL across all metrics, with improved F1 and AUC. The performance gap suggests that the higher-dimensional MHR representation ($D = 381$) provides richer temporal cues for the difference branch, whereas the more compact SMPL body rotation matrix ($D = 207$) trades overall discrimination for better sensitivity.

6. Conclusion

We presented I-Robot, a motion-based framework for distinguishing humanoid robots from humans. We constructed the HvH dataset covering 38 humanoid platforms, and 11 action classes consist of human pose parameters. Our proposed I-Robot, which processes both raw pose and temporal difference sequences in parallel, consistently outperformed all standard sequence classifiers, confirming that how a subject moves is a more reliable cue than where its joints are at any single frame.

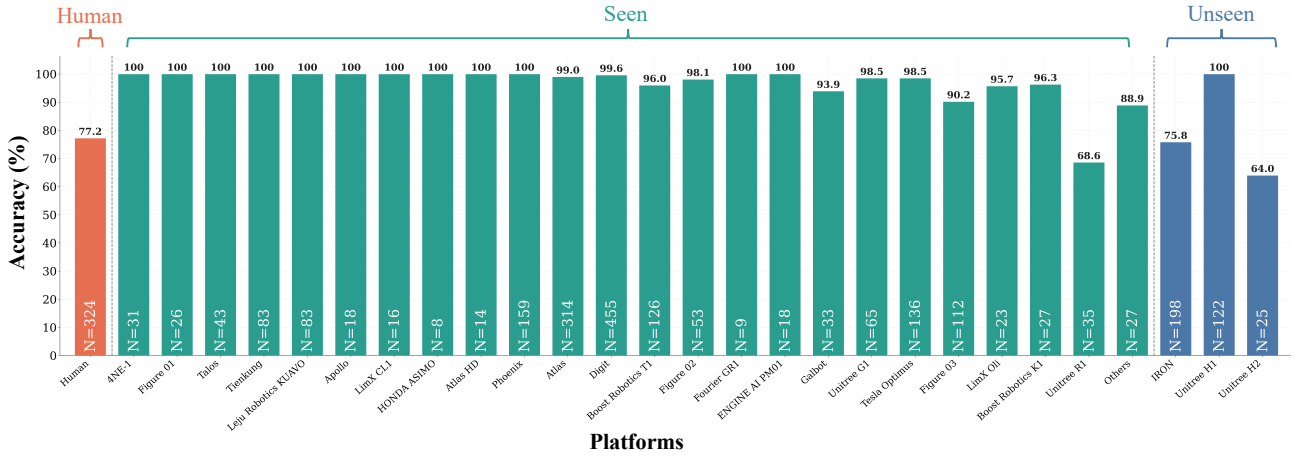

Figure 5. Per-platform accuracy of I-Robot on MHR parameters.

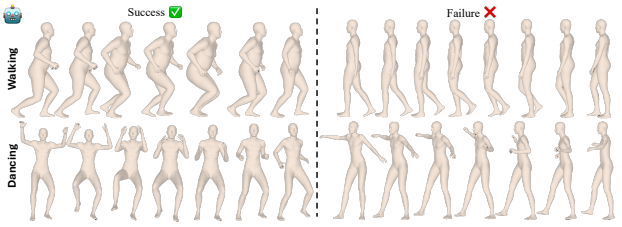

Figure 6. Rendered results of MHR pose parameters for walking and dancing action classes.

Limitations and future work. The current dataset is limited in size and collected entirely from YouTube, which introduces domain bias from demonstration-oriented footage, edited videos, favorable viewpoints, platform-specific recording conditions, and non-surveillance camera settings. Thus, our results do not fully verify generalization to realistic surveillance settings with low resolution, compression, occlusion, crowded scenes, or unusual viewpoints. I-Robot also depends on off-the-shelf pose estimators. Pose errors can directly affect prediction, especially for fast, irregular, or occluded motions. This may explain the lower performance on dynamic actions such as dancing and jumping. Future work should evaluate robustness across pose estimators and explore stronger temporal encoders, including transformer-based motion models. Finally, our formulation focuses on binary human-versus-humanoid classification. We do not yet evaluate whether the learned representation generalizes to unseen robot morphologies or partially non-humanoid agents, which remains an important direction for future work.

Discussion

Physical deepfake. This work is motivated by the need to study the problem before potential misuse of humanoids, such as alibi fabrication or surveillance evasion, becomes common. We refer to AI-driven motion that impersonates human movement in the physical world as a *physical deepfake*, in analogy to pixel-space and audio-space deepfakes that fabricate human appearance or voice. We do not claim that such misuse is already widespread, nor do we evaluate adversarially optimized humanoid motions; rather, we introduce human-versus-humanoid identification as an initial step toward studying this emerging problem—motivated by the precedent that digital deepfake detection became progressively harder only after synthesis quality had already advanced.

Beyond detection: broader applications. I-Robot is not limited to human-versus-humanoid detection; it can also serve as a motion scoring tool for human-robot interaction, industrial safety, and motion generation. As a scoring tool, its output can be interpreted as a continuous measure of how human-like or robot-like the motion appears. While concurrent research evaluates human-likeness by measuring how well humanoid robots imitate humans (Li et al., 2026), our work provides a complementary perspective: whether such motion remains distinguishable from natural human motion. This dual use is useful in human-robot interaction, where systems may need to respond differently to humans and robots (Abdulatif et al., 2018), in industrial safety, where identifying the type of moving agent can support risk-aware monitoring, and in motion generation, where the score can provide an automatic quality check of motion realism (Feng et al., 2026).

Acknowledgements

This work was partly supported by Institute of Information & Communication Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02263841, Devel-

opment of a Real-time Multimodal Framework for Comprehensive Deepfake Detection Incorporating Common Sense Error Analysis; RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)).

References

Abdulatif, S., Wei, Q., Aziz, F., Kleiner, B., and Schneider, U. Micro-doppler based human-robot classification using ensemble and deep learning approaches. In *2018 IEEE radar conference (RadarConf18)*, pp. 1043–1048. IEEE, 2018.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

Chao, H., He, Y., Zhang, J., and Feng, J. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8126–8133, 2019.

Chen, X., Xie, Y., Ma, H., Li, N., Liang, Y., and Guo, J. Crossgap: Unified face anti-spoofing via cross-modal global-aware prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3208–3215, 2025.

Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., and Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 13339–13348. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01311. URL https://doi.org/10.1109/ICCV48922.2021.01311.

Cho, K., Van Merriënboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Dariush, B., Gienger, M., Arumbakkam, A., Zhu, Y., Jian, B., Fujimura, K., and Goerick, C. Online transfer of human motion to humanoids. *International Journal of Humanoid Robotics*, 6 (02):265–289, 2009.

Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., and Yu, S. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9707–9716, 2023.

Feng, L., Jin, Y., Hu, H., and Chen, W. Biomechanical comparisons reveal divergence of human and humanoid gaits. *arXiv preprint arXiv:2602.21666*, 2026.

Ferguson, A., Osman, A. A., Bescos, B., Stoll, C., Twigg, C., Lassner, C., Otte, D., Vignola, E., Prada, F., Bogo, F., et al. Mhr: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025.

Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., and Chen, X. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1060–1069, 2022.

He, T., Luo, Z., Xiao, W., Zhang, C., Kitani, K., Liu, C., and 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.

Heess, N., Tb, D., Sriram, S., Lemmon, J., Merel, J., Wayne, G., Tassa, Y., Erez, T., Wang, Z., Eslami, S., et al. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*, 2017.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jeong, Y., Kim, D., Lee, J., Hong, M., Hwang, S., and Choi, J. mtofnet: Object anti-spoofing with mobile time-of-flight data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 38–47, 2022.

Jin, X., He, T., Zheng, K., Yin, Z., Shen, X., Huang, Z., Feng, R., Huang, J., Chen, Z., and Hua, X.-S. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14278–14287, 2022.

Jung, S., Lee, K., Jeong, Y., Noh, H., Lee, J., and Choi, J. Group-wise scaling and orthogonal decomposition for domain-invariant feature extraction in face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13372–13381, 2025.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7122–7131, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Kim, K., Kim, Y., Cho, S., Seo, J., Nam, J., Lee, K., Kim, S., and Lee, K. Diffface: Diffusion-based face swapping with facial guidance. *Pattern Recognition*, 163:111451, 2025.

Kocabas, M., Athanasiou, N., and Black, M. J. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5253–5263, 2020.

Kocabas, M., Huang, C.-H. P., Hilliges, O., and Black, M. J. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11127–11137, 2021.

Kolotouros, N., Pavlakos, G., Black, M. J., and Daniilidis, K. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2252–2261, 2019.

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., and Hager, G. D. Temporal convolutional networks for action segmentation and detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July*

21-26, 2017, pp. 1003–1012. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.113. URL https://doi.org/10.1109/CVPR.2017.113.

Li, M., Liu, M., Wu, Z., Lin, X., Zhang, J., Yan, M., Xie, Z., Zhang, C., Wen, C., Xu, L., et al. Towards motion turing test: Evaluating human-likeness in humanoid robots. *arXiv preprint arXiv:2603.06181*, 2026.

Li, Y., Lin, Y., Cui, J., Liu, T., Liang, W., Zhu, Y., and Huang, S. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In *9th Annual Conference on Robot Learning*, 2025.

Liao, R., Yu, S., An, W., and Huang, Y. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020.

Liu, Y., Jourabloo, A., and Liu, X. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 389–398, 2018.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Nakazawa, A., Nakaoka, S., Ikeuchi, K., and Yokoi, K. Imitating human dance motions through motion structure analysis. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pp. 2539–2544. IEEE, 2002.

Newell, A., Yang, K., and Deng, J. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pp. 483–499. Springer, 2016.

Peng, X. B., Berseth, G., Yin, K., and Van De Panne, M. Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning. *Acm transactions on graphics (tog)*, 36(4):1–13, 2017.

Peng, X. B., Abbeel, P., Levine, S., and 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.

Peng, X. B., Ma, Z., Abbeel, P., Levine, S., and Kanazawa, A. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

Peng, X. B., Guo, Y., Halper, L., Levine, S., and Fidler, S. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 41(4):1–17, 2022.

Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., and Wu, F. D-fine: Redefine regression task in detrs as fine-grained distribution refinement. *arXiv preprint arXiv:2410.13842*, 2024.

Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.-G., and Xue, X. Long-term cloth-changing person re-identification. In *Proceedings of the Asian conference on computer vision*, 2020.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Rudin, N., Hoeller, D., Reist, P., and Hutter, M. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on robot learning*, pp. 91–100. PMLR, 2022.

Safonova, A., Pollard, N., and Hodgins, J. K. Optimizing human motion for the control of a humanoid robot. *Proc. Applied Mathematics and Applications of Mathematics*, 78:18–55, 2003.

Shiohara, K. and Yamasaki, T. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18720–18729, 2022.

Stathopoulos, A., Han, L., and Metaxas, D. Score-guided diffusion for 3d human recovery. In *CVPR*, 2024.

Sun, K., Xiao, B., Liu, D., and Wang, J. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

Tang, W., Long, G., Liu, L., Zhou, T., Jiang, J., and Blumenstein, M. Rethinking 1d-cnn for time series classification: A stronger baseline. *CoRR*, abs/2002.10061, 2020. URL https://arxiv.org/abs/2002.10061.

Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., and Rigoll, G. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE international conference on image processing (ICIP)*, pp. 2314–2318. IEEE, 2021.

Toshev, A. and Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660, 2014.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12(1), 2016.

Yan, S., Xiong, Y., and Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yang, X., Kukreja, D., Pinkus, D., Sagar, A., Fan, T., Park, J., Shin, S., Cao, J., Liu, J., Ugrinovic, N., Feiszli, M., Malik, J., Dollar, P., and Kitani, K. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

Ze, Y., Chen, Z., Araújo, J. P., Cao, Z.-a., Peng, X. B., Wu, J., and Liu, C. K. Twist: Teleoperated whole-body imitation system. *arXiv preprint arXiv:2505.02833*, 2025.

Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., and Wang, Y. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15085–15099, 2023.