Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083418

ARTICLE

Real-Time Human Interaction Mimicry Teleoperation in Unitree G1 Edu Humanoid Robots Using the RGB Sensor

Yi Wen Tan[1], **Jun Meng Woh**[1], **Ee Sin Yong**[1], **Wai Leong Pang**[1], **Hui Hwang Goh**[1], **Kah Yoong Chan**[2] **and Ari Happonen**[3,*]

[1]School of Engineering, Faculty of Innovation & Technology, Taylor's University, Subang Jaya, Malaysia
[2]Faculty of Artificial Intelligence and Engineering, Multimedia University, Cyberjaya, Malaysia
[3]Software Engineering, School of Engineering Science, LUT University, Lappeenranta, Finland

*Corresponding Author: Ari Happonen. Email: ari.happonen@lut.fi

Received: 03 April 2026; Accepted: 23 June 2026; Published: 13 August 2026

ABSTRACT: Along with the rapid advancement of Artificial Intelligence (AI), humanoid robots are foreseen to have great potential in the service industry, where human interaction is unavoidable. However, current systems face significant hurdles, including Field of View (FoV) problems, markerless real-time mimicry capabilities for humanoid's fingers and arms. This study addresses these hurdles by developing an integrated hardware and software pipeline for the Unitree G1 Edu humanoid robot. A custom 3D-printed helmet and stabiliser interface were designed using FreeCAD and fabricated to house an external Orbbec Gemini 2 RGB-D sensor, optimising the FoV for frontal human-robot interaction by compensating for the robot's height and its downward-facing onboard sensors. Both finger control and arm control use MediaPipe Holistic to extract the coordinates and perform kinematics calculation to obtain the angle commands for each joint, which are then passed through a first-order Exponential Moving Average (EMA) Filter for jitter avoidance. The safety mechanism was verified in MuJoCo simulator, followed by physical implementation on G1 humanoid. Under single-subject test in controlled environment, MediaPipe shows perception robustness of almost 100% tracking reliability for arm-level mimicry across various distances, lighting and occlusion conditions, but experiences serious performance degradation under severe backlighting. Low Root Mean Square Errors (RMSE) of 5.3°–5.6° for shoulder elevation and 8.7°–8.8° for elbow flexion, demonstrating high retargeting reproducibility. Evaluation of the jitter suppression stage indicates that the EMA filter successfully attenuated high-frequency pose estimation noise, yielding a 40.3%–63.4% jitter reduction in shoulder channels and a 34.2%–38.5% reduction in elbow channels while maintaining an optimal balance with tracking responsiveness. Deployment of the algorithm onto G1 Edu humanoid successfully verified the feasibility of the developed algorithm. Combining the algorithm with the customised helmet, this system provides a new framework for human-mimicking robotics development. Future work will focus on sensor fusion, multi-subject and dynamic environment test to better synchronise robot movement with real-world interaction scenario, which are vital for moving the technology from the lab to public service applications.

KEYWORDS: Machine learning; humanoid; field of view; real-time processing; artificial intelligence; dynamic environment; industry 5.0; mimicry capability; RGB-D; human-robot interaction

1 Introduction

With the rapid emergence of Artificial Intelligence (AI), autonomous algorithms and collaborative robot operations with shared tasks with humans, the demand for human-robot interactions has been catalysed to rise. At the same time, Industry 4.0 and 5.0 eras are changing the work environments from human-only to human–robot collaborative environments, where these advances need security and safety considerations [1,2]. These developments have fundamentally transformed the way humans and robots interact. Humanoid robots have been increasingly deployed in roles that are largely tasked with performing direct interactions or cooperation with humans [3], or working as service providers in and for smart city solutions [4], mobility [5], industrial support roles [6], in virtualised robotics solutions, education [7], and so on. They can enhance customer experience, streamline operations, and provide personal assistance in the service industry. This trend reflects its versatility to expand its potential for commercialisation, driving humanoid robots to be integrated across multiple domains [8].

As early as 10 years ago, a study proposed that by 2030, humanoid robots would undertake at least 25% of the tasks within the service industry [9,10]. It has been proven in a survey-based study that humanoid robots can demonstrate their ability to act as a marketing differentiator, attracting unique attention that appeals to customers [11]. This highlights that businesses that integrate humanoid robots into their service will gain a potential competitive advantage [12]. Additionally, the global market value of humanlike service robots was predicted to grow at an annual rate of 25.34% from $11.48 billion in 2018 to $509 billion by 2025 [11]. From an economic perspective, integrating humanoid robots into the servicing sector can offer economic advantages by replacing or supporting human labour to reduce operational costs and offer cost-related benefits, such as solving the issue of high labour expenditures and workforce shortages [13]. However, the widespread adoption of humanoid robots in the service industry still depends on the acceptance and trust in social robots to perceive and integrate them into their everyday lives [14].

The deployment of a humanoid robot in the service industry is inherently dependent on the physical interactions between humans and robots [9]. Humanoid robots must possess visual perception capabilities and effective verbal and non-verbal communication, as they support robots to engage in smoother interactions [15,16]. The acceptance of the integration of humanoid robots in the service industry is ultimately influenced by anthropomorphism, social influences, and their perceived trustworthiness [17]. It was further indicated that different cultural environments will significantly modulate how people attribute anthropomorphism [18]. This overall acceptance towards it is highly tied to the usability of the robot for specific tasks. The work acceptance of robots is also heavily reliant on how the robots respond to humans. The non-verbal cues of the "thinking" and "mindset" of the robots are the attributes that are important for people to accept robots. These complex factors constitutionally emphasise a need for research methods to study the *in-situ* work between robots and humans so that the interactions from the robots can come across naturally and seamlessly [19]. Since human movements are a continuous pipeline of complex dynamical behaviours in a closed-loop manner, it is crucial to address this challenge that affects the accuracy of the detection model [20]. Hence, a system that perceives and processes human action in real-time remains a hurdle that future studies must overcome.

Despite the rapid advancement of technology, the adoption of humanoid robots in the industry remains relatively low [21]. Most existing work or systems focus on recognising static postures in constrained environments. They do not effectively address the challenges of dynamic real-world scenarios where human interactions are continuous, complex, and unpredictable. Variations in human motion, occlusions, and interaction will lead to an unstable landmark representation, affecting the system's ability to understand the motion. Moreover, many humanoid robots do not fully optimise the visual Field-of-View (FoV) due to the placement of onboard sensors tailored for specific use cases. It does not possess the necessary sensor versatility when applied to other use cases, affecting the FoV that is needed within the project. The integration of humanoid robots into social environments that can interact with humans requires a system that is able to perceive and respond to human actions accurately [22,23]. Naturally, the visual FoV and temporal dynamics of human movement are crucial in realising a reliable human-robot interaction system.

In this study, three primary challenges are addressed: (1) the limited research regarding teleoperation frameworks for the dynamic, real-time control of finger articulation; (2) the restrictive dependency on wearable sensors or intrusive tracking infrastructure for human-to-humanoid mimicking; and (3) the inadequacy of the G1 humanoid's downward-facing visual FoV for human-centric interactions. To address these limitations, this research provides the following contributions: (1) the implementation of a markerless joint-position extraction and kinematic mapping approach using MediaPipe, enabling finger mimicry without wearable sensors; (2) the design of a unified vision-based teleoperation framework that synchronises both arm and fine-motor finger articulations to achieve seamless, real-time anthropomorphic mapping in controlled environments; and (3) the development of a custom 3D-printed humanoid head casing holding an external camera in front-view orientation, optimising the visual FoV for frontal human-robot interaction.

The remainder of this paper is organised as follows: Section 2 reviews related works that are relevant to the study, including existing solutions for the optimisation of the FoV in humanoid robots and the humanoid arm and finger mimicking method used. Section 3 details the proposed system and the architecture, while Section 4 presents the experimental results and discussions. Lastly, Section 5 concludes the paper and summarises the key findings and contributions of this paper.

2 Related Works

2.1 Field of View Existing Solutions

Various implemented solutions to address insufficient FoV are mentioned in multiple research papers. The placement of an RGB camera on top of the Light Detection and Ranging (LiDAR) sensor module is proposed, providing an elevated view for the humanoid robot [24], mimicking the added safety on actions, based on elevated view, as also utilized, e.g., in self-driving and robotics cars context too. The humanoid robot's head module consists of an RGB panoramic camera on top and a LiDAR module below the camera sensor. This approach captures the panoramic input video stream of the surroundings while maximising the FoV of the LiDAR module. An elevated input video stream compensates for the insufficient height of the Unitree G1 Edu, aiding the humanoid robot in capturing more human interaction.

Instead of passive FoV solutions, an active head module design is proposed to resemble the sensory-motor capabilities of humans [25]. A certain Degree of Freedom (DoF) is added to the design of the humanoid robot's head module by implementing servo motors for rotational movement, allowing the humanoid robot to shift its gaze around the environment and scan the environment sequentially instead of passively processing imagery that falls within the camera view. Another work also mentions an active vision application on the stereo camera setup [26]. This implementation emphasised the advantage of biomimicry of eye movement mechanics based on the fish eye that allows independent control of each camera sensor, expanding the FoV even further as compared to conventional fixed FoV solutions [24]. With precise calibration performed based on real-time updated extrinsic parameters such as the rotational matrix and translation vector, it eliminates the need for constant recalibration done offline.

While active vision adjusts the humanoid robot's perspective via motorised camera movement, it adds complexity by requiring additional DoF on the robot's hardware. Latency is also introduced when extra components are required. Hence, panoramic sensors such as the LiDAR sensor are used to provide a wide horizontal view and safety extension on the tasks done at a large workplace, even while it is in a stationary setting [27]. The sparse 3D point cloud from the LiDAR sensor is outlined and processed via the Time-Aware Attention Pooling (TAP) mechanism to monitor the object's temporal movement over time, allowing the target object to be detected even if it is out of the RGB-D camera's FoV. Besides, LiDAR allows the detection in low-light environments, resulting in better collision avoidance compared to utilising the RGB-D camera only. Additional RGB-D sensors can be added to the system to provide a panoramic view of the humanoid robot's surroundings, addressing the limited FoV constraints of robot vision [28]. Three monocular camera sensors are incorporated around a planar rotating platform equipped with a front-facing stereoscopic RGB-D camera. The monocular camera acts as a finder in detecting a desired object, followed by high-precision rotational movement of the platform to allow the stereoscopic sensor to capture detailed imagery, which reached 95% of positive detection rate, as well as low latency time ranging between 110 and 160 ms. This approach eliminates the need for LiDAR installation into the robot vision system.

Table 1 presents multiple active and passive solutions for the FoV, along with testing to prove their feasibility for implementation with a humanoid robot. However, all the mentioned solutions are based on a newly designed ground-up build instead of an improvement for the existing model with an external attachment that could be installed on the humanoid robot, which is the Unitree G1 Edu. Thus, the humanoid robot accessories are focused instead of a ground-up design, while adopting the solution for insufficient FoV proposed above. The design of the FoV solution should consider the hardware bottlenecks of the Unitree G1 Edu's Nvidia Jetson Orin on board computer to avoid hardware incompatibility.

Table 1: Summary of FoV solutions.

Work	FoV Solution	Robot Vision System	Contribution to FoV Constraints
[24]	Elevated RGB panoramic camera on LiDAR	Passive	Extends FoV at a higher elevation
[25]	Servo for gaze shifting	Active	FoV for object detection
[26]	Head module biomimicry fisheye camera	Active	Independent control of camera sensor
[27]	LiDAR integration to generate point cloud	Passive	Reduced collision in variable lighting
[28]	Rotating monocular and stereoscopic camera	Active	High-speed 95% detection accuracy

2.2 Humanoid's Arm and Finger Mimicking Method for Human Action-Mimicking Purpose

Finger control teleoperation focuses on the data-capturing method and the data-processing method. For the data-capturing method, most research relies on the sensor equipment. To control the dexterous fingers, the researchers fixed five Arduino-connected bending sensors to a glove to capture the data about the bending motion of fingers, further processed by a weighted Exponentially Moving Average (EMA) filter to smooth the trajectory, and the resulting data will be sent to the joint via serial port [29].

An advanced study integrates both an exoskeleton glove with potentiometers at the joints and a vision-based Leap Motion Controller (LMC) to control the bionic hand [30]. Both sensor readouts from gloves and joint coordinates from LMC pass through feature extraction and a Kalman Filter. The resulting parameters will be processed at the selector and mapped to output bending angles for each joint. The joints are controlled using Pulse Width Modulation (PWM) calculated from the inverse kinematics via MATLAB. Although both methods successfully performed the tasks tested, the glove method has slightly better performance due to the stable and precise feedback from the gloves compared to the vision-based method.

A glove with an Inertial and Magnetic Measurement Unit (IMMU) is utilised to locate at each segment from fingers up to upper arms, instead of joints used to capture the DoF movement for each joint [31]. Along with a multimodal fusion algorithm, each data point from IMMU is processed and passed through an extended Kalman filter, resulting in orientations and positions for each segment. The developed device successfully performs arm and finger teleoperation on 10-DoF and 11-DoF robotic arms.

Another study proposed a whole-body teleoperation using a Virdyn IMU-based full-body motion capture suit that requires the interactor to put on to capture joint parameters for mimicking [32]. The data obtained for each limb's joint are mapped to the nearest functional joint, and the equivalent angles are computed and fed into a lightweight EMA filter to reduce jittering. This method was validated in the MuJoCo simulator and applied to a physical G1 humanoid, with the balancing of G1 needs to be considered during high-speed dynamic motion. While external devices yield higher precision, they require additional cost and physical constraints for deployment in dynamic environments.

Two synchronised RGB cameras are utilised to capture the motions of researchers to control the Quanser QArm [33]. Using MediaPipe and the stereo triangulation method, 3D mapping of finger joints is calculated and mapped to the inverse kinematics to instruct the joint of Quanser QArm to move correspondingly. The paper successfully performs real-time trajectory path tracking and binary gripper actuation at a 200 Hz operational frequency. Another research called AnyTeleop tries to unify the vision-based teleoperation for different robotics arms by proposing a universal architecture [34]. The architecture fits any kind of input methods, such as RGB-D, RGB, and multiple cameras, while the detection for hand pose uses different processing methods. MediaPipe is used for finger extraction, while pixel depth position is extracted for RGB-D setup to obtain the depth information, but for RGB setup, integration of FrankMocap, an additional neural network to estimate depth based on the hand's scale, and MediaPipe to roughly estimate the depth. After detection fusion, human pose retargeting is performed to process the data by decoupling wrist tracking from finger tracking, utilizing an inverse kinematics solver to generate the arm's 6-DoF spatial trajectory, while concurrently executing an optimisation-based keypoint-vector mapping to compute the finger joint angles for collision-free motion, which was successfully deployed on different robotics arms and setups.

Research proposed a markerless vision-based teleoperation framework designed to transfer human dual-hand manipulation skills to a bionic bimanual robot [35]. Data acquisition is achieved through a low-cost monocular setup utilising a single RGB camera to capture unconstrained dual-hand motions. To process the visual stream, the system deploys a dual-hand detection network that embeds upper-body structural constraints to accurately localise and distinguish the hands, alongside a 3D hand pose estimation network optimised via a bone-constraint loss to reconstruct joint configurations. Finally, these extracted parameters are mapped to the robotic platform, leveraging global hand trajectories to drive the dual-arm endpoints while utilising the 3D hand poses to actuate a pair of five-finger dexterous manipulators.

Unlike wheeled robots, bipedal or legged robots are prone to the risks of collision with objects such as obstacles or even a human, especially in heavily occluded environments. This highlights the importance of incorporating a recovery control system into the humanoid to nullify the effect of pushing when a collision occurs and provide higher stability while the bipedal humanoid robot is cruising [36]. Instead of relying on conventional analysis models, the proposed system integrates Deep Neural Network (DNN) with Reinforcement Learning (RL) and algorithms such as Soft Actor-Critic (SAC), allowing data to be mapped directly to commands. This approach eliminates the high computational cost of the hardware, where Deep Reinforcement Learning (DRL) only requires high computation during the training phase, while inference takes near instantaneously to react. However, multiple scenarios of push or collision must be included in the training dataset, which requires millions of trials in simulators, where actual sensor performance might deviate.

Despite how much more freedom of motion from device constraints, most of the markerless approaches require special device calibration knowledge, which is not user-friendly. Hence, this study presents a vision-based, markerless, and user-friendly arm-hand mimicking architecture optimised for humanoid applications.

3 Proposed System

3.1 Proposed Field of View Solution: Humanoid's Helmet as External Camera Housing

Fig. 1 shows the comparison of the height of the G1 humanoid, which is 1.32 m, with the average height of females and males, along with the downward angle of FoV, showing a serious difficulty in capturing the interactor's action in a normal scenario if using the on-board Intel RealSense D435i RGB-D sensor. An external application of the Orbbec Gemini 2 RGB-D sensor is integrated into the G1 humanoid by placing it on top of the Unitree G1 Edu's head module, providing a parallel FoV to capture more human interaction, at the same time compensating for the insufficient height of the humanoid robot.

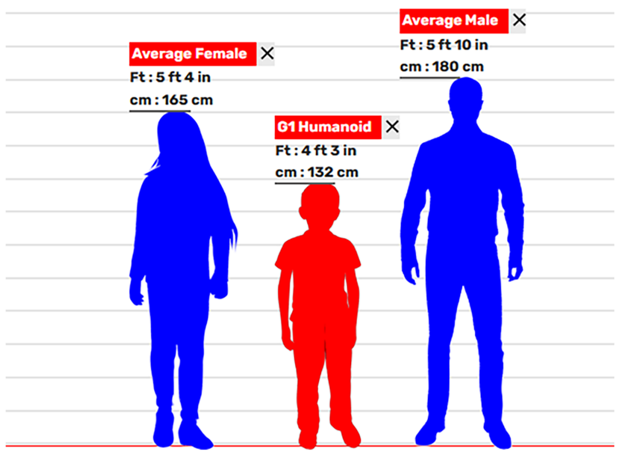


Figure 1: Comparison of G1 humanoid's height with the average height of females and males.

A helmet attachment is designed with open-source 3D Computer-aided Design (CAD) software, FreeCAD. Manual measurement of the head module to allow the designed helmet to fit well on the head, along with a stabiliser bar that locks the helmet on the head module, preventing the helmet from falling off the humanoid robot. As for the camera mounting, it is designed separately according to the official document regarding its measurement, followed by the part assembly function in FreeCAD to assemble both designs into one piece for 3D printing. A hollow space is left at the rear side of the camera mounting to allow passive cooling of the Orbbec Gemini 2, preventing it from overheating during operation.

3.2 Real-Time Kinematic Retargeting System for Action-Mimicking Purpose by Humanoid

The system developed presents a markerless teleoperation system that maps movements from the human upper body and fingers onto a G1 humanoid robot in real time. To realise this system, the robot joint command will be derived continuously, on a per-frame basis, from the live geometry of the human's body.

The pipeline of this system can be differentiated into 2 sections, where one is an arm mimicry pipeline, and the other is a dexterous-hand pipeline. The human pose and hand estimation pipeline begins by capturing human motion through an external RGB sensor and then processed using MediaPipe Holistic. MediaPipe Holistic is a pre-trained inference pipeline that estimates full-body pose and hand articulation from a single RGB stream. The pipeline yields 33 body-pose landmarks and 21 landmarks per hand expressed as a normalised image-plan coordinate. The detection and tracking confidence thresholds are set to 0.5 and 0.6 for the arm and finger pipeline, respectively. The finger pipeline has a stricter confidence threshold because the finger retargeting is more sensitive to landmark noise compared to large-scale arm motions. Moreover, the arm pipeline only processes at every third frame to balance perception latency and reduce inference load while maintaining an update rate that is sufficient for smooth motion. Other than that, a per-landmark visibility threshold of 0.6 is applied so that the motor joints are triggered only when the human landmarks are reliably detected. For safety measures, if a landmark falls below this preset threshold, the motor joints will revert back to a defined pose to prevent any sudden motion that causes commotion.

Once the body and finger landmarks have been extracted, these landmarks need to be translated into commands so that the humanoid can execute them. Since humans and robots differ in coordinate frames, link proportions, and joint conventions, there exists a need to translate the landmarks so that the robot is able to understand and execute the motion. Moreover, the coordinates from MediaPipe are also image-plane coordinates only, rather than joint angles. Therefore, the retargeting stage is necessary to resolve this mismatch by converting geometric relationships among the landmarks into joint targets that are respective to each actuator joint's sign convention. Crucially, this conversion is performed independently for each frame from the instantaneous pose of the human so that the humanoid reproduces continuous motion. Each DoF is treated independently and derived from the landmark as detailed in the following sections.

Starting with the elbow flexion, the elbow target will be obtained from the real geometric bend angle of the human at the elbow joint. Given the landmark positions of the shoulder, elbow, and wrist, the upper-arm vector is defined as $\vec{v_1} = P_{shoulder} - P_{elbow}$, and the forearm vector is defined as $\vec{v_2} = P_{wrist} - P_{elbow}$. Then, the angle between these two vectors will be computed using Eq. (1).

$$\theta_{elbow} = \arccos\left(\frac{\vec{v_1} \cdot \vec{v_2}}{\|\vec{v_1}\| \, \|\vec{v_2}\|}\right) \tag{1}$$

In MediaPipe, it was verified that a fully extended arm will yield an $\theta_{elbow} \approx \pi$. However, the G1 humanoid robot's elbow actuator actually adopts a value of 0 for a fully extended arm. Thus, the conversion between these two conventions will be performed using Eq. (2).

$$q_{elbow} = \pi - \theta_{elbow} \tag{2}$$

where q_{elbow} is the commanded elbow joint target. To ensure robustness against degenerate landmark configurations, a numerical safeguard is applied, whereby if a vector norm falls below a value of 10^{-4}, the target will be defaulted to a straight arm value.

Moving on to the shoulder roll, the shoulder roll governs the elevation of the arms. When experimenting, it was found that a direct mapping based solely on the vertical image displacement of the wrist was unreliable because the wrist will couple to arm elevation with a forward-reaching motion. To decouple these effects, the true elevation angle of the arm will be estimated geometrically from the relative displacement of the wrist to the shoulder. Let the horizontal and vertical wrist-to-shoulder displacements be defined as $dx = x_{wrist} - x_{shoulder}$ and $dy = y_{wrist} - y_{shoulder}$, respectively, the arm elevation relative to the horizontal is stated in Eq. (3).

$$\phi_{arm} = \arctan 2\,(dy, |dx|) \tag{3}$$

where ϕ_{arm} is the angle of elevation of the arm. This angle will then be mapped to the actuator using Eq. (4). As experimented, a horizontal arm corresponds to approximately 1.47 rad, while an arm lowered at the side will correspond to approximately 0 rad.

$$q_{roll} = 1.57 - \phi_{arm} \tag{4}$$

where q_{roll} is the commanded shoulder roll target. The usage of the absolute value of dx ensures that the estimated elevation will be independent of whether the arm moves forward or backwards.

As for the shoulder yaw, it is responsible for capturing the forward and backward horizontal sweep of the arm. It is derived from the normalised horizontal displacement of the wrist relative to the shoulder. To prevent the arm from moving backwards, the shoulder yaw is also scaled and clipped to constrain the commands within a safe operating range, as shown in Eq. (5).

$$q_{yaw} = clip(\frac{-dx}{0.4}, -1, 1) \times 1.5 \tag{5}$$

where q_{yaw} is the commanded shoulder yaw target, and the 0.4 constant is a normalisation width.

Another crucial aspect to take note of is that the left and right arms of the G1 humanoid robot are mirror-symmetric. Thus, for every human motion tracker, an opposite joint sign must be applied on certain joints, depending on the handedness. Additionally, to avoid any unsafe configurations or mechanical stress on the physical robot, every actuator is set to have a physical travel limit that will not be exceeded. Hence, a handedness-dependent sign, $s \in \{-1, +1\}$ is applied. Following that, the actuator range was also clamped to $[q_{min}, q_{max}]$ and is expressed in Eq. (6).

$$q_{cmd} = clip\left(s \cdot q_{raw}, q_{min}, q_{max}\right) \tag{6}$$

where q_{cmd} is the final commanded joint value, q_{raw} is the retargeted joint target before correction, q_{min} denotes the lower actuator limits, and q_{max} corresponds to the upper actuator limit. The signing convention is fixed to each joint and arm. For instance, the right shoulder roll will be fixed to a negative sign for a correct uplift motion, while the left shoulder will be fixed to a positive sign. These clamp bounds will guarantee that any command issued by the retargeting stage beyond its fixed stage will not drive a physical actuator beyond its allowable mechanical range.

Moving on, for finger control, it will be derived from the bend angle at each finger's Proximal Interphalangeal (PIP) joint. It is computed using Eq. (1), but with the appropriate finger landmark instead. For the PIP angle, it will be normalised by a maximum curl $\theta_{max} = 2.5\ rad$ to obtain a curl ratio bounded in $[0, 1]$ as presented in Eq. (7).

$$c_{finger} = clip\left(\frac{\theta_{PIP}}{\theta_{max}}, 0, 1\right) \tag{7}$$

where c_{finger} is the normalised curl ratio of a given finger. The commanded finger value is obtained using Eq. (8).

$$q_{finger} = 1 - c_{finger} \tag{8}$$

where q_{finger} is the commanded finger position. This formulation is applied to the pinky, ring, middle, index fingers and thumb flexion, independently. As for thumb opposition, which is the rotation of the thumb across the palm, it will be treated as a separate DoF as it cannot be captured by a simple bend angle. It will be estimated from the angle between the base of the thumb and the base of the index, where both are measured relative to the wrist. Then, it is mapped such that a small angle corresponds to the closed state and a large angle corresponds to the open state.

After obtaining all the joint targets for all mapped DoF, the pipeline will proceed to a filtering stage prior to issuing the commands to the robot. Since the estimation of raw landmarks will contain frame-to-frame jitter, if they are transmitted directly to the physical humanoid, visible instability can be noticed. Thus, temporal smoothing will be applied where each joint target will be passed through a first-order EMA filter. Given a joint with a smoothing gain of $\alpha \in [0, 1]$, the filter command at frame t is calculated using Eq. (9).

$$q_t = \alpha q_{target} + (1 - \alpha) q_{t-1} \tag{9}$$

where q_t is the filtered command at the instantaneous frame, q_{target} is the retargeted target at the instantaneous frame, q_{t-1} is the filtered command at the previous frame, and lastly α is the smoothing gain. For the smoothing gain, if a larger value is applied, it will result in a more responsive human motion but higher jittering. Whereas, if it is a smaller value, it will produce smaller motion but with a cost of higher latency. Thus, to balance this, the gain was tuned according to the characteristic dynamics for each joint.

3.3 Physical Test Setup

The control algorithms were transitioned into a modular, hybrid-language software architecture to facilitate isolated troubleshooting and system debugging. The initiation of the RGB-D camera, MediaPipe, and kinematics are saved in Python files, while the hardware bridge for arm control and finger control are saved as C++ files. The rationale of the arrangement allows the continuous execution of physical actions with the previous state of the joint angles to cope with the delay caused by the slower kinematics calculations. Data Distribution Service (DDS) middleware is incorporated as the transport medium of the computed skeletal coordinates handled by the Python script that acts as master, whereas the receiving slave C++ script performs movement mimicry on the Dexterous Hand based on the received input.

To enable an Internet connection on the Unitree G1 to install the required dependencies, such as the Software Development Kit (SDK) for the hardware, MediaPipe, and transfer our developed code to the host computer, communication between the Unitree G1's onboard Nvidia Jetson Orin NX and the computer is established via Secure Shell (SSH) Protocol with an Ethernet cable. The humanoid robot is also connected to the Internet connection to allow local installation of dependencies for Advanced RISC Machine (ARM) Instruction Set Architecture (ISA), to avoid incompatibility, as the team's computer is based on X86 ISA.

The Orbbec Gemini 2 external camera is connected to the Nvidia Jetson Orin NX via a USB 3.0 cable that supports up to 10 Gbps of bandwidth, to minimise hardware bottlenecks. Before initiating the camera, the virtual environment (venv) is activated to avoid version conflicts of the newly installed dependencies.

4 Results

The results are discussed for helmet design, mimicking pipeline verification, and physical humanoid result.

4.1 Prototype of Camera Helmet for Humanoid Robot

As demonstrated in Fig. 2a,b, the 3D-printed design of the helmet with camera housing fits on the Unitree G1 Edu humanoid robot, despite having some extra gaps at the side of the helmet that do not contact the surface of the head, which is addressed by the stabiliser bar. The external Orbbec Gemini 2 now has a parallel and elevated view of the fixed FoV, instead of a depressed view found on the built-in Intel RealSense D435i. The RGB-D camera is fixed into its camera housing via two M3 threaded screws.

(a) (b)

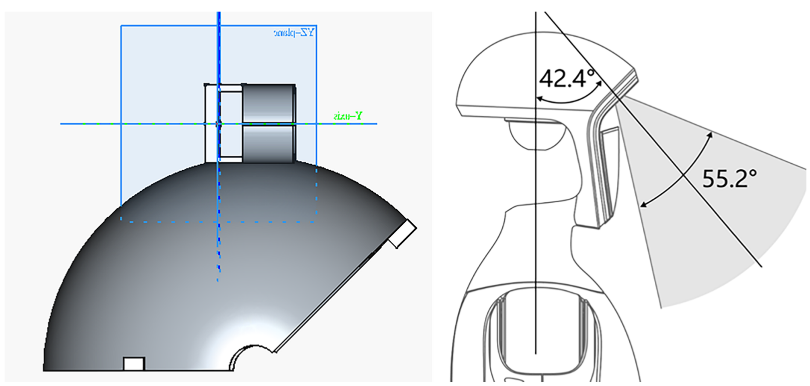
Figure 2: (**a**) Helmet fitting on Unitree G1; (**b**) 3D-printed camera housing with stabiliser bar.

The specifications of the on-board Intel Realsense D435i and the external Orbbec Gemini 2 is tabulated and compared in Table 2, in which the latter excels in providing wider FoV and longer detection, with similar resolution of the RGB camera.

Table 2: Comparison between the FoV and range of the RGB-D cameras.

Aspects	Intel Realsense D435i	Orbbec Gemini 2 RGB
Maximum Range Detection (m)	8.1 FPS	33.6 FPS
Camera's FoV	102.61 ms	107.05 ms
RGB Stream Output's Resolution	2000	2000

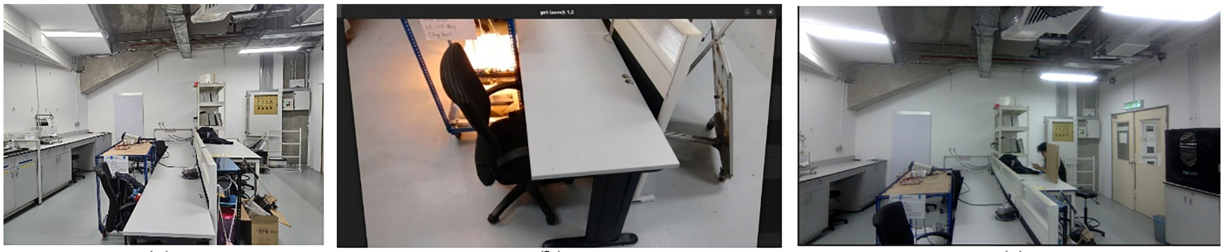
Fig. 3a illustrates the camera placement of the Orbbec Gemini 2 on the Unitree G1 helmet in FreeCAD Computer-Aided Design (CAD) software. The centre of the camera direction forms a parallel line with the Y-axis, thus there is zero angle of depression on the RGB-D camera when mounted on the humanoid robot. The onboard Intel Realsense camera has the camera fixed with an angle of depression of 47.6° as shown in Fig. 3b. Combined with shorter height of the Unitree G1, the latter excels in performing task on table, whereas the former provides parallel FoV for the robot.



(a) (b)

Figure 3: (**a**) The side view of the helmet in FreeCAD; (**b**) FoV of the Intel RealSense D435i.

The Unitree G1 is hung in the laboratory, facing a view with a desk and a rack as shown in Fig. 4a. The on-board Intel RealSense D435i could capture a close-distance view in the lab displayed in Fig. 4b, thus not feasible in human-robot interaction with a standing person. The external Orbbec Gemini 2, however, provides an elevated and parallel view of the same location as displayed in Fig. 4c, allowing the robot to capture skeletal movement of humans and their expression for movement mimicry.



(a) (b) (c)

Figure 4: (**a**) Environmental view of Unitree G1 facing; (**b**) FoV of Intel RealSense D435i; (**c**) FoV of Orbbec Gemini 2.

4.2 Developed Mimicking Pipeline

The proposed system was evaluated using three quantitative tests, including the perception robustness test, retargeting reproducibility, and smoothing behaviour, and one qualitative test comprising a physical robot demonstration. However, the quantitative results will not be available for finger mimicry as there is no simulator available for the specific set of hands. All evaluations were conducted with a single subject within 1 m of the sensor under normal lighting conditions unless stated otherwise.

4.2.1 Perception Robustness

The perception reliability test was carried out on the helmet-mounted RGB sensor across variations in distance, lighting, and partial occlusion. Using MediaPipe Holistic and the same configuration mentioned in Section 3, the reported results reflect the reliability of the deployed system. The robustness evaluation establishes the conditions under which the perception front-end can supply usable input. Table 3 presents the tracking rate across different operating conditions for a single subject. In the retargeting pipeline, the arm-level tracking is the most crucial channel that drives it. It can be observed that at the initial stage, the arm-level tracking remained at almost 100% across the entire tested distance range, under dim lighting, and also under every occlusion condition. The results demonstrated that the body-pose channel responsible for generating the humanoid control commands has remained stable under every variation, even in scenarios where the hand tracking channel experienced significant degradation. The most severe degradation was seen to occur under strong backlighting, where the pose detection fell to a detection accuracy of 88.4%, and the right-hand and left-hand detection dropped to 65.2% and 43.3%, respectively. This is consistent with the fact that a typical RGB-based perception front-end often struggles when foreground contrast is reduced. These findings indicate that the continuous arm-mimicry capability can operate reliably under conditions that are commonly encountered within real-world interaction sessions.

4.2.2 Retargeting Reproducibility

The retargeting reproducibility of the output was evaluated by measuring the steady-state deviation between the generated joint command and comparing it against a recorded reference command that corresponds to the intended posture across three trials. This evaluation quantifies how consistent the targeting pipeline converts the retargeting input into robot commands. However, it is important to note that this does not constitute an assessment of absolute joint-angle accuracy against the physical ground truth, as the equipment was not available during experimentation. Table 4 summarises the reproducibility results obtained across the evaluated trials. Shown in Table 4, the shoulder-roll channel reproduced the joint command with a Root Mean Square Error (RMSE) of approximately 5.3°–5.6°, while maintaining a low Standard Deviation (SD) of no more than 4.64° for both arms. The consistent performance of the left and right arms indicates that the handedness-dependent sign and clamp corrections in the retargeting equations established in Eq. (6) are applied consistently and do not introduce any asymmetry in the retargeting. Thus, it can be inferred that the mapping for elevation can be considered reproducible within a single-subject scope. On the other hand, the elbow channel exhibited lower reproducibility as the RMSE values declined to values of 8.7°–8.8°. Additionally, the right elbow demonstrated the highest variability in the SD of 6.59°. This is due to the fact that elbow flexion is estimated from the angle that is formed by the shoulder, elbow, and wrist landmarks, making it particularly sensitive to the ambiguity of the depth that is inherent to the RGB pose estimation. This sensitivity will become more obvious when the elbow articulation frequently occurs along the camera viewing axis, where the estimation errors of small landmarks can lead to large angular deviations. As a result, the results related to the elbow should not be seen as a deficiency specific, instead, it could be interpreted as an explainable limitation caused by the monocular input.

Table 3: Perception validation tracking rate across varying operating conditions.

Condition	Parameters	Pose (%)	Arm Usable (%)	Both Arms (%)	L-Hand (%)	R-Hand (%)
Distance	0.5 m	100.0	100.0	100.0	100.0	100.0
	1.0 m	100.0	100.0	100.0	100.0	100.0
	1.5 m	100.0	100.0	100.0	100.0	100.0
	2.0 m	100.0	100.0	100.0	87.6	100.0
Lighting	Normal	100.0	100.0	100.0	100.0	100.0
	Dim	100.0	100.0	100.0	97.3	77.1
	Backlit	88.4	88.4	88.4	65.2	43.3
Occlusion	None	100.0	100.0	100.0	100.0	100.0
	Left Arm	100.0	100.0	100.0	81.7	99.4
	Right Arm	100.0	100.0	100.0	100.0	96.3

Table 4: Steady-state retargeting reproducibility results.

Joint	Ref	Reference (def)	RMSE (deg)	SD (deg)
Left Shoulder Roll	0.97	+87.5	5.28	4.64
Right Shoulder Roll	0.97	−88.8	5.59	4.18
Left Elbow	0.97	+18.1	8.84	3.55
Right Elbow	0.97	+14.3	8.75	6.59

4.2.3 Filtering and Jittering Suppression

The jitter suppression capability of the filtering stage was evaluated by measuring the landmark deviation between the raw landmark targets and the smoothed commands sent to the robot during a continuous dynamic motion test. This isolates the EMA filter's attenuation of high-frequency pose estimation noise from its behaviour. Table 5 presents the deviation and jitter reduction results collected across the evaluated trials. As shown in Table 5, the shoulder roll channels that applied a filter gain of 0.25 reduced frame-to-frame jitter by 40.3%–63.4%. The difference between the two arms reflects the differing noise content of the raw landmarks, where it can be seen that for the left roll, it was approximately 1.2 times noisier than the right roll. As for the elbow channels, with a higher filter gain to preserve responsiveness, they reduced jitter by 34.2%–38.5%. This is consistent with the fact that it is expected that a smaller attenuation will result in lighter smoothing. This caused a trade-off between the degree of noise suppression and responsiveness on a faster elbow articulation.

Table 5: Deviation and jitter reduction results.

Joint	Filter Gain (α)	Raw vs. Fil RMSE (deg)	Jitter Reduction (%)
Left Shoulder Roll	0.25	14.43	63.4
Right Shoulder Roll	0.25	11.86	40.3
Left Elbow	0.60	4.56	34.2
Right Elbow	0.60	5.97	38.5

4.2.4 Runtime Performance

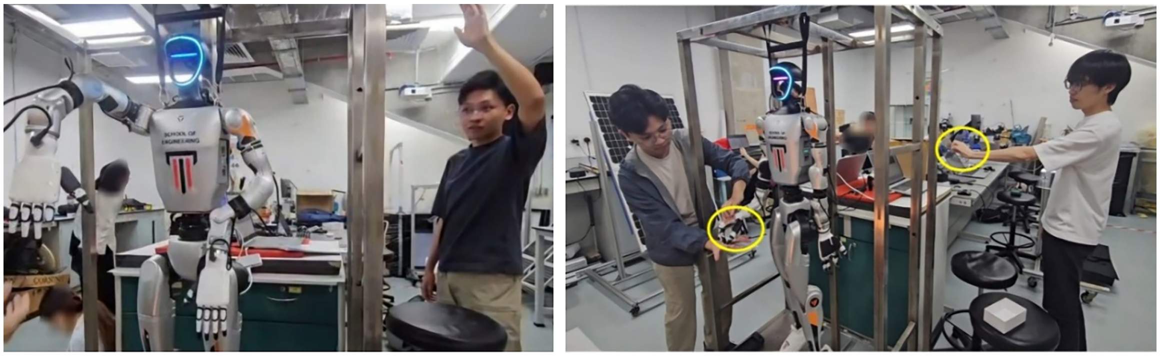
The runtime performance of the retargeting pipeline was evaluated on the G1 onboard computer to validate its real-time feasibility. The performance was assessed by comparing 2 configurations, whereby the baseline pipeline is executed on every frame, whereas the optimised pipeline is executed every third frame for computational purposes. This can maintain a control rate that is sufficient for upper-body teleoperation. Table 6 summarises the runtime results logged across 2000 frames. Based on Table 6, the optimised pipeline achieved a mean effective control rate of 33.6 FPS. This demonstrates that the system can operate within the real-time threshold on the G1 onboard hardware. Conversely, the baseline pipeline achieved only 8.1 FPS as the computational load of the onboard processor is limited. This four-fold difference demonstrates that the optimised baseline can accommodate the necessary inference speed that is needed for real-time deployment. Furthermore, the average vision inference latency remained constant at approximately 107 ms, which shows that per-frame inference cost is stable.

Table 6: End-to-end pipeline runtime on the Unitree G1 onboard computer.

Metric	Baseline Pipeline	Optimised Pipeline
Mean FPS	8.1 FPS	33.6 FPS
Average vision inference latency	102.61 ms	107.05 ms
Frames logged	2000	2000

4.2.5 Physical Demonstration of G1 Humanoid Robot

The mimicry pipeline was deployed and demonstrated on the physical Unitree G1 Edu humanoid robot. Fig. 5a depicts the main frame in which the operator raises his arm and the humanoid robot reproduces the corresponding arm configuration. This demonstration provides direct evidence that the proposed mimicry pipeline is capable of converting upper-body human motion into physically executable humanoid joint commands under real-life scenarios. Furthermore, Fig. 5b illustrates the finger control, where it can be seen that the operator's hand closing and the humanoid robot reproduce the action to grab an object.



(a) (b)

Figure 5: Physical demonstration on G1 humanoid. (**a**) Raised arm configuration; (**b**) Closed hand configuration.

5 Conclusion

The research has successfully established a functional pipeline for vision-based arm and finger mimicking on the Unitree G1 humanoid platform. By identifying the height and FoV limitations of the Unitree G1 Edu, the project successfully designed and fabricated a custom 3D-printed helmet and stabiliser interface that optimises the visual range for frontal interactions. The integration of the Orbbec Gemini 2 camera allowed the system to overcome the depressed view of the on-board sensor for effective human motion capturing. Future research will explore LiDAR-based sensor fusion to enhance arm velocity estimation and situational awareness, which are critical for developing safety mechanisms in applications. The tracking framework utilising MediaPipe and an EMA filter was validated across perception robustness, retargeting reproducibility, and jitter suppression trials. The arm-tracking algorithm achieved near 100% accuracy under varied occlusion states and operational distances from 0.5 to 2.0 m, demonstrating robust performance independent of operator proximity. Environmental illumination, however, significantly impacted tracking fidelity. While normal lighting yielded perfect detection, dim conditions reduced separate left- and right-hand tracking to 97.3% and 77.1%, respectively. The most critical degradation occurred under backlighting, where accuracy dropped to 88.4% for the body pose, 65.2% for the left hand, and 43.3% for the right hand, underscoring the necessity of a controlled ambient lighting environment to sustain high-accuracy vision-based teleoperation. The retargeting reproducibility of the mapping pipeline was validated across multiple evaluation trials to quantify the consistency of the generated joint commands against intended reference postures. The shoulder-roll channels demonstrated highly stable tracking, yielding RMSE of 5.3°–5.6° while maintaining a low SD of no more than 4.64° for both arms, which verifies that the handedness-dependent sign and clamp equations successfully prevent kinematic asymmetry. However, the RMSE values of the elbow channel rose to 8.7°–8.8°, and the right elbow demonstrated the highest variability with an SD of 6.59°. This variance indicates the limitation of the monocular vision setup, where estimating elbow flexion from sequential shoulder, elbow, and wrist landmarks introduces geometric sensitivity to depth ambiguities when movements occur directly along the camera viewing axis. The jitter suppression stage validated the attenuation capabilities of the first-order EMA filter against high-frequency tracking noise. At a conservative filter gain of 0.25, the shoulder channels achieved a 40.3% to 63.4% reduction in frame-to-frame jitter, with raw noise variations causing the left shoulder to be 1.2 times noisier than the right. To preserve tracking responsiveness during rapid joint movements, the elbow channels utilised a higher filter gain of 0.60, resulting in a lower jitter reduction of 34.2% to 38.5%. These findings highlight a critical engineering trade-off where minimising signal latency for fast arm articulation restricts absolute noise suppression. The runtime evaluation also shows that it runs at 33.8 FPS, achieving real-time performance.

To validate its operational functionality, the proposed algorithm was successfully deployed on the physical Unitree G1 Edu humanoid platform. The seamless execution of real-time upper-limb and finger mimicry demonstrates that the markerless tracking framework is highly viable for intuitive human-robot interaction and remote teleoperation applications. Since the proposed mimicry pipeline is tested only under a controlled environment for a single subject, future studies should validate the proposed pipeline under dynamic scenarios to understand the performance and reliability of the pipeline under noisy and complex real-time scenario. Multi-subject detection is also a potential research direction to explore the dynamic environment sensing and perception analysis involving two or more interactors. Future work could compare the success of related research in sensor fusions and image analysis development research for further improvement, e.g., low-light environments and human gesture detection. Hence, the developed system enhances human-robot interaction through more intuitive communication, paving the way for greater public acceptance of robotics in healthcare and customer service.

Acknowledgement: The collaboration was supported by the LUT Internationalization Project. The industrial usability and technical supported by LOBRA Project, co-funded by the European Union. This research is supported by Taylor's University, Malaysia, through Taylor's Internal Research Grant Scheme-Impact Lab Grant.

Funding Statement: This research is supported by Taylor's University, Malaysia, through Taylor's Internal Research Grant Scheme-Impact Lab Grant (TIRGS-ILG/1/2024/SOE/001). LOBRA Project, co-funded by the European Union, supported open access publishing.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Jun Meng Woh and Hui Hwang Goh; methodology, Ee Sin Yong and Ari Happonen; software, Yi Wen Tan, Jun Meng Woh and Ee Sin Yong; data curation, investigation, Ee Sin Yong, Yi Wen Tan, Hui Hwang Goh and Kah Yoong Chan; writing—original draft, review and editing, Ee Sin Yong, Jun Meng Woh, Ari Happonen, Hui Hwang Goh, Kah Yoong Chan and Yi Wen Tan; research guidance, Ari Happonen; supervision, Wai Leong Pang; proof reading, Ari Happonen, Hui Hwang Goh and Kah Yoong Chan; corresponding author, Ari Happonen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets supporting this article are not currently available due to ongoing analyses. Requests for access may be directed to the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Islam MT, Sepanloo K, Woo S, Woo SH, Son YJ. A review of the industry 4.0 to 5.0 transition: exploring the intersection, challenges, and opportunities of technology and human-machine collaboration. Machines. 2025;13(4):267. doi:10.3390/machines13040267.
2. Rahman MM, Khatun F, Jahan I, Devnath R, Bhuiyan MA. Cobotics: the evolving roles and prospects of next-generation collaborative robots in industry 5.0. J Robot. 2024;2024(1):2918089. doi:10.1155/2024/2918089.
3. Vianello L, Penco L, Gomes W, You Y, Anzalone SM, Maurice P, et al. Human-humanoid interaction and cooperation: a review. Curr Robot Rep. 2021;2(4):441–54. doi:10.1007/s43154-021-00068-z.
4. Chand R, Sharma B, Kumar SA. Systematic review of mobile robots applications in smart cities with future directions. J Ind Inf Integr. 2025;45(6):100821. doi:10.1016/j.jii.2025.100821.
5. Ahmad Usmani U, Happonen A, Watada J. Revolutionizing transportation: advancements in robot-assisted mobility systems. In: ICT infrastructure and computing. Singapore: Springer; 2023. p. 603–19. doi:10.1007/978-981-99-4932-8_55.

6. Paluch S, Wirtz J, Kunz WH. Service robots and the future of services. In: Marketing weiterdenken. Wiesbaden, Germany: Springer; 2020. p. 423–35. doi:10.1007/978-3-658-31563-4_21.
7. Lampropoulos G. Social robots in education: current trends and future perspectives. Information. 2025;16(1):29. doi:10.3390/info16010029.
8. Rojas-Quintero JA, Rodríguez-Liñán MC. A literature review of sensor heads for humanoid robots. Robot Auton Syst. 2021;143(3):103834. doi:10.1016/j.robot.2021.103834.
9. Moriuchi E, Murdy S. The role of robots in the service industry: factors affecting human-robot interactions. Int J Hosp Manag. 2024;118:103682. doi:10.1016/j.ijhm.2023.103682.
10. Denny J, Elyas M, D'Costa A, Donate R, Souza D. Humanoid robots-past, present and the future. Eur J Adv Eng Technol. 2016;3:8–15.
11. Skubis I, Mesjasz-Lech A, Nowakowska-Grunt J. Humanoid robots in tourism and hospitality—Exploring managerial, ethical, and societal challenges. Appl Sci. 2024;14(24):11823. doi:10.3390/app142411823.
12. Khoa DT, Gip HQ, Guchait P, Wang CY. Competition or collaboration for human-robot relationship: a critical reflection on future cobotics in hospitality. Int J Contemp Hosp Manag. 2023;35(6):2202–15. doi:10.1108/ijchm-04-2022-0434.
13. Intahchomphoo C, Millar J, Gundersen OE, Tschirhart C, Meawasige K, Salemi H. Effects of artificial intelligence and robotics on human labour: A systematic review. Leg Inf Manag. 2024;24(2):109–24. doi:10.1017/s1472669624000264.
14. Massaguer Gómez G. Should we trust social robots? Trust without trustworthiness in human-robot interaction. Philos Technol. 2025;38(1):24. doi:10.1007/s13347-025-00850-3.
15. Watanabe T, Nishida K, Kumazaki H. Initial study of verbal and nonverbal communication training through the collaborative operation of a humanoid robot for individuals with autism spectrum disorder. Asian J Psychiatry. 2025;106(14):104423. doi:10.1016/j.ajp.2025.104423.
16. Bin T, Yan H, Wang N, Nikolić MN, Yao J, Zhang T. A survey on the visual perception of humanoid robot. Biomim Intell Robot. 2025;5(1):100197. doi:10.1016/j.birob.2024.100197.
17. Jessup SA, Alarcon GM, Harris KN, Lee MA. The influence of robot anthropomorphism and trust violation types on trustworthiness perceptions and trust behaviors. Int J Soc Robot. 2025;17(8):1437–52. doi:10.1007/s12369-025-01295-6.
18. Roselli C, Lapomarda L, Datteri E. How culture modulates anthropomorphism in human-robot interaction: A review. Acta Psychol. 2025;255(6):104871. doi:10.1016/j.actpsy.2025.104871.
19. Sarvghadi H, Reinhardt A, Semmelhack EA. A survey of wearable devices to capture human factors for human-robot collaboration. Pervasive Mob Comput. 2025;110(1):102048. doi:10.1016/j.pmcj.2025.102048.
20. Sharif Razavian R. Human-aware control for physically interacting robots. Bioengineering. 2025;12(2):107. doi:10.3390/bioengineering12020107.
21. Del Giudice M, Scuotto V, Ballestra LV, Pironti M. Humanoid robot adoption and labour productivity: a perspective on ambidextrous product innovation routines. Int J Hum Resour Manag. 2022;33(6):1098–124. doi:10.1080/09585192.2021.1897643.
22. Roychoudhury A, Khorshidi S, Agrawal S, Bennewitz M. Perception for humanoid robots. Curr Robot Rep. 2023;4(4):127–40. doi:10.1007/s43154-023-00107-x.
23. Cao L. Humanoid robots and humanoid AI: review, perspectives and directions. ACM Comput Surv. 2025;58(4):1–37. doi:10.1145/3770574.
24. Zhang Q, Zhang Z, Cui W, Sun J, Cao J, Guo Y, et al. HumanoidPano: hybrid spherical panoramic-LiDAR cross-modal perception for humanoid robots. arXiv:2503.09010. 2025.
25. Asfour T, Welke K, Ude A, Azad P, Dillmann R. Perceiving objects and movements to generate actions on a humanoid robot. In: Unifying perspectives in computational and robot vision. Boston, MA, USA: Springer; 2008. p. 41–55. doi:10.1007/978-0-387-75523-6_4.
26. Zhou Y, Wang X. Biomimetic active stereo camera system with variable FOV. Biomimetics. 2024;9(12):740. doi:10.3390/biomimetics9120740.

27. Qu P, Li Z, Jia Y, Liu Z, Zhu L, Li H, et al. OmniDP: beyond-FOV large-workspace humanoid manipulation with omnidirectional 3D perception. arXiv:2603.05355. 2026.
28. Al-Tawil B, Candemir A, Jung M, Al-Hamadi A. Mobile robot navigation with enhanced 2D mapping and multi-sensor fusion. Sensors. 2025;25(8):2408. doi:10.3390/s25082408.
29. Xia J, Li Y, Zhu Y, Wu H, Wang Y. A teleoperation system for dexterous hand control. In: Proceedings of the 2024 4th International Joint Conference on Robotics and Artificial Intelligence. New York, NY, USA: ACM; 2025. p. 216–21. doi:10.1145/3696474.3708465.
30. Fu J, Poletti M, Liu Q, Iovene E, Su H, Ferrigno G, et al. Teleoperation control of an underactuated bionic hand: comparison between wearable and vision-tracking-based methods. Robotics. 2022;11(3):61. doi:10.3390/robotics11030061.
31. Fang B, Sun F, Liu H, Guo D, Chen W, Yao G. Robotic teleoperation systems using a wearable multimodal fusion device. Int J Adv Rob Syst. 2017;14(4):1729881417717057. doi:10.1177/1729881417717057.
32. Durrani HA, Khan S. Real-time whole-body teleoperation of a humanoid robot using IMU-based motion capture with Sim2Sim and Sim2Real validation. arXiv:2605.12347. 2026.
33. Featherson DG. Real time markerless 3D hand tracking for intuitive robotic arm control [master's thesis]. San Luis Obispo, CA, USA: California Polytechnic State University; 2025.
34. Qin Y, Yang W, Huang B, Wyk K, Su H, Wang X, et al. AnyTeleop: a general vision-based dexterous robot arm-hand teleoperation system, Daegu, Republic of Korea. In: Proceedings of the Robotics: Science and Systems 2023; 2023 Jul 10–14. Daegu, Republic of Korea. doi:10.15607/rss.2023.xix.015.
35. Gao Q, Deng Z, Ju Z, Zhang T. Dual-hand motion capture by using biological inspiration for bionic bimanual robot teleoperation. Cyborg Bionic Syst. 2023;4(9):0052. doi:10.34133/cbsystems.0052.
36. Aslan E, Ali Arserim M, Uçar A. Development of push-recovery control system for humanoid robots using deep reinforcement learning. Ain Shams Eng J. 2023;14(10):102167. doi:10.1016/j.asej.2023.102167.