PepperPose: Full-Body Pose Estimation with a Companion Robot

Chongyang Wang
wangchongyang@tsinghua.edu.cn
Tsinghua University
Beijing, China

Siqi Zheng
zheng-sq21@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Lingxiao Zhong
zhonglx20@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Chun Yu
chunyu@tsinghua.edu.cn
Tsinghua University
Beijing, China

Chen Liang
lliangchenc@163.com
Tsinghua University
Beijing, China

Yuntao Wang
yuntaowang@tsinghua.edu.cn
Tsinghua University
Beijing, China

Yuan Gao*
gaoyuan@cuhk.edu.cn
Shenzhen Institute of Artificial Intelligence and Robotics for Society, The Chinese University of Hong Kong, Shenzhen
Shenzhen, China

Tin Lun Lam
tllam@cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China

Yuanchun Shi
shiyc@tsinghua.edu.cn
Tsinghua University
Beijing, China
Qinghai University
Xining, Qinghai, China

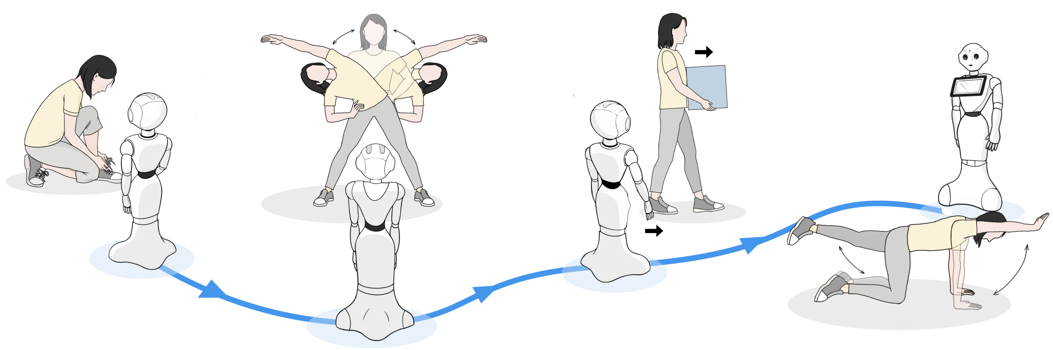

Figure 1: PepperPose is a companion robot system that optimized to estimate the pose of a user when they move and act diversely in an open space. The magic lies in its ability of actively tracking a person and finding the optimal viewpoint for pose estimation. With PepperPose, the user does not need to wear any devices for accurate action sensing results, and such a capacity opens up new opportunities in embodied interaction and intelligence.

ABSTRACT

Accurate full-body pose estimation across diverse actions in a user-friendly and location-agnostic manner paves the way for interactive applications in realms like sports, fitness, and healthcare. This task becomes challenging in real-world scenarios due to factors like the user's dynamic positioning, the diversity of actions, and the varying acceptability of the pose-capturing system. In this context, we present PepperPose, a novel companion robot system tailored for optimized pose estimation. Unlike traditional methods, PepperPose actively tracks the user and refines its viewpoint, facilitating enhanced pose accuracy across different locations and actions. This allows users to enjoy a seamless action-sensing experience. Our evaluation, involving 30 participants undertaking daily functioning and exercise actions in a home-like space, underscores the robot's promising capabilities. Moreover, we demonstrate the opportunities that PepperPose presents for human-robot interaction, its current limitations, and future developments.

*Corresponding Author.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3642231

CCS CONCEPTS

• **Human-centered computing → Interaction devices**; • **Computing methodologies → Motion capture**; **Robotic planning**.

KEYWORDS

pose estimation, human-robot interaction

ACM Reference Format:
Chongyang Wang, Siqi Zheng, Lingxiao Zhong, Chun Yu, Chen Liang, Yuntao Wang, Yuan Gao, Tin Lun Lam, and Yuanchun Shi. 2024. PepperPose: Full-Body Pose Estimation with a Companion Robot. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642231

1 INTRODUCTION

Many renowned research works [39, 50, 66] have underscored the significance of accurate full-body pose estimation, particularly in contexts where actions involving multiple body parts become the essential channel for information exchange. This is especially applicable in fields such as athlete training [47], exercise coaching [41], and sports rehabilitation [11, 58]. In these situations, the ability to extract detailed kinematic features from a full-body pose is critical for the effective operation of these interactive systems. However, implementing a pose-capturing system in an open and real-world environment poses a considerable challenge. This is largely due to the unpredictability of the target's movements across various spatial locations and the diversity of their actions. Furthermore, it is crucial to take into account the acceptability of naive users, particularly when they are required to wear devices or stay within a specific area to enjoy the service.

To reach a balance between user comfort and pose estimation accuracy, we seek a versatile, flexible, and interactive co-pilot that can actively perceive the skeletal poses of the user when they move and act in an open area. Given the recent advancement in robotics, employing a visual robot for this purpose emerges as a promising solution. Nonetheless, this poses unique challenges and questions in driving the robot with its visual system. In this explorative work, we target one central question: **how to enable a visual robot to adaptively adjust its position and viewpoint for optimal pose estimation across different spatial positions and action types?** This is critical for vision-based systems, as the occlusion caused by a fixed viewing angle and diverse facing directions of the user can significantly reduce the accuracy.

Addressing these issues, this paper presents PepperPose, a pose estimation-centric robotic system integrated with the humanoid Pepper robot [6]. We trained the robot to actively track the target user when they move, and adjust the viewpoint to improve pose estimation results. Consequently, PepperPose can function as a fundamental action-sensing platform that eliminates the need for users to wear additional devices or remain within a restricted area. We evaluated the performance of this system in a real-world experiment that involves 30 participants. Particularly, we quantify its pose estimation accuracy by leveraging the synchronized high-fidelity pose obtained from the participant's full-body motion capture suit integrating Inertial Measurement Units (IMUs), its track losing rate, and speed in moving onto the optimal observation position in response to various participant actions. While the current cost of such a robot may be unaffordable, we highlight the potential of a robotic pose estimation solution that could provide richer interaction opportunities with minimal impact on user experiences. By working closely with the industry and public interest groups (e.g., hospital, gyms, and sports teams), we anticipate that early applications of PepperPose are on the horizon.

2 RELATED WORK

Human pose estimation has been a subject of extensive research over the past decade, with solutions utilizing a variety of devices (e.g., camera of different capacities, stand-alone IMUs, mobile phone, bracelet, smartwatch, earbud, Wi-Fi, and mm Wave etc.) deployed for different purposes and under diverse conditions. Here, we first review the pose estimation studies in two categories: those that adopt external devices, and those that apply wearable sensors. We further review relevant advances in active perception using robots. Through this literature review, we identify a gap in the research: limited efforts have been made to liberate the user from the need to wear equipment while also providing accurate pose estimations as they move across different locations and act diversely. This forms the basis of our motivation and the direction of our research.

2.1 Estimate Pose with External Devices

2.1.1 With Stationary Devices. Studies focusing on pose estimation using vision-oriented systems typically share a common characteristic: they aim for, and often require, the captured pose to be accurate. Such precision could better drive their downstream applications in fields such as rehabilitation [58], VR [68], digital human [31], and so on. The devices employed in these studies range from monocular RGB [13, 20, 43, 72] and RGB-D (Kinect [1], Intel RealSense [7]) cameras to professional ones like OptiTrack [5] and Vicon [9]. Although these solutions offer high-accuracy pose estimation, the use of vision-captured systems is significantly restricted by their stationary positioning and coverage, which can limit their effectiveness and adaptability in open environments. We believe that a user-friendly, mobile platform equipped with a camera could provide a promising solution to balance pose estimation quality with mobility. Additionally, there is an emerging trend of utilizing wireless sensing devices (e.g., Wi-Fi [22, 49, 71] and mm Wave [40, 52]) for full-body pose estimation. However, these studies are still in their exploratory stages, and the pose estimation provided by their systems tends to be less accurate.

2.1.2 With Dynamic Drones. There are several studies that employ drones (also referred to as aerial robots) to capture the human action [24, 28, 32]. An earlier work by Zhou et al., [76] takes the advantage of using a drone to actively record the action of a user in the wild and developed algorithms to reconstruct 3D body pose data from the video. The work by Cheng et al. [16] reconstructs the 3D mesh of human body using an aerial robot mounted with a depth camera. Given a participant in a static posture (standing still and punching posture in their experiments), the robot was able to find the shortest flying route surrounding the person to capture the 3D body mesh. This proved faster than its previous work, FlyCap [65], which adopts a fixed flying trajectory during the capture. In more dynamic settings, such as when the user is walking or running, Tallamraju et al. [56] presented a model that is able to put the target person within the center of the captured frame when they walk in diverse directions. They further analyzed the impact of moving speed and distance between the drone and the user on pose estimation accuracy. Boonsongsrikul et al. [12] conducted a more relevant study, where the drone follows the user closely as they walk freely and capture the full-body pose simultaneously. While these works support the idea of using a mobile device mounted with a camera for less constrained action capturing, they primarily focus on simple situations where the target is static or merely moving around. Additionally, the use of drones in domestic settings raises concerns in safety and comfort of the user. The noise produced by these devices has been criticized by researchers working on human-robot interaction [53, 61, 63]. By contrast, our study utilizes PepperPose, which operates on the ground level and captures poses of the participant while they move and perform various actions. In the future, its voice interface and robot arms have the potential to provide richer interaction with the user, in comparison to drones.

2.2 Estimate Pose with Wearable Sensors

When the application scenario of the pose-tracking system extends beyond a pre-defined area, acquiring accurate full-body pose becomes rather challenging. In such cases, the most practical solution often involves the use of wearable systems equipped with inertial sensors. Notably, commercial products from companies like Movella Xsens [2], Noitom [3], and Rokoko [8] offer solutions in the form of suits embedded with numerous IMUs (typically 17). These suits, usually wireless in today's market, offer the user increased freedom in terms of mobility and orientation. Nevertheless, these inertial sensor systems are not without their drawbacks. Long-standing issues with pose-shifting, where errors accumulate over time, persist. Additionally, the practicality of wearing such a suit for everyday use is questionable due to potential discomfort and inconvenience. For the latter issue, recent efforts have been directed towards using less IMUs for pose estimation. By leveraging SMPL [42], a physics-restraint human body model, significant advancements have been made in using 4 to 17 IMUs to approach full-body pose reconstructions [25, 57, 69]. Mollyn et al. [44] proposed a system that utilizes IMUs present in commodity devices (mobile phones, smartwatches, and earbuds) for full-body pose estimation. This approach significantly improves user comfort since it eliminates the need for specialized sensors. However, pose estimation with fewer sensors usually results in less accuracy. Specifically, the system has to infer the movement of body parts to which sensors are not attached. Another promising approach involves the use of soft fabric-based devices [75]. This method has shown positive strides in pose reconstruction, although it is still in its developmental phase within the lab setting. A recent research has enabled the use of a VR headset for ego-body pose estimation [37], envisioning the scenario of user's wearing the device and moving around in a domestic environment.

2.3 The Active Perception of a Robot

We find the following studies in the area of active perception-oriented robot control that are informative to our work. The capacity of active perception is the basis for a visual robot to understand the physical environment [36]. Therein, its perception may include object recognition for manipulation [19] and scene recognition for navigation [26]. More recently, researchers start to introduce the human action data for the robot to imitate the human behaviors. Zimmermann et al. [77] trained action models for the robot given body pose data, as such the robot learns to interact with objects in a way similar to the human demonstrator. Weigend et al. [62] mapped the arm gesture into the trajectory that a robot arm could follow. To the best of our knowledge, there is no study done on active full-body pose estimation across locations and diverse human actions using a visual robot placed on the ground. In this work, we transform a popular visual humanoid robot, pepper, into an active pose estimation machine, which automatically track and adjust its viewpoint for optimal estimation results. Such a robot can also observe and imitate human behaviors on their own, without the effort to provide them with human actions using extra devices.

3 VIEWPOINTS IN VISUAL POSE ESTIMATION

For vision-captured system, including the visual robot system like PepperPose, a major issue that affects their pose estimation qualities is the occlusion caused by external objects, the facing direction and postures of the user, and proportion of their body in the frame, especially when the camera is put at a fixed **viewpoint**. To provide a clear picture of this problem and construct prior knowledge to aid the functioning of PepperPose, we conducted a preliminary experiment with a user conducting everyday functioning and exercise actions. These actions also compose our following main evaluation. We first fixed the distance between the camera and the user, resulting in an approximate 80% vertical proportion of the standing-up human body in the captured frame, to analyze the impact of different observational angles, and then adopted the *optimal* angle learned therein to drive the analysis on the distance.

3.1 View Angles

We captured the footage from 8 view angles split by 45° degrees surrounding the participant (158 cm tall) at a distance of 1.5 meters per each complete action execution. The camera of a mobile phone that captures videos in 1080P@60Hz was used. By default, the degree is set as 0 for the facing direction of the participant, which increases along the counterclockwise circle. To calculate the pose estimation quality, we adopted confidence scores output by the 2D pose estimator of PoseFormerV2 [73], an open-sourced tool for 3D full-body pose estimation. Specifically, the metric is calculated as the average confidence scores of all the body joints for the action's duration per each view angle. It is notable that this method is a representative of 3D full-body pose estimation with monocular RGB cameras [38, 60, 70, 74], where their inputs are 2D pose sequences that are either provided by the dataset or acquired using off-the-shelf estimators such as HRNet [55], CPN [15], and stacked hourglass network [45].Therefore, results presented in this evaluation should be informative about the viewpoint issues existing in current vision-captured system, and to our design on using a robot for this task.

Figure 2 illustrates the results. It is clear that there is a range of view angles that could return pose estimation results with better qualities, and such a range varies per action type. Particularly, we can see that the quality is mostly affected by orientations of the body as well as the occlusions caused by movements of limbs. Thereon, we categorize these 14 actions used in this study into three groups given their similarities in postures, where each group

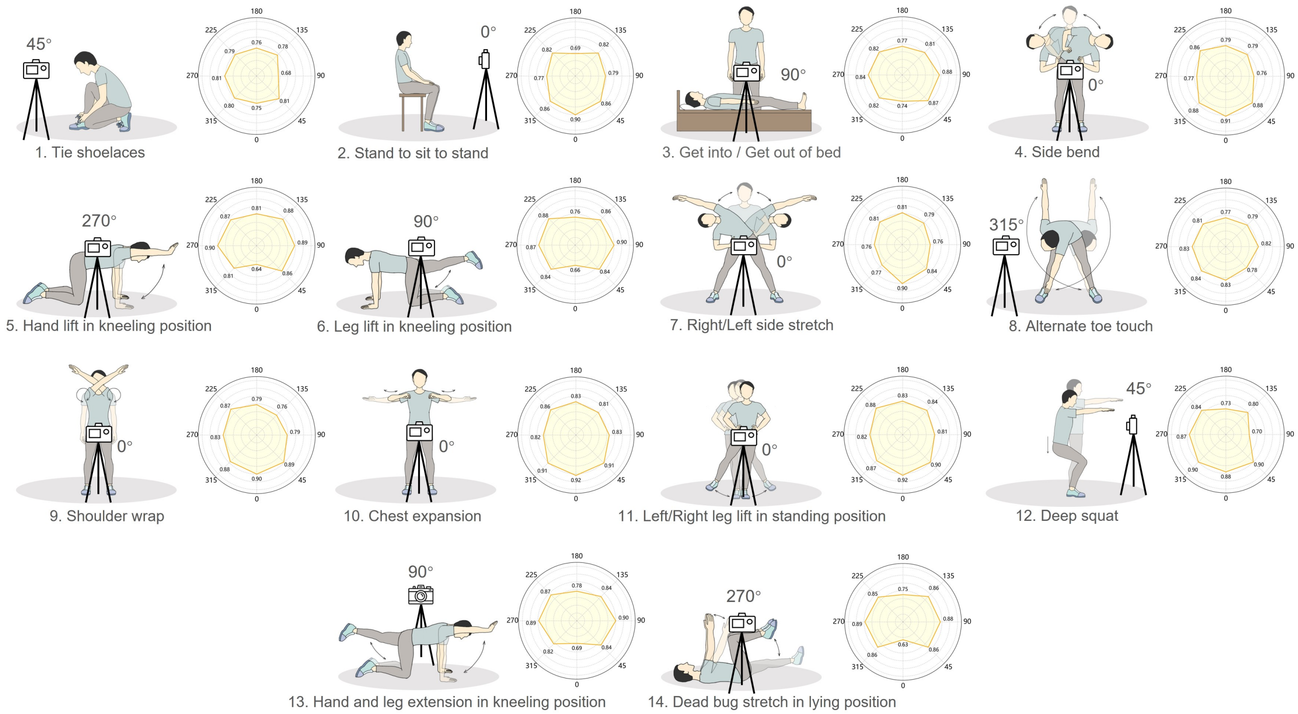

Figure 2: The results of analyzing the impact of viewing angles on pose estimation qualities, together with the illustration of actions that are considered in this work. Whilst there is a range of angles per action type that provide pose estimation results with high qualities, the view angle that returns the *best* result is highlighted by placing the camera icon at the respective observational position.

has a similar range of suitable view angles. We have *standing* (with actions of number 2, 4, 7, 9, 10, 11), *bending* (with actions of number 1, 8, and 12), and *reclining* (with actions of number 3, 5, 6, 13, and 14). We believe these three categories reasoned here shall apply to actions that are not covered by this study, making room for the generalization capacity of PepperPose to unseen actions in real life. The suitable view angle is 0° (facing the person) for standing, 45° or 315° for bending, and 90° or 270° for reclining. This will serve as important prior knowledge to support the functioning of PepperPose.

3.2 View Distances

The distance between the user and the robot (camera) affects the pose estimation quality [56], as it leads to different proportions of the target in captured frames. Through another analysis, we build such a prior knowledge to aid the control of PepperPose in terms of the distance it should keep away from the user. We directly use the *optimal* view angle acquired in the last subsection per each action to conduct the analysis on distance. The same mobile phone is used as the camera. By controlling the vertical proportion of the standing participant (180 cm tall) in the captured image, we adopt the same quality metric used above to show the impact of different distances that result in proportions ranging from 90% (camera put close to the subject, approximately 1 meter in our experiment) to 20% (camera put far from the subject, approximately 5 meters in our experiment).

Figure 3 reports the results. Across all action types, we observed that a distance resulting in the target's vertical proportions occupying 50% to 80% of the captured frame is optimal. This range appears to offer promising pose estimation quality. Given the diverse heights of our participants, we follow this range to adopt a distance range of 1.5 meters to 2.5 meters for PepperPose's operation in this work. This also meets our safety requirements, ensuring that the robot does not interfere with the user during intense exercise or movement in the space.

It's worth noting that this preliminary analysis on view angles and distances could get slightly affected by the performance of this participant, e.g., for the action of leg lift in kneeling position, the pose estimator may find the view angle of 270° to be better than 90° when their left leg lifts lower than the right one. Additionally, for a proper estimation upon a skinny person, the view distance could be shorter than what for a person with a bigger size. Nevertheless, it is natural, since the individual bias is another challenging factor to vision-captured pose estimation methods. Moreover, the distance is also determined by the Field of View (FOV) of the camera, i.e., a camera with a narrower view range needs a longer distance to put the subject in its capture frame. In addition, in our evaluation below, we also consider mobile actions, e.g., carrying objects (e.g., a cardboard box with loads) in walking and sweeping the floor, which are left out in this analysis. This is mainly because, for these

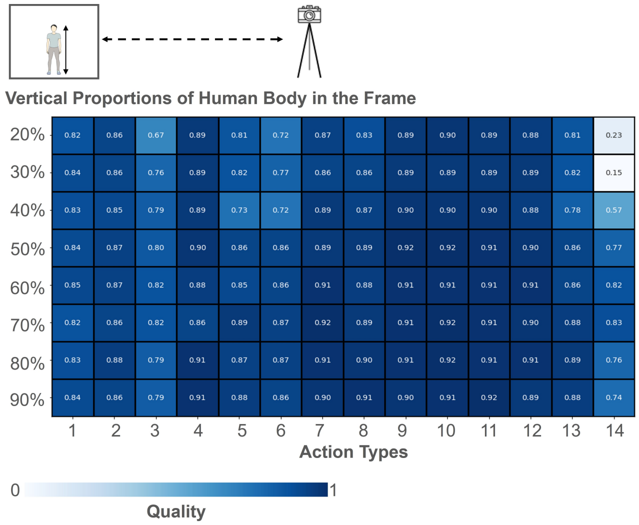

Figure 3: The results of analyzing the impact of distances between the camera and the user. We test the distances that result in different vertical proportions of a standing-up human in the capture frame, ranging from 90% to 20%. The *optimal* view angle for each action is used. The action types are denoted by numbers, as shown in Figure 2.

actions, a proper *capture* is achieved by *tracking* the user at the front -right or -left position.

4 IMPLEMENTATIONS

A digital twin of the Pepper robot is first trained in a simulated environment using online reinforcement learning, where it uses the active visual perception capacity to interact with animated people driven by the action data collected from a real-world experiment.

4.1 Data Collection for Robot Training

We used the simulation environment of Nvidia's Omniverse [4] to conduct the training of PepperPose. In this training, we aim to refine the action space (i.e., all the possible actions) of the robot, and help establish its kinematics model that controls the robot's linear and angular velocities and orientations of its body and head. This learning-based method aims to enable the robot to operate in a natural and smooth manner, and is more efficient and effective than directly manipulating APIs. To drive the virtual people' action in this environment, we used data collected from 100 diverse participants in a home-like environment using a motion capture suit mounted with 17 IMUs from Noitom [3]. Actions we collected are the ones shown in Figure 2, while the action of carrying a suitcase or a cushion during walking is additionally added. We designed a natural and continuous experimental procedure, where the participant conducted each action on their own, with a basic instruction shown on the TV informing the type of action and number of repetitions should be conducted. A complete experimental session lasts for approximately 15 to 20 minutes. Figure 4 demonstrates the data-collection environment, the suit we made to improve user's comfort, the collected 3D pose data, and the interaction between Pepper and

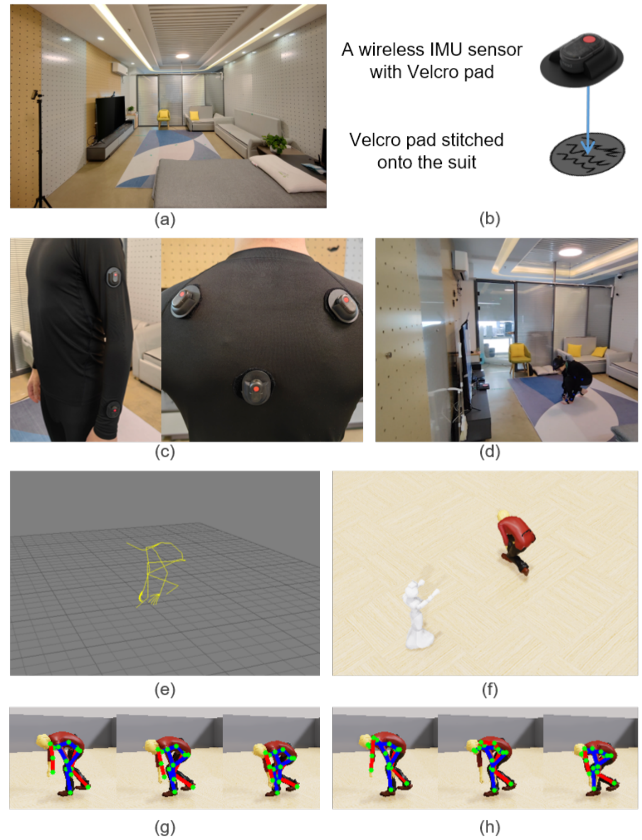

Figure 4: To collect realistic data for the training of PepperPose, (a) we made our best to transform the lab space into a home-like environment and (b-c) self-make comfortable suits with Velcro pads to accommodate the IMUs. (d-f) PepperPose is trained to interact with the *people* in Omniverse driven by such full-body 3D pose data. (g-h) By referring to the ground truth pose, the pose estimation module of PepperPose functions well on the simulated people.

people in Omniverse. As shown, the pose estimation module of PepperPose functions well on capturing the intermediate 2D pose of the virtual people, suggesting that Omniverse is a perfect place for the training.

4.2 The PepperPose Framework

Figure 5 presents an overview of the PepperPose framework. By using its visual system (i.e., the integrated monocular RGB camera of Pepper that captures video at 360P@10fps), the robot is trained to actively track the user and refine its viewpoint for better pose estimations. Specifically, the functioning of PepperPose relies on three modules. First, operating on the captured video frames, the pose estimation module extracts the 3D full-body poses, with 2D poses as the intermediates. Second, given the poses, the pose classification module classifies the action of the user into one of the three groups (i.e., standing, bending, and reclining) that characterize the coarse postures of actions considered in this study, as are discussed in

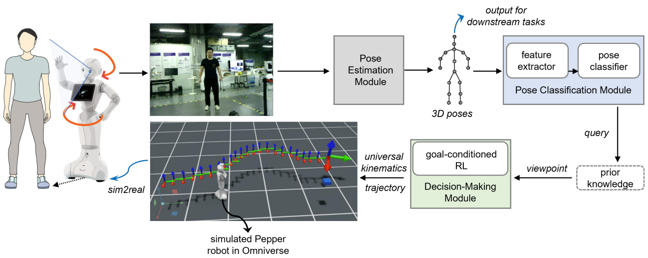

Figure 5: The framework of PepperPose.

Section 3.1. Thereon, the robot is able to retrieve the knowledge of the range of viewing angles that can lead to better pose estimation results. With such knowledge, the decision-making module plans the route to move to that position. It is notable that, by operating on poses, PepperPose demonstrates strong generalization capabilities from the training in Omniverse to real-world environments.

4.2.1 Pose Estimation Module. We use the off-the-shelf 3D pose estimator, namely PoseFormerV2 [73] to acquire 3D full-body pose estimations from the captured frame. To prepare the input sequence for this method in real time, we duplicate each captured frame by the robot to acquire frame-wise pose estimation results. The estimated poses are the output of PepperPose for its initial functioning, which are then the input to the following modules that drive the robot to actively find the suitable viewpoint. Here, we would like to note that this pose estimation module is actually run on the graphic processing unit (GPU) of an additional machine, given the high computational loads it creates that overwhelm the current hardware capacity of Pepper. We made some engineering efforts to reduce the latency caused by transmission control protocol (TCP) communication (i.e., resulting in an interval of approximately 300 ms between sending a single frame to the machine and receiving the estimated pose). We believe this temporary limitation is trivial, given the fast development of compute in mobile platforms as well as cloud computing.

4.2.2 Pose Classification Module. Given the three categories of actions (i.e., standing, bending, and reclining), when the robot recognizes one of them, it would be able to locate the view angle quickly. In addition, another advantage of downsampling the 14 actions into these three groups during the process of viewpoint searching is that it facilitates a better fit with pose data retrieved from *suboptimal* viewpoints, since a simpler classification task on the three categories tends to be more compatible with such less accurate pose data. We conducted an offline training for our pose classification module with the data collected in the home-like environment. It should be noted that the data used for this offline training does not overlap with the data used in the online training of PepperPose in Omniverse. Since the three categories of actions are built based on their characteristics of postures, from standing to reclining, we look into the angle between each part of body and the ground plane to represent such a posture shift. That is, we compute the angles between the vectors formed by every two adjacent nodes in the 3D human skeleton and their projections onto the ground plane (i.e., setting the Y component of the vectors to zero). The use of angular features computed per frame within approximately 4ms benefits the real-time operation of a robot, and eliminates the need for data normalization. We employ a simple yet efficient support vector machine (SVM) model using a radial basis function (RBF) kernel as the classifier, which produces a promising accuracy with a macro F1 score of 0.9489 on the test set during training. In general, the pose classification module receives skeletal data provided by the pose estimation module, computes the angular features, and then returns the classification result. The result is usually acquired by applying majority-voting among multiple frames when necessary.

4.2.3 Decision-Making Module. With the classified action category as a query to retrieve the pre-defined knowledge on viewpoints, via goal-conditioned reinforcement learning [14], the decision-making module locates the viewpoint for better pose estimations and plans the shortest moving path. As is shown in Figure 6, this module is based on a simple real-world 2D coordinate system $< x, y >$ embedded in the navigation system of Pepper robot. Specifically, during this decision-making process, PepperPose needs to leverage and/or compute the following information:

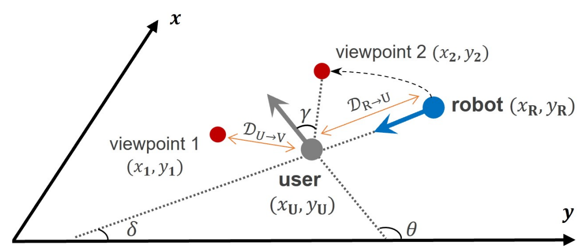

Figure 6: The projection figure of our decision-making process in searching for the viewpoint. For the current action of this user, two suitable viewpoints are acquired according to the prior knowledge, which are defined by the angle γ against the facing direction of the user (clockwise and counterclockwise) and the distance $\mathcal{D}_{U\to V}$ in-between. Pepperpose establishes a simple real-world $< x, y >$ coordinate system to build the movement space, where it estimates its distance $\mathcal{D}_{R\to U}$ from the user and the orientation θ of the user against the x+ direction. Thereon, it plans the route towards the viewpoint.

- **The prior knowledge on viewpoints**, including the view angle γ against the facing direction of the user given the current category of the user's action, and the safe distance $\mathcal{D}_{U\to V}$ the robot should keep from the user, which are already set given our analysis discussed in Section 3.
- **Positioning of the robot**, including its coordinate (x_R, y_R) and orientation δ created by the facing direction of the robot and the x+ direction of the real-world coordinate system, which are provided by the robot's navigation system directly. Specially, the robot is learned to put the user at the center of its captured frame, which ensures that it always face the user.
- **Positioning of the user**, including their coordinate (x_U, y_U) and orientation θ created by the facing direction of the user and the x+ direction of the real-world coordinate system. To compute these, Pepperpose first estimates its distance $\mathcal{D}_{R\to U}$ from the user given

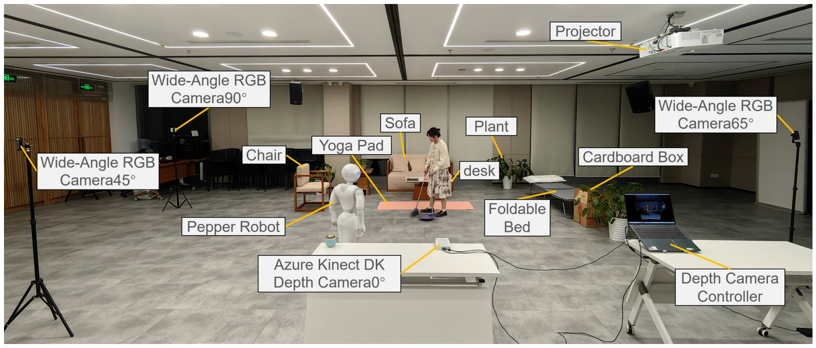

Figure 7: The layout of our real-world experimental space. We put the essential equipment for actions apart from each other, as to help the participant to move and face diversely during the experiment, challenging the functioning of pose estimation systems. The RGB and depth cameras are added as baseline methods for comparison.

the average proportion in the frame of lengths of several action-invariant bones, e.g., the distance between the joints of *head* and *neck*. Then, the orientation of the user is computed by mapping the Z+ direction of the captured pose in the camera coordinate system onto the real-world $< x, y >$ coordinate system. The Z+ direction of the user is simply computed as the norm of the plane formed by two vectors, namely $\mathbf{J}_{head} \rightarrow \mathbf{J}_{right\_shoulder}$ and $\mathbf{J}_{right\_shoulder} \rightarrow \mathbf{J}_{left\_shoulder}$, where $\mathbf{J}$ denotes the joint coordinate.

- **Inferring the viewpoint**, given the above information, the coordinates $(x_n, y_n), n \in [1, 2, ..., N]$ of the N doable viewpoints can be computed using geometric positioning methods.

With the computation of positioning information of the user and potential viewpoints given their current action, the decision-making module further generates universal kinematics for the robot to navigate to the target position. Particularly, when tracking is lost, PepperPose will rotate in place to recapture the user. Please refer to Appendix for detailed information regarding the reinforcement learning process implemented in Omniverse, as well as the simulation-to-real (sim2real) transformation of the system to the real-world scenario.

5 EVALUATION

We conduct an experiment with 30 participants (20 female, 10 male) aged from 19 to 30 (M: 24.1, SD: 2.65) using standard benchmark metrics for measuring pose estimation accuracies, and collect their self-reported user experiences towards the use of PepperPose as a companion robot in real life. Our participants have an average height of 168.34 cm (SD=9.28 cm), and an average BMI of 22.08 (SD of 3.14). Before their arrival, 10 of them reported Neutral for their frequency and proficiency in using robots, 12 are infrequent users and have limited knowledge, and only 8 reported more frequent and proficient use experiences with robots. This study is approved by the Institutional Review Board (IRB) of the University.

5.1 The Design of a Real-World Experiment

As an embodied system, we look into the interaction of PepperPose with the physical environment and real users. Here, we present a real-world experiment that simulates the situation of using PepperPose to acquire the 3D body pose of a user when they perform diverse actions and change their locations and facing directions in a 4m×6m home-like space.

5.1.1 Devices and Equipment. We adopt the commercial robot Pepper [6], with its internal flat 2D RGB camera operating in 320P@10fps with FOV of $54.4° \times 44.6°$, the battery lasting for approximately 10 hours, and a height of 120cm. For ground truth body poses, we use 17 wireless wearable IMU sensors from Noitom [3] together with our self-made suits. The software from the manufacturer provides the 3D pose data. To aid the real-world experiment on daily functioning and exercise actions, aside from the necessary furniture (e.g., a desk, chair, sofa, yoga pad, and bed), we additionally added some small objects (e.g., a check board and fruits) on the desk, a broom with dustpan, a cardboard box with 5kg load, and a few plants. A projector with a screen is used to show instructions during the experiment. Figure 7 demonstrates the arrangement of these elements. Throughout the experiment, we intentionally altered the directions of the chair and the yoga pad to test the adaptability of PepperPose.

5.1.2 Experimental Procedure. In this experiment, we ask each participant to perform the actions (as shown in Figure 2), with walking and changing the facing direction added in between. We also introduce tasks such as walking while carrying a cardboard box and sweeping the floor. Each of the actions that is position-dependent is repeated three times. Before the data collection stage, we provide a brief overview of all the actions to help the participants get familiar with them. To collect natural and continuous data for our evaluation, we opt for moderate experimental control rather than detailed instructions. This involves displaying the number of repetitions and the type of remaining actions with a projector and a screen positioned adjacent to the experimental space. Please kindly refer to Appendix for a sample of the slides used for instruction. Participants were also asked to maintain a moving speed slower than 1 meter per second, in line with the best moving speed (approximately 0.5 meter per second) of Pepper robot used in this study. The distance moved between different actions ranges from 2 to 5 meters.

5.1.3 Baseline Methods. We compare the performance of PepperPose against three stationary wide-angle RGB cameras (operating in 720p, 30Hz, FOV of $106^\circ \times 90^\circ$) and an Azure Kinect DK depth camera [1] (operating in WFOV 2×2 Binned mode, 512×512 pixels, 15Hz, FOV of $120^\circ \times 120^\circ$). It should be noted that this experiment does not account for scene changes that typically occur when a user moves across different rooms or locations. Such changes render the use of stationary cameras less effective and not directly comparable to PepperPose. We position the four cameras at the boundaries of the experimental space, at a distance that ensures they capture the entire scene. In the following, we refer to the camera put in front of the user at their initial position as Depth Camera0°, the camera put at the position lateral to the user as Camera90°, and the rest cameras put at the respective angles as Camera45° and Camera65°. For RGB camera, the same pose estimator used by PepperPose, namely PoseFormerV2 [73], is used to acquire the pose data of the user. For depth camera, the official Azure Kinect Body Tracking API [1] is used to acquire the 3D pose data.

5.2 The Questionnaire for User Study

We adapt the Negative Attitudes towards Robots Scale (NARS) [46] to design a 5-likert scale questionnaire, in order to gain a deeper understanding of user perceptions regarding the integration of a robot for action sensing in their everyday environments. The questions we included are as follows, where each question corresponds to a particular dimension of the user's potential attitude towards a robot. The options are: *"Very Disagreed, Disagreed, Neutral, Agreed, Very Agreed."*. To avoid the influence of irrelevant factors, we ask the users to not consider the potential cost of having such a robot for their personal use.

- **Acceptability**: *I feel comfortable to have a robot system like PepperPose to use in real life*;
- **Usefulness**: *I find the functioning of PepperPose useful*;
- **Expectation**: *I would like to see more applications built on PepperPose given my needs*;
- **Trust**: *I would follow the advice from a robot expert like PepperPose if they are made under the guidance of domain professionals* (personal coach, clinical physiotherapist, psychologist, etc.);
- **Preference**: *If needed, I prefer to receive the support from a real person instead of a robot*;
- **Concern**: *I am worried about the negative influence of this kind of robot to our society.*

We further conducted non-structured interviews to gather their extended feedback, providing insights for the next-step development of this embodied interaction research.

5.3 Evaluation Protocol

Through this real-world experiment with users, we aim to assess the effectiveness and efficiency of PepperPose in accurately capturing the human pose. Therefore, in alignment with prior research [13, 20, 43, 44, 72], we use and/or propose the following metrics to evaluate the performance of PepperPose:

- **Mean Per Joint Position Error** (MPJPE, in centimeters, cm): MPJPE measures the error between the data of two human body poses as the mean Euclidean distance between each corresponding pair of joints; given the ground truth pose returned from IMUs, for each input pose, we implemented the following strategies to maintain a fair comparison, i) the exclusion of frames where the robot directly affected the estimation, e.g., occluding the subject from the camera, or the estimator wrongly recognized the robot as the human, ii) the design of a strict frame-wise normalization process, including skeleton matching and normalization, trajectory removal, Z+ normalization, and root (pelvis) alignment; additionally, since the camera of Pepper used in this study operates at 10Hz, we first synchronize the poses from different devices and sample the respective frames from the stationary cameras and wearable system for a proper comparison;
- **Track Losing Rate** (percent, %): We count the ratio of frames where the method does not even detect the existence of a human, a common problem for vision-captured MoCap system;
- **Reaction Speed** (second, s): Particularly for PepperPose, we measure the time spent on moving to the suitable viewpoint after a user starts an action; the moving actions (i.e., sweeping the floor and carrying a cardboard) are left out in this evaluation since the robot is closely tracking the user when they move; we manually compare the confidence of poses to what acquired from the better viewpoints listed in Section 3.1 to determine the time spent on moving to the better viewpoint.

For the first two metrics, we conduct Friedman test and post-hoc Wilcoxon Signed-Rank test with Bonferroni corrections to analyze the statistical significance. For visualizations, we present representative samples collected from our researchers instead of the data collected from real participants during the experiment, complying with our ethical requirements. Please kindly refer to the video figure of using PepperPose in different scenes for a more vivid understanding of its performance.

6 RESULTS

We first compare the performances of PepperPose with that of stationary RGB and depth cameras. Then, we look into the self-reported feedbacks from participants.

6.1 The Comparison of Performances

6.1.1 Accuracy in Pose Estimation. Figure 8 presents the pose estimation accuracies, measured by MPJPE (cm), of PepperPose and stationary cameras against the ground truth pose. The average accuracy computed across the frames per participant of each method differs significantly between each other ($\mathcal{X}^2(4) = 101.28, p = 5.25\times 10^{-21}$). The post-hoc Wilcoxon Signed-Rank test with Bonferroni corrections shows that PepperPose (M=18.58, SD=2.07, p <0.01/5) is significantly better than Camera45° (M=31.38, SD=3.06), Camera65° (M=28.30, SD=3.63), Camera90° (M=36.15, SD=3.08), and Depth Camera0° (M=25.91, SD=3.46). Camera65° and Depth Camera0° are significantly better than Camera45° and Camera90° (p <0.01/5). However, Camera65° and Depth Camera0° do not differ significantly from each other (p =0.023). Figure 9 presents some qualitative visualizations of the poses estimated from one of the researchers in the experimental space with these devices. As also highlighted in earlier sections, we can see that the increase of errors of stationary cameras is mainly caused by the occlusion of body parts (e.g.,

[1]https://microsoft.github.io/Azure-Kinect-Body-Tracking/release/1.1.x/index.html

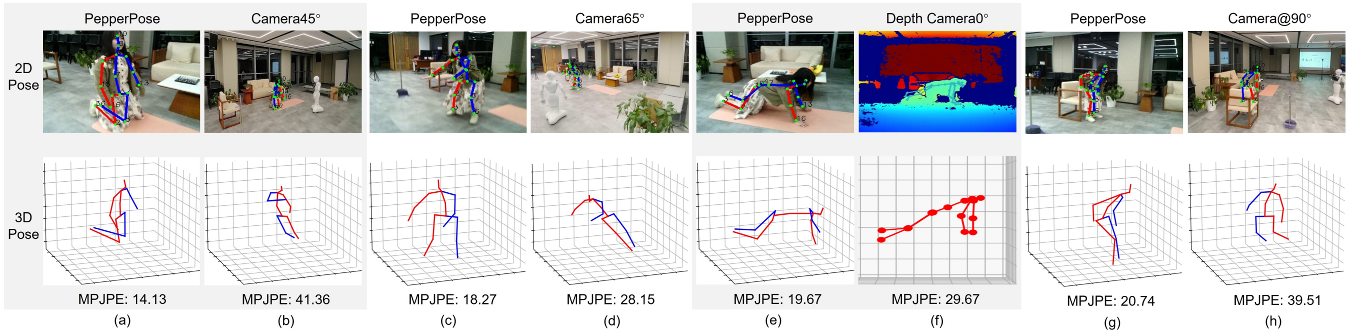

Figure 9: Visualizations of the pose estimation results. Unlike PepperPose (a, c, e, and g) that can actively move and refine its viewpoint, stationary cameras are largely affected by unwanted orientations of the user (b, d, and f) and occlusions that caused by external objects in the environment (h).

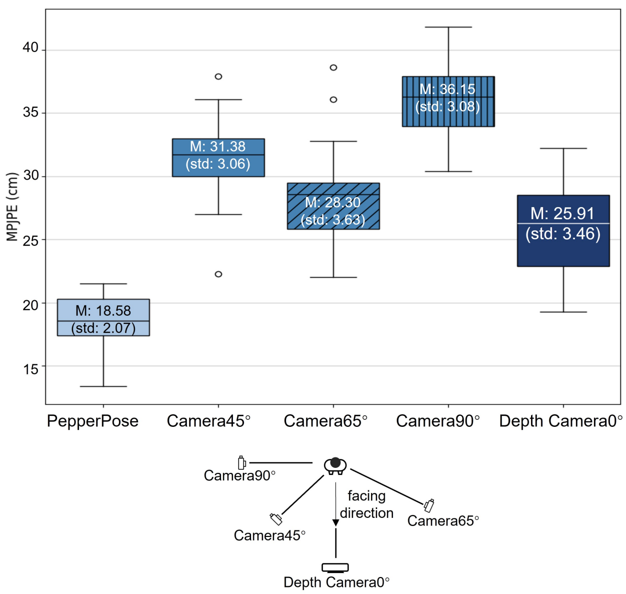

Figure 8: Left: The comparison of pose estimation performances using PepperPose, stationary RGB and depth cameras. Right: The position of RGB and depth cameras. The mean (with standard deviation) is added to each box.

caused by undesired orientation of the user) and external objects (a common situation when the user is not put in an empty space). Generally, this result demonstrates the great potential of using PepperPose for active pose estimation, which provides the user with more freedom on acting and moving in an open space.

6.1.2 Track Losing Rate. Figure 10 (a) reports the track losing rates of PepperPose and the cameras. While these methods significantly differ from each other ($\chi^2(4) = 30.22, p = 4.42 \times 10^{-6}$), the significance is only found between the depth camera (M= 5.87%, SD=3.62%, p <0.01/5) and each of the others. It should be noted that, for most operating time of all the methods, the track losing rate is low, with a ratio of less than 10%. By checking with the captured depth videos,

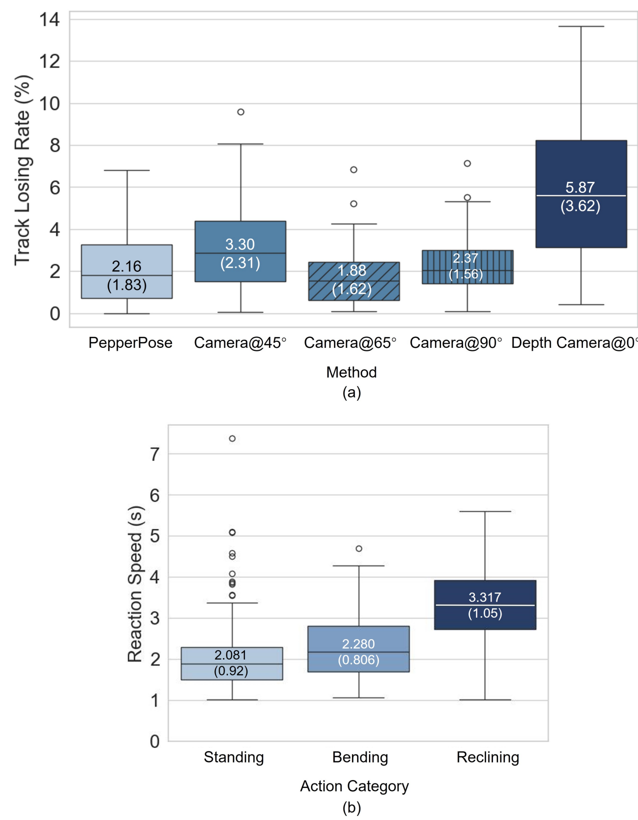

Figure 10: (a) The comparison of track losing rates (%) of PepperPose, stationary RGB and depth cameras; (b) The reaction speed (second, s) of PepperPose in different action categories. The mean (with standard deviation) is added to each box.

we found two major issues that may account for the comparably higher track losing rate of the depth camera: i) given the humanoid design of Pepper robot, the depth sensor tends to wrongly recognize the robot as human more often than the RGB cameras; ii) by switching to the wide-angle mode (e.g., WFOV 2×2 binned mode),

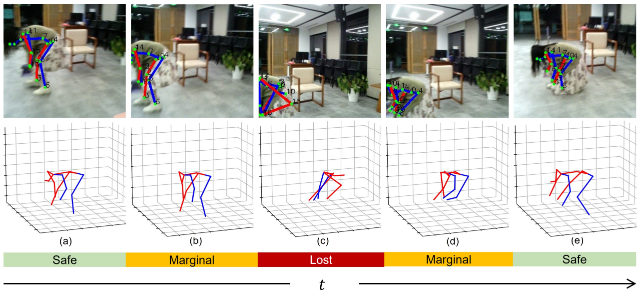

Figure 11: A visualization of PepperPose's captured sequence of a user who suddenly changed from standing to bending, approaching the marginal of captured frames (c and e), and eventually the track is lost (d).

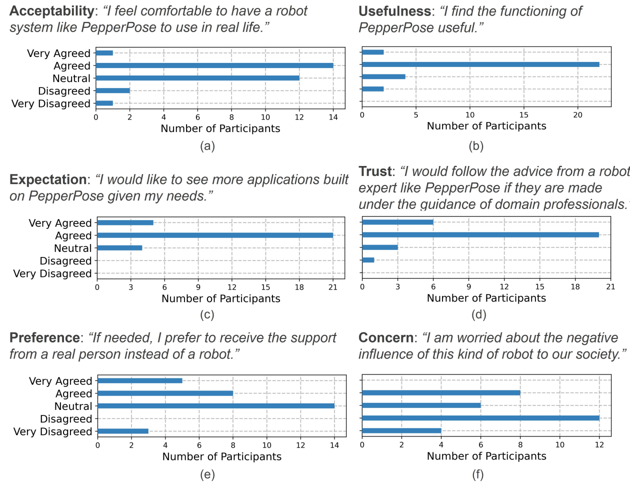

Figure 12: Questionnaire results of the user study, presented per each dimension as defined in Section 5.2

.

the operating distance is reduced from approximately 5 meters to 2 meters, which in our case would cause track losing when the subject is sitting on the sofa when the WFOV mode is used to include the whole experimental space.

Figure 11 provides the visualization of a sequence of tracking results of PepperPose. While the robot is trying to track the user when they move, the sudden change of action categories from standing to bending caused the tracking loss. For stationary cameras, occlusions are still the main reason for causing the lost in tracking.

6.1.3 Reaction Speed. We compute the reaction speed of PepperPose per each category of actions, with results shown in Figure 10 (b). At most, PepperPose is able to find the proper viewpoints after the user's start of an action within 5 seconds in this space. Such a speed is largely affected by the Pepper robot used in this study, which only has a maximum straight line speed of 0.5m/s. By looking at the on-site videos, we also notice the following factors that could impact the reaction speed: i) **The Speed of the User**, when the user moves too fast, our adopted Pepper robot needs a longer interval before getting to the better viewpoint; ii) **Network Traffic**, we used Wi-Fi to transfer the captured frames to an external GPU for pose estimation, which was used to plan the route to better viewpoints; thus, the reaction of PepperPose in this setting could get largely affected when the communication is too busy.

6.2 Insights from the User Study

Here, we first look into the questionnaire feedback from users regarding their real-world experience with a robotic system such as PepperPose, taking into account their prior knowledge and the experience gained during the experiment. Then, we present the comments from some of the participants who provided interesting opinions therein.

6.2.1 Questionnaire. Figure 12 reports results of the questionnaire. It is encouraging to learn that, most participants see values in using PepperPose (24/30=80%), expect to see more downstream applications built on PepperPose (26/30=86.67%). Only 3 participants (3/30=10%) found it less comfortable to have a robot to use in real life. Whilst this positive result could be attributed to the young participants recruited in this study, it is meaningful too since only 8 of them reported to be more frequent and proficient robot users. The educational background seems not to be a clear factor on the acceptability of robots, since all of them are at least undergraduate students. In addition, most participants (26/30=86.67%) expressed that a robot built under professional guidance is a trustworthy source of advices to follow, showing the importance of collaborating with domain experts (e.g., gym coach, clinical physios, etc.) in future development. In the comparison of receiving supports from a human and a robot, the preference of people become more diverse. Participants who are willing to take advices from robot experts also prefer to have supports from a real person. More generally, a few participants (8/30=26.67%) think the use of robots may pose a negative influence on the society.

6.2.2 Interview. To reveal more insights from the questionnaire results, we conducted a non-structured interview with the participants that showed interesting opinions above. We believe the opinions received from these prospective users are rather valuable to guide the development of future research and manufacturing of the industry. We report the interview as follows:

- In terms of the **Acceptability** of using a robot like PepperPose in real life, people mainly talked about the match between the functioning of PepperPose with their needs, as well as the privacy issues typically associated with similar visual systems. Participants reported that: "*It was not that comfortable to be watched by a humanoid robot at the beginning, which became more acceptable when I understood that this robot could act as a physio to improve my health*", "*I felt quite comfortable in this experiment, the robot gave me a sense of care when I was acting in the living room*". These point to the importance of operating the robot to meet necessary needs, rather than a pose estimation platform alone.
- We also collect the **Expectation** towards the future development of PepperPose, which is informative and inspiring. Most of them desire an interactive robot coach for fitness training, and one participant put it even more clear that: "*I would expect this robot can teach me new actions, by displaying the demo on a screen, and*

provide me with instant feedback". Another participant highlights the usefulness of multimodal sensing and contextual service recommendations: "*It would be great if this robot can read my emotions according to not only my actions but also my physiological signals, e.g., by connecting with my smartwatch, and plan my daily routine given the outside weather and traffic status*". Indeed, by using PepperPose, the next step can be swiftly moved onto the design of a fully-interactive embodied agent, and empowering it with multi-modalities.

- For **Trust** and **Preference** of robots vs. human, while most of our participants expressed trust towards the functioning of PepperPose in the future, they tend to have different preferences given their diverse personalities and appreciations of specific roles a robot or human can play in their life. Some of them value the lower cost of using robots for long-term care and monitoring than human, and some highlight the simpler social attribute of robots over human: "*Interacting with robots can waive myself from cognitive loads in dealing with human, since they sidestep awkward moments and unwanted socializing, especially for introverted individuals; in this way, we can focus on functionality*". Whereas, several participants expressed a preference for human interaction, noting: "*When it comes to emotional comfort and hands-on care, the reliable and vivid presence of a human is needed, especially when the reliability of a robot is uncertain*". While the community is working hard to improve the naturalness and sufficiency of robots, we recognize that the decision to utilize robots is highly personal and dependent on the specific context.
- Towards the **concerns** about potential negative social impacts, participants showed two opposite views. Some once again mentioned the security, privacy, and ethical issues, and one said that: "*If this leads to job losses among specialists, we might face situations where professional expertise is unavailable when needed*"; whereas, one participant highlighted that: "*I believe robots lead to reduced cost and more comprehensive sensing capabilities than human in certain tasks, which can reduce unnecessary repetitive works for human*".

7 OPPORTUNITIES WITH PEPPERPOSE

PepperPose could provide exciting interaction experiences to the user in many downstream applications. Here, we provide some potential use cases of PepperPose in its following development.

7.1 A Robot Physio and Coach for Fitness and Rehabilitation

There is a growing interest in providing people with a virtual physio for exercise and rehabilitation guidance at home [27, 33, 35, 59]. However, similar to the findings reported in our earlier sections, existing methods are also limited by their action-sensing capacities. They usually struggle with the granularity of the captured action that could lead to different levels of feedback and guidance to the user, vs. user's comfort in wearing extra devices and/or staying in a constrained area. Furthermore, their form of feedback is mostly limited to visualizations that report the progress of the program and evaluation score of the performed action. Figure 13 demonstrates the skill of our physio partner in understanding the action of a user and responding with professional feedback. We argue that such a

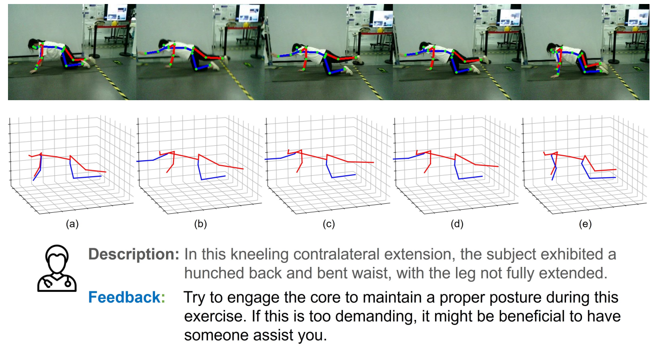

Figure 13: The estimated poses of a user conducting hand and leg extension in kneeling position, together with the understanding and feedback from a physio in natural language.

vivid interaction between the user and the physio can be established using PepperPose soon: First, for action understanding, the latest development on action-language modeling [21, 30] has pointed to the possibility of establishing expert-like description of the action of users, which calls for accurate estimation of full-body poses, matching the basic skill of PepperPose. Second, the large language model can act properly to provide professional feedback by having access to the knowledge, e.g., retrieval-enhanced language modeling [23, 48, 54], while the voice interface of this kind of humanoid robot can work properly to deliver the message.

7.2 Context-Aware Action Sensing for Vulnerable Populations

The objects and scenes a user interacts with provide a full picture of their actions [10, 17]. In comparison to non-visual systems, vision-captured MoCap methods are able to acquire these contextual information for a comprehensive understanding of actions. In this way, a visual MoCap system, e.g., PepperPose, is able to provide richer opportunities for downstream tasks. For instance, once it recognizes the heavy suitcase a person wants to carry, it may remind the person to squat first and lift the bag to avoid hurting the back. Here, we present a proof-of-concept example about the feasibility of transforming PepperPose into such a contextual sensing interface given the recent advances in multimodal language modeling. We use mPLUG-Owl [67], a state-of-the-art framework that empowers a large language model with the multimodal capacity, as a plug-and-play visual parsing model that has the potential to work seamlessly with PepperPose in the future. We directly use the online demo[2] to process our collected video frames. In the online demo, we use the following hyperparameters: 512 for max output tokens, 1 for temperature, 3 for Top K, 0.9 for Top p, 1 for length_penalty, 1 for beam size, 2 for no_repeat_ngram_size, and 28 for number of frames with sampling selected. Figure 14 presents the contextual sensing results. As the participant is conducting the same action (e.g, picking up an object and lying down), the natural language output returned by mPLUG-Owl provides a comprehensive picture of the activity. Although the current capacity of this multimodal

[2]https://www.modelscope.cn/studios/damo/mPLUG-Owl/summary/

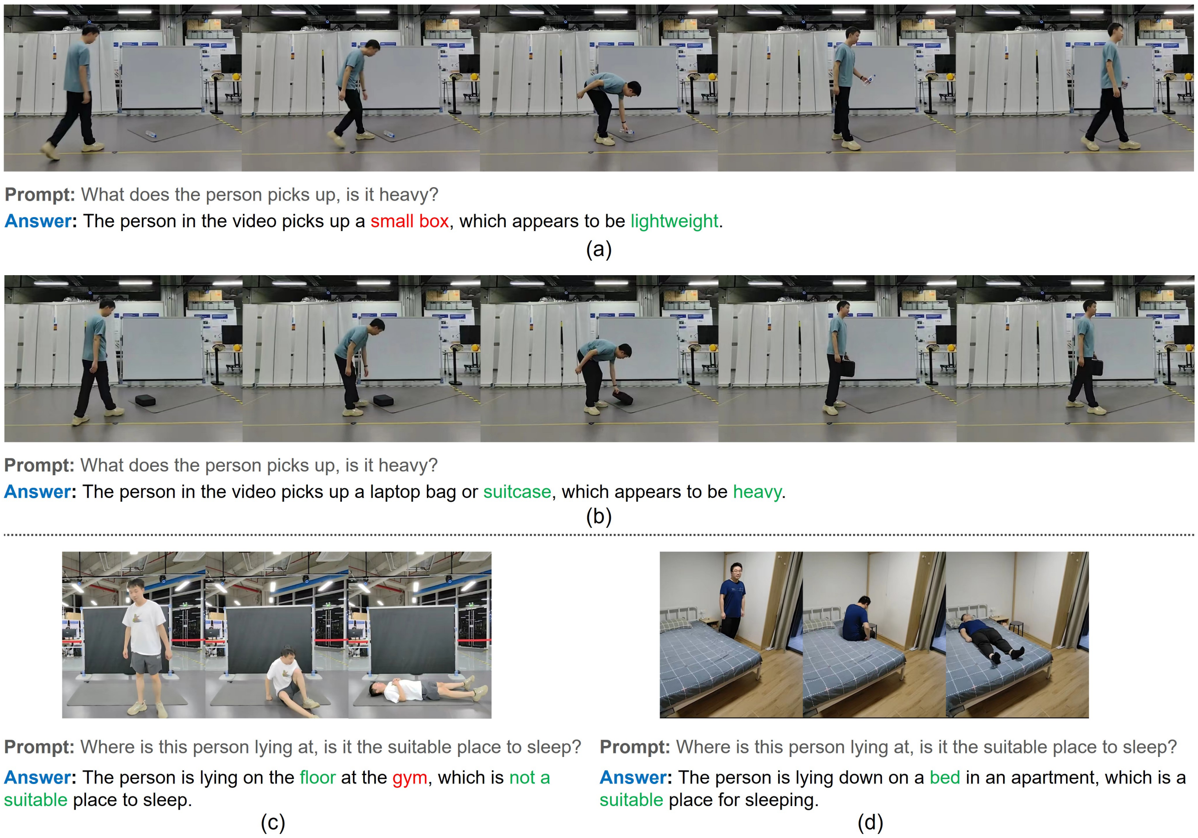

Figure 14: Results of visual parsing with mPLUG-Owl, a multimodal language model. By simply sending pre-defined prompts given the action type, this model helps the system to understand the comprehensive context of the action a person is involved with, which could looks the same by only referring to the pose data.

model is not perfect, e.g., it recognizes the bottle of water as a small box and the lab space as the gym, the general insights about the weight of the item and the judgement on suitability derived from the visual input are correct. Additionally, this online demo only took approximately 2 to 4 seconds before starting the output, which could be easily improved by local deployments or advanced cloud computing. Driven by this capacity, we additionally envision the following two use cases.

7.2.1 A Functional Partner for Children. Humanoid robot is appealing to children, and this new space for interaction with these younger users is also attracting some attention recently. Such a robot can help the children with their homework to increase their motivation and engagement [66]. Alternatively, the robot can simply act as a toy that can actively participate in the physical activities of children [29]. Some studies further look into the medical impact of a humanoid robot on supporting the children with autism [51]. Thereon, Pepperpose can fully exploit its capacity in understanding the actions of young users, and provide support in various forms under the guidance of domain experts. For this case, the use of a robot for action sensing becomes rather appropriate, since asking a child to wear devices is not feasible, and they may naturally find it more acceptable to have such a robot friend as a company.

7.2.2 Monitoring Disease Development at Home. Last but not least, PepperPose can operate closely with patients that benefit from a long-term monitor of their motor capacities. In the latest works by Ricotti et al. [50] and Kadirvelu et al. [34], researchers demonstrate how a motion capture system using wearable suits can help monitor the development of motor-impactful diseases like Duchenne muscular dystrophy and Friedreich's ataxia, respectively. This is important, as for these conditions, the patient usually needs to visit their physicians at a certain frequency to report their latest status and inform the doctors' decision of interventions. A motion capture system deployed at home can largely reduce such an effort, and offers a more convenient platform for patient-doctor interactions. Moreover, wearing a full-body motion capture suit can be challenging for these people, and PepperPose may act as a promising alternative to carry out such a monitoring task.

8 LIMITATIONS AND DEVELOPMENT

Our exploration of using a mobile visual robot, Pepper, for active pose estimation has revealed limitations of existing hardware and the proposed framework. Here, we describe them in detail as to open the space for future work.

8.1 The Hardware of PepperPose

The functioning of PepperPose requires three core pieces of hardware, the mobile platform, the camera, and the GPU. In our current practice with the Pepper robot, we only tested with its flat 2D RGB camera operating at 360P@10Hz and have to communicate via TCP with an external GPU to conduct pose estimations. These sometimes resulted in the lost of tracking, given the limited range of viewing angles and the low processing frequency. From another perspective, current algorithms and models for 3D pose estimation are not well adapted for real-time operations, and pose a high demand on compute. While our experiments have showcased promising results in active pose estimation, we acknowledge the need for both hardware and software enhancements before applying it to a range of downstream tasks.

8.2 Fitting with Diverse Environments

Real-world environments often present unexpected obstacles and challenges, making the safe and efficient operation of PepperPose a complex task. For instance, ground-moving robots like Pepper typically require a flat surface, a condition not always guaranteed due to common household features like blankets and stairs. While these issues are prevalent for the daily use of robots, potential solutions may involve enhanced navigation strategies and integration with other sensing systems, such as stationary cameras. For PepperPose specifically, adopting a robot with a smaller size, improved mobility, and a wide-angle camera could be a viable solution. This approach would better accommodate constrained and crowded spaces while minimizing collision risks. Importantly, **PepperPose is not confined to the Pepper robot alone**; our proposed framework is designed to be lightweight and effective, suitable for any mobile visual platform. In parallel, we recognize the critical need to develop user-protection strategies in future developments, particularly for downstream tasks. This entails collaborating with domain experts to clearly delineate its functional boundary and implementing robust privacy protection measures.

8.3 The Extension to Multi-Agent Scenarios

The present configuration of PepperPose does not account for scenarios involving multiple users. During our real-world experiments, we had to conceal ourselves and isolate the experimental space with white boards to prevent the robot from mistakenly tracking someone else within its field of view. One solution to this issue is to integrate person identification algorithms, which would enable the robot to consistently focus on the designated target user. Alternatively, adapting the robot for multi-user environments presents another avenue for development. In such cases, employing multi-agent reinforcement learning [18] or advanced multi-agent large language models [64] could significantly enhance the system's capability. These technologies would allow PepperPose to navigate complex interactions with various users, each having unique requirements and behaviors. This approach would not only rectify current limitations but also expand the system's applicability to more dynamic and varied user interactions.

9 OPEN SOURCE

In the development of this work, one of the major challenges was the scarcity of reference materials, open-source tools, and datasets for training within the newly unveiled Omniverse environment. This platform is vital for our community, as it offers extensive support for human-robot interaction research through its comprehensive APIs and tools. Consequently, to support future development of this community on robot-involved HCI, we release the technical document detailing how to run simple-to-complex robot-human interaction experiments in Omniverse, action data from real participants that can aid the replication of PepperPose, all the essential codes, assets, and useful tools. Please refer to [Hidden for Anonymity] for more details.

10 CONCLUSION

This paper presents PepperPose, a companion robot system developed to track a user's movements and adapt its viewpoint to various actions for active full-body pose estimations.PepperPose eliminates the need for a user to wear special devices or remain within a restricted area, while still delivering high-quality pose estimation. The robot's training utilizes realistic action data in a simulation environment geared towards human-robot interaction. We have showcased its effective performance through a home-like experiment involving 30 participants engaged in a range of actions and orientations. In the future, Pepperpose could serve as the fundamental embodied interactive agent to drive rich applications using its active pose estimation capacity, diverse interactive channels (e.g., its robot arm and voice interface), and the advancement on semantics understanding and language generation of multimodal language models.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China under Grant No. 62132010, Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park No.Z221100006722018, National Key R&D Program of China under Grant No. 2020YFB1313300, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2022A1515110787, and by the Shenzhen Science and Technology Program under Grant No.JSGGKQTD20221101115656029.

REFERENCES

[1] 2023. Kinect. https://azure.microsoft.com/en-us/products/kinect-dk
[2] 2023. Movella Xsens. https://www.movella.com/
[3] 2023. Noitom. https://www.noitom.com/
[4] 2023. Omniverse. https://developer.nvidia.com/omniverse
[5] 2023. OptiTrack. http://optitrack.com
[6] 2023. Pepper. https://us.softbankrobotics.com/pepper
[7] 2023. RealSense. https://www.intelrealsense.com
[8] 2023. Rokoko. https://www.rokoko.com/
[9] 2023. Vicon. https://vicon.com
[10] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. 2020. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6033–6040.
[11] Md Atiqur Rahman Ahad, Anindya Das Antar, and Omar Shahid. 2019. Vision-based Action Understanding for Assistive Healthcare: A Short Review.. In *CVPR Workshops*. 1–11.
[12] Anuparp Boonsongsrikul and Jirapon Eamsaard. 2023. Real-Time Human Motion Tracking by Tello EDU Drone. *Sensors* 23, 2 (2023), 897.

[13] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.
[14] Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. 2021. Goal-conditioned reinforcement learning with imagined subgoals. In *International Conference on Machine Learning*. PMLR, 1430–1440.
[15] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7103–7112.
[16] Wei Cheng, Lan Xu, Lei Han, Yuanfang Guo, and Lu Fang. 2018. IHuman3D: Intelligent Human Body 3D Reconstruction Using a Single Flying Camera. In *Proceedings of the 26th ACM International Conference on Multimedia* (Seoul, Republic of Korea) *(MM '18)*. Association for Computing Machinery, New York, NY, USA, 1733–1741. https://doi.org/10.1145/3240508.3240600
[17] Tomohiro Fujita and Yasutomo Kawanishi. 2023. Future Pose Prediction from 3D Human Skeleton Sequence with Surrounding Situation. *Sensors* 23, 2 (2023), 876.
[18] Yuan Gao, Junfeng Chen, Xi Chen, Chongyang Wang, Junjie Hu, Fuqin Deng, and Tin Lun Lam. 2023. Asymmetric Self-Play-Enabled Intelligent Heterogeneous Multirobot Catching System Using Deep Multiagent Reinforcement Learning. *IEEE Transactions on Robotics* 39, 4 (2023), 2603–2622. https://doi.org/10.1109/TRO.2023.3257541
[19] Brent Griffin. 2023. Mobile Robot Manipulation using Pure Object Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 561–571.
[20] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7297–7306.
[21] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*. Springer, 580–597.
[22] Xiaonan Guo, Jian Liu, Cong Shi, Hongbo Liu, Yingying Chen, and Mooi Choo Chuah. 2018. Device-free personalized fitness assistant using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.
[23] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.
[24] Cherie Ho, Andrew Jong, Harry Freeman, Rohan Rao, Rogerio Bonatti, and Sebastian Scherer. 2021. 3D human reconstruction in the wild with collaborative aerial cameras. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5263–5269.
[25] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–15.
[26] Galadrielle Humblot-Renaux, Letizia Marchegiani, Thomas B Moeslund, and Rikke Gade. 2022. Navigation-oriented scene understanding for robotic autonomy: learning to segment driveability in egocentric images. *IEEE Robotics and Automation Letters* 7, 2 (2022), 2913–2920.
[27] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermudez i Badia. 2023. Design, development, and evaluation of an interactive personalized social robot to monitor and coach post-stroke rehabilitation exercises. *User Modeling and User-Adapted Interaction* 33, 2 (2023), 545–569.
[28] Martin Jacobsson, Jonas Willén, and Mikael Swarén. 2023. A Drone-mounted Depth Camera-based Motion Capture System for Sports Performance Analysis. In *International Conference on Human-Computer Interaction*. Springer, 489–503.
[29] Shomik Jain, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, and Maja J Matarić. 2020. Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Science Robotics* 5, 39 (2020), eaaz3791.
[30] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint arXiv:2306.14795* (2023).
[31] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. 2022. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European Conference on Computer Vision*. Springer, 443–460.
[32] Qingyuan Jiang and Volkan Isler. 2023. Onboard View Planning of a Flying Camera for High Fidelity 3D Reconstruction of a Moving Actor. *arXiv preprint arXiv:2308.00134* (2023).
[33] Yu Jiang, Zhipeng Li, Ziyue Dang, Yuntao Wang, Yukang Yan, Y Zhang, Xinguang Wang, Yansong Li, Mouwang Zhou, Hua Tian, et al. 2022. Facilitating Self-monitored Physical Rehabilitation with Virtual Reality and Haptic feedback. *arXiv preprint arXiv:2209.12018* (2022).
[34] Balasundaram Kadirvelu, Constantinos Gavriel, Sathiji Nageshwaran, Jackson Ping Kei Chan, Suran Nethisinghe, Stavros Athanasopoulos, Valeria Ricotti, Thomas Voit, Paola Giunti, Richard Festenstein, et al. 2023. A wearable motion capture suit and machine learning predict disease progression in Friedreich's ataxia. *Nature Medicine* 29, 1 (2023), 86–94.
[35] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2022. Enabling AI and robotic coaches for physical rehabilitation therapy: iterative design and evaluation with therapists and post-stroke survivors. *International Journal of Social Robotics* (2022), 1–22.
[36] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. 2016. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research* 17, 1 (2016), 1334–1373.
[37] Jiaman Li, Karen Liu, and Jiajun Wu. 2023. Ego-Body Pose Estimation via Ego-Head Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17142–17151.
[38] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. 2022. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia* 25 (2022), 1282–1293.
[39] Yanan Li, Gerolamo Carboni, Franck Gonzalez, Domenico Campolo, and Etienne Burdet. 2019. Differential game theory for versatile physical human–robot interaction. *Nature Machine Intelligence* 1, 1 (2019), 36–43.
[40] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. 2020. Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4, 4 (2020), 1–28.
[41] Jingyuan Liu, Nazmus Saquib, Zhutian Chen, Rubaiat Habib Kazi, Li-Yi Wei, Hongbo Fu, and Chiew-Lan Tai. 2022. PoseCoach: A Customizable Analysis and Visualization System for Video-based Running Coaching. *IEEE Transactions on Visualization and Computer Graphics* (2022).
[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics* 34, 6 (2015).
[43] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 2023. 3D Human Mesh Estimation From Virtual Markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 534–543.
[44] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12.
[45] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 483–499.
[46] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 7, 3 (2006), 437–454.
[47] Claudio Pizzolato, Monica Reggiani, Luca Modenese, and David G Lloyd. 2017. Real-time inverse kinematics and inverse dynamics for lower limb applications using OpenSim. *Computer methods in biomechanics and biomedical engineering* 20, 4 (2017), 436–445.
[48] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083* (2023).
[49] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2022. GoPose: 3D human pose estimation using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–25.
[50] Valeria Ricotti, Balasundaram Kadirvelu, Victoria Selby, Richard Festenstein, Eugenio Mercuri, Thomas Voit, and A Aldo Faisal. 2023. Wearable full-body motion tracking of activities of daily living predicts disease trajectory in Duchenne muscular dystrophy. *Nature medicine* 29, 1 (2023), 95–103.
[51] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics* 3, 19 (2018), eaao6760.
[52] Panneer Selvam Santhalingam, Al Amin Hosain, Ding Zhang, Parth Pathak, Huzefa Rangwala, and Raja Kushalnagar. 2020. mmasl: Environment-independent asl gesture recognition using 60 ghz millimeter-wave signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.
[53] Beat Schäffer, Reto Pieren, Kurt Heutschi, Jean Marc Wunderli, and Stefan Becker. 2021. Drone noise emission characteristics and noise effects on humans—a systematic review. *International journal of environmental research and public health* 18, 11 (2021), 5940.
[54] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652* (2023).
[55] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5693–5703.

[56] Rahul Tallamraju, Nitin Saini, Elia Bonetto, Michael Pabst, Yu Tang Liu, Michael J Black, and Aamir Ahmad. 2020. AirCapRL: autonomous aerial human motion capture using deep reinforcement learning. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6678–6685.

[57] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, Vol. 36. Wiley Online Library, 349–360.

[58] Chongyang Wang, Yuan Gao, Akhil Mathur, Amanda C De C. Williams, Nicholas D Lane, and Nadia Bianchi-Berthouze. 2021. Leveraging activity recognition to enable protective behavior detection in continuous data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–27.

[59] Hanchen David Wang and Meiyi Ma. 2023. PhysiQ: Off-site Quality Assessment of Exercise in Physical Therapy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–25.

[60] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2020. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*. Springer, 764–780.

[61] Ziming Wang, Ziyi Hu, Björn Rohles, Sara Ljungblad, Vincent Koenig, and Morten Fjeld. 2023. The Effects of Natural Sounds and Proxemic Distances on the Perception of a Noisy Domestic Flying Robot. *ACM Transactions on Human-Robot Interaction* (2023).

[62] Fabian C Weigend, Shubham Sonawani, Michael Drolet, and Heni Ben Amor. 2023. Anytime, Anywhere: Human Arm Pose from Smartwatch Data for Ubiquitous Robot Control and Teleoperation. *arXiv preprint arXiv:2306.13192* (2023).

[63] Anna Wojciechowska, Jeremy Frey, Sarit Sass, Roy Shafir, and Jessica R Cauchard. 2019. Collocated human-drone interaction: Methodology and approach strategy. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 172–181.

[64] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).

[65] Lan Xu, Yebin Liu, Wei Cheng, Kaiwen Guo, Guyue Zhou, Qionghai Dai, and Lu Fang. 2017. Flycap: Markerless motion capture using multiple autonomous flying cameras. *IEEE transactions on visualization and computer graphics* 24, 8 (2017), 2284–2297.

[66] Weipeng Yang, Haoran Luo, and Jiahong Su. 2022. Towards inclusiveness and sustainability of robot programming in early childhood: Child engagement, learning outcomes and teacher perception. *British Journal of Educational Technology* 53, 6 (2022), 1486–1510.

[67] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* (2023).

[68] Yongjing Ye, Libin Liu, Lei Hu, and Shihong Xia. 2022. Neural3Points: Learning to Generate Physically Realistic Full-body Motion for Virtual Reality Users. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 183–194.

[69] Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2021. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

[70] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. 2020. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 507–523.

[71] Feng Zhang, Chenshu Wu, Beibei Wang, Hung-Quoc Lai, Yi Han, and KJ Ray Liu. 2019. WiDetect: Robust motion detection with a statistical electromagnetic model. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.

[72] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. 2023. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8877–8886.

[73] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. 2023. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8877–8886.

[74] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 2021. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11656–11665.

[75] Bo Zhou, Daniel Geissler, Marc Faulhaber, Clara Elisabeth Gleiss, Esther Friederike Zahn, Lala Shakti Swarup Ray, David Gamarra, Vitor Fortes Rey, Sungho Suh, Sizhen Bian, et al. 2023. MoCaPose: Motion Capturing with Textile-integrated Capacitive Sensors in Loose-fitting Smart Garments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–40.

[76] Xiaowei Zhou, Sikang Liu, Georgios Pavlakos, Vijay Kumar, and Kostas Daniilidis. 2018. Human motion capture using a drone. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2027–2033.

[77] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. 2018. 3d human pose estimation in rgbd images for robotic task learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1986–1992.

A APPENDIX

A.1 Goal-Conditioned Reinforcement Learning in Omniverse

We used goal-conditioned reinforcement learning (RL) with Omniverse simulator to refine the action space (i.e., the set of all possible actions of a robot) of Pepper robot and build its kinematics model. In this way, after the simulation, the robot is able to track the user in a smooth and safe manner. This approach requires us to consider the movement of the Pepper robot as a Markov decision process, allowing the use of RL techniques such as domain randomization. Through assigning random goals, the agent can gradually learn to reach goals around them with consideration of the kinematics of the robot. Here, we report the goal-conditioned RL used in this study that involves three major components, namely environment setup, observation, and reward, as follows:

- **Environmental Setup:** As stated previously, we use Nvidia Omniverse for constructing our simulated learning environment, which includes parallel learning agents. The robot model is trained using the Proximal Policy Algorithm within the Orbit library framework [3], an RL training extension of Omniverse designed for simulations with models like Pepper and human characters. The virtual human model is initially created in Blender [4] using our collected MoCap data and subsequently imported into Omniverse. Similarly, the virtual Pepper robot is integrated by converting its official Unified Robotics Description Format (URDF) asset [5] into Universal Scene Description (USD) models.
- **Observation:** The observation data pertaining to the Pepper robot encompasses various elements. These include the precise coordinates and orientation of the robot itself, and the randomly sampled viewpoints.
- **Reward:** Given a goal-conditioned policy, we train the Pepper in simulation to go to any given position, driven by the kinematics dynamics it has been assigned to. The general reward is defined as the negative Euclidean distance between the position of the pepper robot and the position of the goal it has been assigned to.

A.2 Sim2Real Deployment

After the training, the pepper robot can reach the provided goal as a position swiftly. During the testing stage, we provide the corresponding location and orientation of the virtual human model, the relative distance between the robot and the human. Furthermore, we also consider the visual information captured by the Pepper robot in the form of images. To get a first-person view, we mounted a camera element to the simulated Pepper robot and compare the differences between the pose estimated by utilizing the estimator

[3] https://github.com/NVIDIA-Omniverse/Orbit
[4] https://www.blender.org/
[5] https://github.com/ros-naoqi/pepper_robot/blob/master/pepper_description/urdf/pepper1.0_generated_urdf/pepper.urdf

referred to as PoseformerV2 [73] and the ground truth pose of the realistic action data that is used to drive the action of simulated people. The captured frames in this environment are realistic, making it possible for us to directly test the performance of the proposed framework in simulation and tune parameters for controllers and planners, e.g., goal assignment per time step for the RL model, path planner given the prior knowledge on viewpoints, and evaluate action spaces and kinematics model of the robot. Figure 4 illustrates the ground truth and the estimated poses. In each parallel environment, a random human action is sampled, while the task of the Pepper robot is to optimize the overall estimation accuracy. After selecting the action space and the establishment of the kinematics model, the next step is to deploy these with a Pepper robot in the real world. This includes using a simple yet efficient Proportional Integral Derivative (PID) controller to control the robot's velocity, allowing the robot to reach the desired viewpoint suggested by other modules of PepperPose.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Figure 15: The slides used in our real-world experiment, which act as the instruction to inform the participant about which action to conduct, changing orientation, and the remaining repetitions.