

mmHRI: Towards Privacy-Preserving Human-Robot Interaction with Millimeter-Wave Radar

Junqiao Fan¹, Yuxuan Hu², Bofan Lyu², Yanshuo Lu², Pengfei Liu²,
Jiarui Zhang¹, Fangqiang Ding³, Lihua Xie¹, Gen Li^{4,†}, Jianfei Yang^{2,†}

Abstract—Assistive robots increasingly operate in many human-centered environments and perform various human-robot interaction (HRI) tasks, such as object delivery. However, most existing HRI systems rely on RGB cameras that continuously observe humans to respond to non-verbal commands, such as hand gestures. This raises privacy concerns in privacy-critical environments, such as hospital wards or restaurants, where direct camera observation of humans is restricted. To develop privacy-preserving HRI, we leverage millimeter-wave (mmWave) radar, which can sense human motion through privacy barriers without identifiable imagery. We propose mmHRI, the first multi-modal robot manipulation framework that achieves mmWave radar-guided privacy-preserving HRI. mmHRI introduces two key designs to mitigate the sparsity and temporal inconsistency of radar data in cluttered robot manipulation environments. First, we propose a dual-stream architecture that jointly learns from unfiltered raw radar tensors and radar point clouds to estimate both human actions and 3D poses. To mitigate signal inconsistency, mmHRI further incorporates a memory-based state-space model (MSSM) that retains historical radar features to reduce abrupt changes in pose/action. These estimated human states are then converted into structured textual robot instructions, which control a vision-language-action (VLA) policy for closed-loop robot manipulation and human-aware reactions. Our evaluation covers human action recognition and closed-loop delivery and retrieval. In the privacy-preserving curtain setting, mmHRI achieves 85.09% action-recognition accuracy, outperforming existing radar-based alternatives. Robot trials further demonstrate successful delivery and retrieval under visual occlusion, with stable task performance across unseen subjects, clutter configurations, and environments.

I. INTRODUCTION

Recent advances in perception, manipulation [1], and robot foundation models [2], [3] have accelerated the deployment of service robots in hospitals, restaurants, retail stores, and other human-centered environments. In these applications, effective human-robot interaction (HRI) requires robots to understand not only object states but also human intent, including where and when users expect a response. Although verbal instructions can specify intent, humans also convey much of their intent through non-verbal cues [4], such as hand gestures. For example, during object delivery, a user may gesture toward one side to indicate a preferred delivery location or wave a hand to signal readiness.

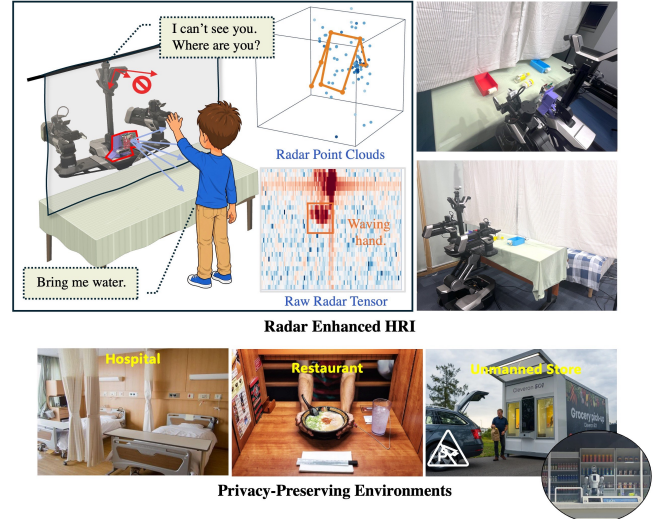


Fig. 1. Motivation of privacy-preserving mmHRI using mmWave radar for human sensing. Unlike conventional vision-based HRI that fails under occlusion, mmHRI senses human intent through privacy curtains to coordinate object delivery without directly observing humans. This supports privacy-sensitive applications such as hospitals, restaurants, and unmanned stores.

Most existing HRI systems rely on RGB-D cameras to perceive non-verbal human interaction cues. However, RGB-based systems require direct visual observation of the interacting person, leading to two practical limitations. First, cameras capture faces, clothes, and body images that raise **privacy concerns**. Therefore, direct visual observation is often restricted in privacy-sensitive environments [5], [6], [7]. For example, privacy curtains in hospital wards intentionally shield patients from observation during rest and treatment. Similarly, wooden partitions in some Japanese restaurants (e.g., ICHIRAN) separate customers from staff and other diners to ensure dining privacy. These privacy barriers make it more challenging for robots to perceive user intent and respond promptly without compromising privacy. Moreover, varying illumination conditions, such as darkness and strong sunlight, may also degrade visual observations and reduce the reliability of human state perception [8], [9].

Millimeter-wave (mmWave) radar has emerged as a promising and affordable complement to existing line-of-sight (LoS) camera-based systems. Sensitive to moving targets, mmWave radar captures human motion by transmitting and receiving radio-frequency (RF) signals without revealing identifiable visual information (e.g., facial appearance or clothing) [10]. Its signals can penetrate certain visually opaque materials (e.g., curtains or wooden partitions), and

¹School of Electrical and Electronic Engineering, Nanyang Technological University. ²School of Mechanical and Aerospace Engineering, Nanyang Technological University. ³The Hong Kong University of Science and Technology (Guangzhou). ⁴School of AI and Robotics, Hunan University. [†]Corresponding authors: ligen@g.skku.edu; jianfei.yang@ntu.edu.sg.

is more robust to illumination changes [11]. These properties have motivated various radar-based human sensing applications [12], [13], [14]. However, most existing radar systems are developed for clean, uncluttered monitoring environments. Using radar for HRI and robot manipulation is more challenging. First, commercial radar has lower resolution and produces **sparse radar point clouds (RPC)** around moving human body parts [15]. Detected body parts may occasionally disappear due to specular reflections [16], causing temporal inconsistencies in RPCs. Second, robot-manipulation environments introduce additional **sensing interference (clutter)** compared with conventional monitoring environments. Furniture (e.g., tables), delivered objects, privacy curtains, and moving robot arms can generate multipath reflections and false ghost targets [17], further increasing temporal inconsistency in human perception.

To achieve privacy-preserving HRI, we propose **mmHRI**, the first learning-based radar-vision multimodal HRI framework that integrates mmWave radar as an additional human-perception modality. mmHRI is built on a vision-language-action (VLA) backbone and can be implemented with any off-the-shelf alternatives. Its RGB cameras observe only the robot working area for fine-grained object manipulation. Meanwhile, mmWave radar captures human non-verbal intent (e.g., gestures and poses) to determine where and when the robot should react. To compensate for RPC sparsity, we first propose dual-stream radar-based human perception (DRP), exploring the raw radar tensor and its fusion with RPC. Specifically, the raw Range-Doppler-Time radar heatmap preserves more unfiltered temporal micro-velocity information for human action/gesture estimation, while the filtered 3D RPC retains 3D geometry information for human pose estimation. To further mitigate temporal inconsistency in radar signals, we introduce a memory state-space model (MSSM) to track human states over time and memorize historical radar patterns. This prevents occasional curtain motion or robot-arm motion from being misinterpreted as human commands. Finally, human-aware text reasoning (HTR) converts the radar-estimated human states, including gestures and human poses, into text-based robot reaction descriptions. These structured text commands are fed into a VLA policy for closed-loop manipulation that adapts to real-time human intent.

We evaluate mmHRI using both real-world robot trials and our collected human perception dataset for privacy-preserving HRI. mmHRI achieves higher accuracy and robustness than existing radar-based systems. The robot trials further demonstrate that mmHRI is more robust under occlusion than existing RGB-based systems. Extensive experiments also show robustness across unseen subjects, clutter, and environments. The main contributions of this work are as follows:

- 1) We present **mmHRI**, the first multimodal HRI framework that integrates mmWave radar human perception into a robot manipulation foundation model, achieving closed-loop privacy-preserving object delivery.
- 2) We introduce a dual-stream radar perception (DRP)

architecture and a memory state-space model (MSSM) for radar-based human perception, tailored for cluttered robot manipulation environments.

- 3) We conduct systematic experiments using our collected HRI-oriented human perception dataset and real-world robot trials to evaluate human perception performance and its downstream effects on robot decision-making.

II. RELATED WORK

A. Radar-Based Human Perception

Human action recognition (HAR), human pose estimation (HPE), and other human-perception technologies are important computer-vision tasks for various applications, such as virtual reality [18], surveillance [19], and HRI [20]. Most existing methods rely on RGB-D camera systems [21], [22], which remain limited by visual occlusion, changing illumination [8], and privacy concerns [23], [7]. Radar has therefore been adopted as a complementary or alternative modality for human sensing. Early works focus on coarse, clustering-based human tracking [13], [24] from sparse radar point clouds (RPCs). Subsequent works [12], [25] explore RPC-based action and gesture recognition using deep-learning algorithms. Recent methods further explore higher-fidelity pose estimation using RPCs [11], [8], [14] or raw radar tensors [9], [26]. Nevertheless, these methods typically assume clean, uncluttered sensing environments, whereas HRI and robot-manipulation scenarios are usually severely cluttered. Several works have integrated mmWave radar into robotic tasks: WaveMan [25] provides RPC-based recognition of four gestures for teleoperating robots in an open room. OmniVLA [27] incorporates radar to locate non-line-of-sight objects inside boxes. It applies a simple masked overlay of radar measurements and RGB observations to identify which object to grasp. Yet, these methods remain restricted to static objects or relatively clean and uncluttered environments.

B. Learning-Based Human-Robot Interaction

Traditional human-robot interaction methods [28], [29], [30] typically require RGB-based pose estimation and subsequent action recognition to estimate the human state. They typically rely on predefined robot responses for tasks such as human following [31] and object delivery [29], [30], [32]. Recently, learning-based robot policies have achieved better generalization in object manipulation. Diffusion Policy learns visuomotor control directly from visual observations [1], while OpenVLA [2] and $\pi_{0.5}$ [3] connect visual observations and language instructions to robots. However, these policies generally lack human understanding and rely on detailed language descriptions, which is inefficient for HRI. HABIT [33], Gaze2Act [34] and more [32], [35], [36] incorporate human non-verbal information into VLA policies for HRI, but require direct RGB observation of the human or a clear view for gesture and gaze recognition. Consequently, these methods fail when humans are outside the visual line of sight under privacy-sensitive scenarios. Several methods [37] design wearable sensing systems to support non-line-of-sight interaction but require strong user compliance with

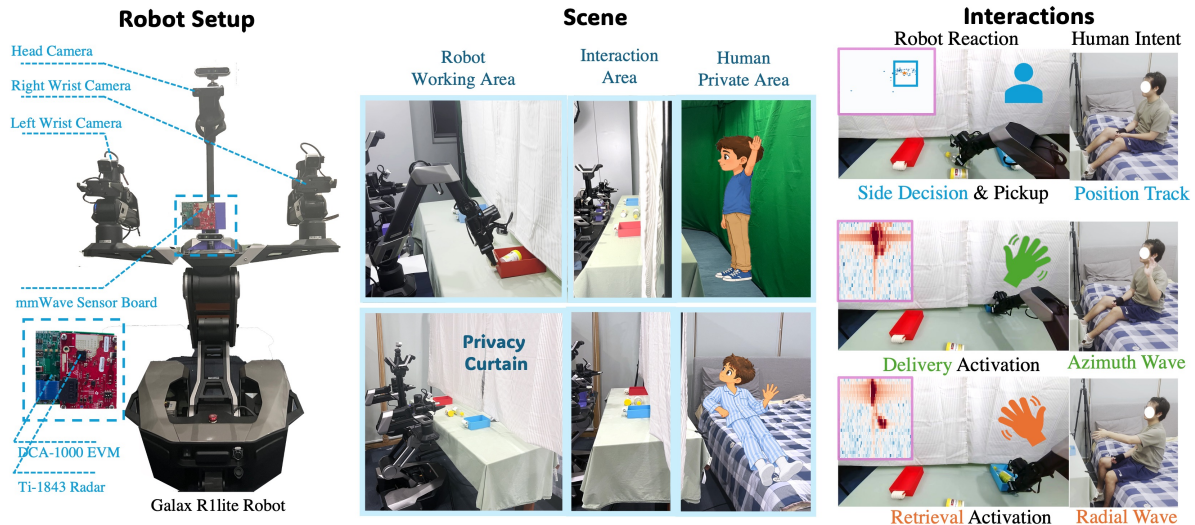


Fig. 2. Left: Physical robot system setup and the mmWave sensing unit. Middle: Two privacy-preserving scenes, i.e., hospital ward and office. Right: Three gesture/pose based human-robot-interaction decision-making tasks. We show different radar signals that the robot may refer to for different reactions.

wearing the devices. Therefore, incorporating radar into HRI is promising to address remote privacy-preserving human perception under visual occlusion, but its real-world deployment for HRI remains underexplored.

III. SYSTEM OVERVIEW

Robot Setup. As presented in Figure 2, our platform is built on a R1 Lite dual-arm mobile robot. The robot is equipped with a head camera that focuses on the working table. Two wrist cameras are mounted on the robot’s left and right arms to support fine-grained object manipulation. Human perception is achieved by a mmWave radar sensing board mounted on the robot’s chest, which can operate under darkness or through occlusions. It consists of a TI IWR1843 mmWave radar (3TX 4RX) and a DCA1000 data-capture board for 10 Hz raw data acquisition. The sensing board and robot transmit synchronized radar and camera observations to a Linux policy server, which performs perception, reasoning, and policy inference. It then generates action chunks and sends them back to the robot for execution.

Scene Setup. Our scenes are set up for privacy-preserving human-robot interaction (HRI) in two different environments, simulating hospital and office settings. As illustrated in Figure 2, each scene comprises three areas: (1) The robot working area is observed by the robot’s RGB cameras to support object manipulation. The fields of view of all onboard cameras are restricted to this area; (2) The interaction area is accessible by both the robot and the human, including a table for conducting object delivery and a curtain for privacy protection; and (3) The human private area is occupied only by the human and cannot be directly monitored by RGB cameras. This physical setup inevitably contains several sources of radar interference, including the table, robot hardware, curtain, and nearby objects.

Problem Formulation. We study a classic human-robot interaction task: tabletop/bedside object delivery. The robot

is tasked to deliver human-specified objects to their reachable region, understanding gesture intent, such as where and when the delivery should take place. First, the robot identifies the requested object and uses the user’s location and gestures to select the delivery side (Where). Second, after an azimuth-wave greeting gesture (When), it grasps the object, places it in a delivery box, and pushes the box within the user’s reach. Finally, it waits and retrieves the delivery box until it recognizes a radial wave (When). Throughout the task, it continuously monitors the user’s position and pauses actions whenever the user is near the robot to avoid collisions.

We formulate this HRI task as a learning problem. Given radar observations O_{radar} , RGB observations O_{rgb} , and a simple spoken task description T_{task} , the mmHRI policy π outputs a 20-step action chunk A for the two robot arms: $A = \pi(O_{\text{radar}}, O_{\text{rgb}}, T_{\text{task}})$, $A \in \mathbb{R}^{T_{\text{chunk}} \times 2N_{\text{arm}}}$, where $T_{\text{chunk}} = 20$ and $N_{\text{arm}} = 7$ in our physical robot setup. The task description specifies the requested object, e.g., “Bring me water.” The radar observations O_{radar} are raw I/Q signals from the mmWave sensor board for real-time human-state perception. The RGB observations O_{rgb} are multi-view images from robot-mounted cameras for object manipulation.

IV. METHOD

As presented in Figure 3, we first preprocess the input raw radar signals into a radar tensor and a radar point cloud (Sec. IV-A). We then use a dual-stream radar perception network to estimate human actions, 3D pose, and root location (Sec. IV-B). Next, we design a human-aware text reasoning (HTR) module that converts the estimated human state and task description into a structured text prompt (Sec. IV-C). Finally, this prompt conditions a single VLA policy to generate different robot reactions (Sec. IV-C).

A. Radar Signal Preprocessing

We process the raw complex ADC signals into two representations for human-state estimation: a range-Doppler-time

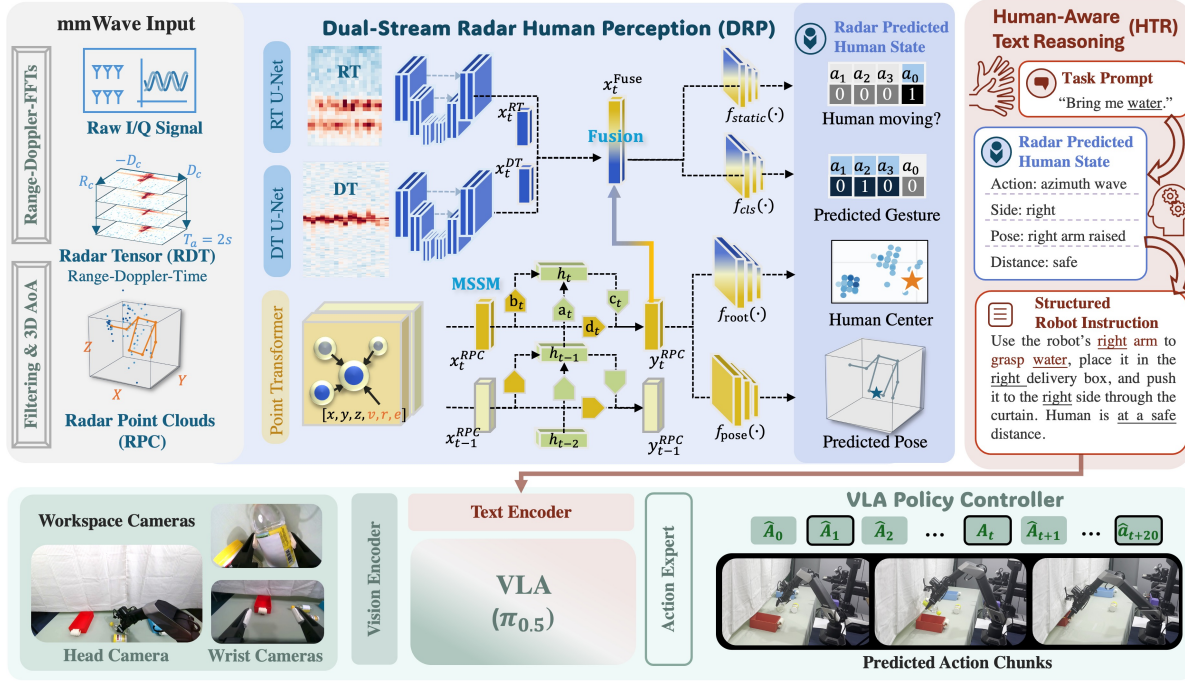


Fig. 3. Overview of mmHRI. Radar measurements are first preprocessed into RDT tensors and RPC. The dual-stream radar human perception (DRP) then extracts motion and geometry patterns from both modalities, which are fused to jointly predict action and 3D poses. The human-aware text reasoning (HTR) then converts the estimated human states into structured robot instructions, controlling the downstream VLA policy for different actions.

(RDT) tensor and a radar point cloud (RPC). To construct the raw radar RDT tensor, we first generate range–Doppler (RD) maps by applying a range FFT and a Doppler FFT to the ADC signals. We apply mean pooling across all virtual antennas to reduce computation. Following mmMesh [11], we perform chirp-wise mean subtraction on the range maps to remove background clutter. The RD map produces a heatmap that is sensitive to micro-Doppler motion responses, providing information about how far the human is from the robot and how the human moves. Finally, we stack $T_a = 20$ RD maps spanning 2 s to construct the RDT tensor. To obtain the RPC, we use top- K ($K = 128$) energy selection to detect targets with the most salient velocity responses. Then, we calculate the angle of arrival (AoA) to obtain the azimuth and elevation angles of the detected targets, generating the 3D RPC. We observe that human motion produces stronger reflections in the RD map, whereas multipath reflections generally lose energy. Therefore, we apply an additional energy threshold to further suppress multipath points. Finally, following previous works [14], we interpolate RPCs from four historical frames into one frame to reduce sparsity.

B. Dual-Stream Radar-Based Human State Estimation

Dual-stream radar-based human perception uses both the unfiltered RDT tensor O^{RDT} and the filtered RPC O^{RPC} for human-state estimation. It designs a multi-task framework, simultaneously performing action classification, root tracking, and pose estimation. Specifically, we use the RDT tensor for high-level, spatially agnostic action recognition. Meanwhile the filtered RPC supports pose estimation and root tracking because it preserves 3D spatial information.

RDT Feature Extraction. The RDT stream receives O^{RDT} as inputs. The RDT tensor is higher-dimensional and generally retains more unfiltered motion-related information than the standard RPC. However, directly applying 3D convolution over the entire tensor is computationally inefficient. Therefore, we reshape the tensor and apply two 2D convolution U-Nets along two viewing directions, Range-Time (RT) and Doppler-Time (DT). The additional dimensions are treated as feature channels. As illustrated in Figure 4, the RT and DT representations capture subtle micro-motion responses for different hand-waving gesture actions. Specifically, we use efficient three-layer RT and DT U-Nets for feature extraction, producing x_t^{RT} and x_t^{DT} at time t .

Memory State-Space-Model (MSSM). At time t , the RPC stream receives O_t^{RPC} and applies a Point Transformer [38] to perform point-wise self-attention and extract the RPC feature x_t^{RPC} . To mitigate temporal inconsistency in radar signals, we design a state-space model (ssm) to enhance the RPC feature with historical memory. The memory state h_t stores both current-frame x_t^{RPC} information and historical information from previous h_{t-1} . As shown in Figure 3, the state is updated using four coefficients that depend on the current RPC feature x_t^{RPC} :

$$\begin{aligned} a_t &= g_a(x_t^{\text{RPC}}), & b_t &= g_b(x_t^{\text{RPC}}), \\ c_t &= g_c(x_t^{\text{RPC}}), & d_t &= g_d(x_t^{\text{RPC}}), \end{aligned} \quad (1)$$

where g_a , g_b , g_c , and g_d are learnable one-layer MLPs. The memory state evolves as:

$$h_t = a_t h_{t-1} + b_t x_t^{\text{RPC}}. \quad (2)$$

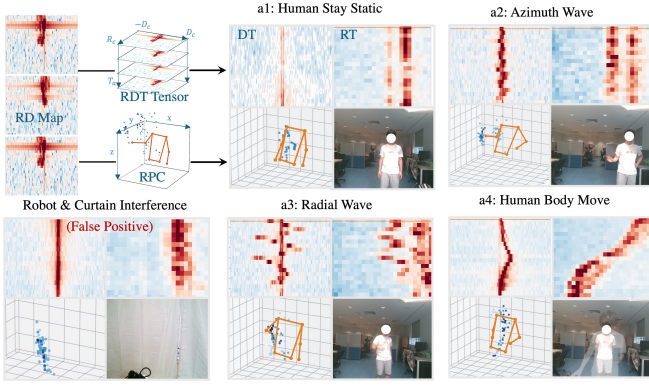


Fig. 4. Visualization of the complementary radar modalities used by the dual-stream model. The DT and RT maps preserve motion-sensitive Doppler patterns, whereas the RPC retains 3D spatial structure for pose and root estimation. Synchronized RGB images are shown for visual reference.

The output of MSSM y_t^{RPC} is then calculated as:

$$y_t^{\text{RPC}} = c_t h_t + d_t x_t^{\text{RPC}}. \quad (3)$$

This recurrence design carries information across historical frames while retaining current spatial features, preventing subtle changes caused by occasional signal inconsistency.

Dual-Stream Feature Fusion. We concatenate the RDT features with the MSSM output along the feature dimension:

$$x_t^{\text{Fuse}} = [x_t^{\text{RT}}; x_t^{\text{DT}}; y_t^{\text{RPC}}]. \quad (4)$$

This feature combines both unfiltered motion information from RDT and spatial information from RPC. As shown in Figure 3, the fused feature is then utilized by four MLP heads to estimate four different human states: First, a static head f_{static} predicts whether a non-static human action is present. Meanwhile, an action classification head f_{cls} predicts the logits for three non-static actions: azimuth wave, radial wave, and body move:

$$\begin{aligned} \hat{p}_{\text{static},t} &= \text{sigmoid}(f_{\text{static}}(x_t^{\text{Fuse}})), \\ \hat{p}_{\text{act},t} &= \text{softmax}(f_{\text{cls}}(x_t^{\text{Fuse}})). \end{aligned} \quad (5)$$

To obtain the 3D human position, the root tracking head f_{root} first predicts the coordinates of the human root:

$$\hat{\tau}_t = f_{\text{root}}(x_t^{\text{Fuse}}) \in \mathbb{R}^3. \quad (6)$$

Finally, the pose estimation head f_{pose} predicts the root-normalized 3D positions of eight upper-body joints:

$$\hat{P}_t = f_{\text{pose}}(y_t^{\text{RPC}}) \in \mathbb{R}^{8 \times 3}. \quad (7)$$

Together, $\hat{p}_{\text{static},t}, \hat{p}_{\text{act},t}, \hat{\tau}_t, \hat{P}_t$ form the structured human states and are used by the downstream VLA policy.

C. Human-State-Conditioned VLA Policy

Human-Aware Text Reasoning (HTR). As illustrated in Figure 3, we convert the task text and radar-derived human state into a structured robot instruction. The instruction template contains five fields: the requested object from T_{task} , and the action, receiving side, pose, and safety distance from $S_{\text{human}} = \{\hat{p}_{\text{static},t}, \hat{p}_{\text{act},t}, \hat{\tau}_t, \hat{P}_t\}$. The action class decoded

from action predictions determines the current interaction stage (e.g., delivery or retrieval); the root and hand locations determine the receiving side (e.g., left or right); and the human distance indicates whether execution is safe. These fields are instantiated as an imperative instruction specifying which arm to use, which object to grasp, which delivery box to use, and whether the robot may act. For example, the task text “Bring water” and the human-state descriptions trigger the structured instruction T_{robot} shown in Figure 3. This structured robot prompt T_{robot} directly controls downstream VLA policy for different robot action execution.

VLA Control Policy. We adopt $\pi_{0.5}$ [3] as the VLA backbone. At time t , the visual observation contains one head-camera image and two wrist-camera images, $O_t^{\text{robot}} = \{O_{\text{head},t}, O_{\text{lwrist},t}, O_{\text{rwrist},t}\}$. The PaliGemma vision-language model (VLM) encodes the images using a SigLIP-So400m visual encoder and T_{robot} using a Gemma-2B text encoder. The resulting visual and language features are jointly provided to a transformer-based flow-matching action head, which models the conditional distribution of a 20-step action chunk for the two 7-DoF robot arms:

$$A_{t:t+20} \sim \pi_{0.5}(\cdot | O_t^{\text{robot}}, T_{\text{robot}}), \quad A_{t:t+20} \in \mathbb{R}^{20 \times 14}. \quad (8)$$

Here, $A_{t:t+20}$ denotes the action chunk over the interval $[t, t+20)$. This interface enables a single policy to perform object selection, grasping, basket placement, delivery, waiting, and retrieval according to the current human state.

V. EXPERIMENTS

We design our experiments to answer three core questions.

Q1. How does mmWave radar benefit privacy-preserving and robust HRI in real-robot trials under clear vision and visual occlusion? **Q2.** How does the proposed mmHRI radar-perception model compare with existing radar-based methods? **Q3.** How robust is mmHRI under unseen human subjects, environments, and tabletop clutter setups?

A. Experimental Details

Implementation Details. The VLA policy is initialized from the pretrained $\pi_{0.5}$ and fine-tuned with full parameters. We collect 160 real-world teleoperated demonstrations with synchronized human-state descriptions (e.g., action, pose, root), structured robot instructions T_{robot} . The policy is optimized using AdamW for 1000 epochs with a learning rate of 1×10^{-8} and a batch size of 64. Training takes 26 hours on four NVIDIA A800 GPUs. The DRP module adopts three-layer CNN U-Nets for the RDT tensor encoders, and the human state decoder are implemented with two-layer MLPs. The entire DRP is pre-trained separately using AdamW for 10 epochs with a learning rate of 1×10^{-4} and a batch size of 64. We collect a radar perception dataset for DRP training and validation. The dataset includes 12K frames of a clear-visibility office scene (Clear Visibility); 2K testing frames of the same curtain-occluded office scene (Occlusion); and 2K testing frames of a curtain-occluded hospital ward scene (Occlusion Cross-Env). We use 80% of Clear Visibility data for training and the remaining for testing.

TABLE I: **Performance on the radar perception dataset.** HPE is evaluated under clear visibility, and HAR is evaluated under clear visibility, privacy occlusion, and cross-environment occlusion.

Methods	HPE		Clear Visibility				Occlusion (Privacy)				Occlusion (Cross-Env)			
	MPJPE (cm)	TE (cm)	ACC (%)	FPR (%)	Static-F1 (%)	HPE Jitter	ACC (%)	FPR (%)	Static-F1 (%)	HPE Jitter	ACC (%)	FPR (%)	Static-F1 (%)	HPE Jitter
RGB														
RGB (YOLOv.) + AGCN	.	.	96.02	4.55	97.06	.	.	.	.	.	.	.	.	.
mmWave Radar														
PointTrans. (RPC) + AGCN	7.40	6.26	86.18	16.18	89.58	3.52	72.49	23.11	88.36	3.91	31.06	40.84	67.91	6.08
RadHAR (RPC)	.	.	87.06	11.95	77.11	.	62.18	26.09	82.50	.	27.61	33.10	61.81	.
Waveman (RT)	.	.	91.65	4.47	91.33	.	79.65	12.67	86.95	.	62.31	0.00	82.28	.
Ours														
mmHRI (Fusion)	7.54	6.01	92.56	4.55	95.14	3.54	81.45	5.00	88.69	3.45	68.94	20.42	85.83	5.15
mmHRI (Fusion + MSSM)	4.25	3.12	96.51	1.38	97.25	2.83	85.09	9.90	90.17	3.25	74.91	11.27	86.09	4.42

TABLE II: **Performance on real-world robot trials.** PA and Success report the perception-only and end-to-end stage success rates. MA reports the success rate over 30 manipulation-only trials with ground-truth human states.

Methods	Side Decision & Grasping			Object Delivery			Box Retrieval			Collision Avoidance		
	PA	MA	Success	PA	MA	Success	PA	MA	Success	PA	MA	Success
Clear Visibility												
RGB + pi05	30/30	26/30	26/30	30/30	27/30	27/30	27/30	30/30	27/30	30/30	30/30	30/30
mmWave+pi05	30/30	26/30	26/30	29/30	27/30	26/30	29/30	30/30	29/30	30/30	30/30	30/30
Occlusion												
RGB+pi05	0/30	26/30	0/30	0/30	27/30	0/30	0/30	30/30	0/30	0/30	30/30	0/30
mmWave + pi05	28/30	26/30	25/30	28/30	27/30	25/30	29/30	30/30	29/30	30/30	30/30	30/30
Occlusion (Cross-Environment)												
RGB+pi05	0/30	23/30	0/30	0/30	26/30	0/30	0/30	28/30	0/30	0/30	30/30	0/30
mmWave + pi05	28/30	23/30	21/30	25/30	26/30	22/30	26/30	28/30	25/30	25/30	30/30	25/30

Evaluation Metrics. For human action recognition (HAR), we report classification accuracy (ACC), false-positive rate (FPR) and F1-score. The FPR measures the proportion of incorrect activation of static samples. The F1-score measures the performance over precision and recall. For human pose estimation (HPE), mean per-joint position error (MPJPE) measures the average Euclidean error of root-normalized 3D joints, translation error (TE) measures the Euclidean error of the human root, and jitter measures the frame-wise instability of the predicted pose; all three are reported in centimeters. In real-robot trials, perception accuracy (PA) measures the success rate of human intent detections during the end-to-end trials, including side decisions, delivery/retrieval activation, and collision detections. Manipulation accuracy (MA) separately reports the VLA policy manipulation success rate over 30 manipulation-only trials using ground-truth human states. Finally, Success indicates the overall stage success rate during end-to-end trials.

B. Radar Perception Discussion

As presented in Table I, we compare mmHRI with existing radar-based perception methods. The RGB baseline first estimates human pose and then performs skeleton-based action recognition [22]. The RGB baseline achieves the best HAR performance under clear visibility, but cannot produce valid HAR or HPE predictions when humans are non-line-of-sight. In contrast, all mmWave radar-based methods perform

HAR under both clear-visibility and occlusion conditions. Traditional RPC-based RADHAR [12] shows poor generalization, with ACC of only 62.18% under occlusion and 27.61% across environments. Replacing point-cloud input with a radar tensor and using the WaveMan [25] backbone improves ACC to 79.65% under occlusion and 62.31% across environments. This suggests that Doppler-rich motion patterns captured by RDT tensors are more informative and robust under cluttered HRI environments. However, radar tensors sacrifice spatial information and cannot support 3D spatial tasks such as HPE and tracking. mmHRI addresses this limitation through DRP fusion, which achieves joint 3D HPE and HAR. We observe that mmHRI achieves the best 85.09% ACC under occlusion and 74.91% ACC across environments, and reduces the FPR to below 11.27%. The lower FPR reduces the risk that clutter or radar multipath interference is misclassified as an interaction gesture, which could trigger premature delivery/retrieval. In addition, the MSSM module reduces HPE error and temporal jitter. By incorporating historical states and signal information, it mitigates abrupt changes and radar’s temporal inconsistency, resulting in more stable spatial reasoning.

C. Real-World Robot Trial Discussion

As shown in Table II, our real-world robot trials evaluate both PA and MA. PA measures whether the robot recognizes human intents, and MA evaluates how many failures

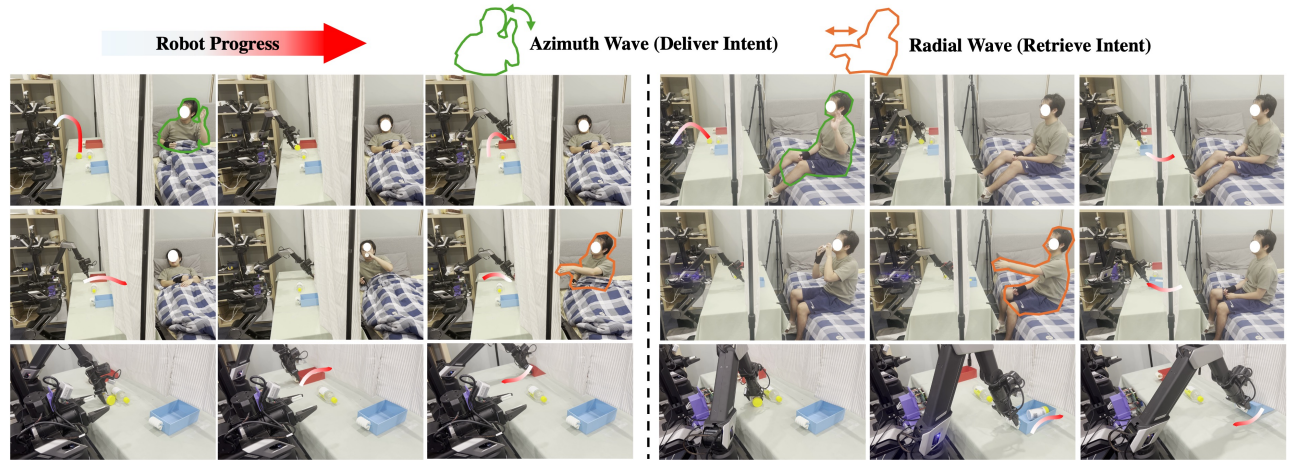


Fig. 5. Qualitative examples of the human states and corresponding robot operations. The columns show an azimuth wave, waiting, approaching, and a radial wave; the camera views show object grasping, tray delivery, and tray retrieval during the interaction.

are attributed to the VLA policy. The RGB-based method performs well under clear visibility but fails under privacy-preserving occlusion scenes. In contrast, the mmWave radar-based method remains robust under these occluded conditions. Its PA performance also approaches RGB’s under clear visibility. We observe that the radar method achieves slightly higher retrieval-activation PA. The radial waving can cause self-occlusion in RGB images, whereas radar captures hand motion patterns that are less affected by self-occlusion. For occlusion and cross-environment cases, the curtain motion may occasionally cause false activations and premature delivery/retrieval. The proposed DRP and MSSM modules aim to suppress these false positives and limit the performance degradation. For robot manipulation MA, occasional failures mainly occur when the relatively heavy water bottle loses balance and slips, or the delivery box collides with the table. The failure rate slightly increases in cross-environment trials due to changes in illumination. Since our mmHRI can be connected with any vision-language policies, the MA may improve with other alternatives. The radar-based method also achieves reliable collision detection, even when humans are blocked by privacy curtains, supporting safer HRI.

D. Robustness and Generalization

To evaluate generalization to unseen scenarios and assess deployment reliability, we conduct two robustness tests for the perception module: cross-subject and cross-table-clutter

TABLE III: Generalization to unseen subjects.

Subject	Delivery Activation	Retrieve Activation	Collision Avoidance
Seen Subject	28/30	29/30	30/30
Unseen Subject 1	30/30	30/30	30/30
Unseen Subject 2	27/30	26/30	30/30

TABLE IV: Generalization to unseen tabletop clutter.

Clutter	Delivery Activation	Retrieve Activation	Collision Avoidance
No Additional Clutter	28/30	29/30	30/30
Metal Box	28/30	26/30	30/30
Tall Cup Box	26/30	30/30	30/30

evaluation. As shown in Figure 6 and Tables III & IV, we introduce unseen clutter that interferes with the radar signal, including a metal box filled with daily objects and a large paper box that blocks the sensor’s line of sight. The success ratios are only slightly affected in both cases. We also evaluate two unseen subjects with different heights, body shapes, and waving habits to verify that the perception model does not overfit the participants in the training dataset; performance remains stable across these subject changes.

VI. CONCLUSION

This work presents mmHRI, a privacy-preserving human-robot interaction framework that integrates mmWave radar-based human perception with a vision-language-action policy. By combining motion-sensitive raw radar representations with the geometric information in radar point clouds, mmHRI jointly estimates human actions, 3D poses and locations, and converts the resulting human state into structured instructions for closed-loop robot control. The proposed DRP and MSSM mechanism further improve robustness to curtain motion and robot-arm interference. Experiments

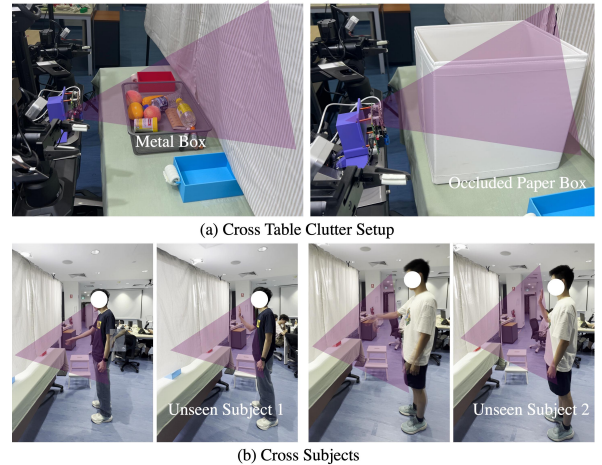


Fig. 6. Real-world robustness evaluation setups. (a) Cross-table-clutter settings with a metal box and an occluding paper box. (b) Cross-subject settings with two unseen subjects.

show that mmHRI achieves reliable action recognition: 85.09% accuracy under occlusion and 74.91% accuracy cross-environment. Real-world trials further demonstrate robust object delivery and retrieval across unseen subjects, clutter configurations, and environments. These results demonstrate the potential of mmWave radar as a complementary sensing modality for privacy-sensitive HRI.

Limitations. (1) The current system does not consider physical interaction with the curtain, such as opening curtains before delivery. Instead, it focuses on perceiving human intent through the curtain and satisfying human requests. Such interaction could be extended with more demo data. (2) The current system still cannot grasp objects using mmWave radar alone or handle multiple subjects. Future work could explore radar-only HRI or multi-subject human perception.

REFERENCES

- [1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *Int. J. Robot. Res.*, vol. 44, no. 10-11, pp. 1684–1704, 2025.
- [2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Open-VLA: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.
- [3] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, " $\pi_{0.5}$: A vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.
- [4] R. L. Birdwhistell, *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania press, 2010.
- [5] A. Baselizadeh, M. Z. Uddin, W. Khaksar, D. S. Lindblom, and J. Torresen, "Prima-care: Privacy-preserving multi-modal dataset for human activity recognition in care robots," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 233–237.
- [6] X. Liang, Z. Liu, K. Lin, E. Gu, R. Ye, T. Nguyen, C. Hsu, Z. Wu, X. Yang, C. S. Y. Cheung *et al.*, "Openrobocare: A multimodal multi-task expert demonstration dataset for robot caregiving," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 2661–2668.
- [7] A. Liu, N. K. Banerjee, and S. Banerjee, "Understanding worker perceptions toward privacy when working with collaborative robots that provide as-needed assistance," in *2025 11th International Conference on Automation, Robotics, and Applications (ICARA)*. IEEE, 2025, pp. 30–34.
- [8] A. Chen, X. Wang, S. Zhu, Y. Li, J. Chen, and Q. Ye, "mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3501–3510.
- [9] Y.-H. Ho, J.-H. Cheng, S. Y. Kuan, Z. Jiang, W. Chai, H.-W. Huang, C.-L. Lin, and J.-N. Hwang, "Rt-pose: A 4d radar tensor-based 3d human pose estimation and localization benchmark," in *European Conference on Computer Vision*. Springer, 2024, pp. 107–125.
- [10] J. Zhang, R. Xi, Y. He, Y. Sun, X. Guo, W. Wang, X. Na, Y. Liu, Z. Shi, and T. Gu, "A survey of mmwave-based human sensing: Technology, platforms and applications," *IEEE Communications Surveys & Tutorials*, 2023.
- [11] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su, "mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 269–282.
- [12] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 2019, pp. 51–56.
- [13] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "mid: Tracking and identifying people with millimeter wave radar," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2019, pp. 33–40.
- [14] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing," *arXiv preprint arXiv:2305.10345*, 2023.
- [15] J. Fan, J. Yang, Y. Xu, and L. Xie, "Diffusion model is a good pose estimator from 3d rf-vision," in *European Conference on Computer Vision*. Springer, 2024, pp. 1–18.
- [16] F. Ding, Z. Luo, P. Zhao, and C. X. Lu, "milliflow: Scene flow estimation on mmwave radar point cloud for human motion sensing," in *ECCV*. Springer, 2024, pp. 202–221.
- [17] Y. Sun, Z. Huang, H. Zhang, Z. Cao, and D. Xu, "3drimr: 3d reconstruction and imaging via mmwave radar based on deep learning," in *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*. IEEE, 2021, pp. 1–8.
- [18] M. Keller, K. Werling, S. Shin, S. Delp, S. Pujades, C. K. Liu, and M. J. Black, "From skin to skeleton: Towards biomechanically accurate 3D digital humans," *ACM TOG*, vol. 42, no. 6, pp. 1–12, 2023.
- [19] S. Kim, K. Yun, J. Park, and J. Y. Choi, "Skeleton-based action recognition of people handling objects," in *WACV*. IEEE, 2019, pp. 61–70.
- [20] B. Reily, F. Han, L. E. Parker, and H. Zhang, "Skeleton-based bio-inspired human activity prediction for real-time human-robot interaction," *Auton. Robots*, vol. 42, no. 6, pp. 1281–1298, 2018.
- [21] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-Pose: Enhancing YOLO for multi-person pose estimation using object keypoint similarity loss," in *CVPR*, 2022, pp. 2637–2646.
- [22] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019, pp. 12026–12035.
- [23] S. An and U. Y. Ogras, "Mars: mmwave-based assistive rehabilitation system for smart healthcare," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 20, no. 5s, pp. 1–22, 2021.
- [24] T. Gu, Z. Fang, Z. Yang, P. Hu, and P. Mohapatra, "Mmsense: Multi-person detection and identification via mmwave sensing," in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 2019, pp. 45–50.
- [25] Y. Hu, K. Zuo, B. Ma, S. Li, Z. Xia, F. Xu, and J. Yang, "Wave-man: mmwave-based room-scale human interaction perception for humanoid robots," *arXiv preprint arXiv:2601.07454*, 2026.
- [26] H. Xue, Q. Cao, C. Miao, Y. Ju, H. Hu, A. Zhang, and L. Su, "Towards generalized mmwave-based human pose estimation through signal augmentation," in *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 2023, pp. 1–15.
- [27] H. Guo, S. Wang, R. Ma, S. Jiang, Y. Ghasempour, O. Abari, B. Guo, and L. Qiu, "Omnivla: Physically-grounded multimodal vla with unified multi-sensor perception for robotic manipulation," *arXiv preprint arXiv:2511.01210*, 2025.
- [28] K. Strabala, M. K. Lee, A. D. Dragan, J. Forlizzi, S. S. Srinivasa, M. Cakmak, and V. Micelli, "Toward seamless human-robot handovers," *J. Human-Robot Interact.*, vol. 2, no. 1, pp. 112–132, 2013.
- [29] T. L. Phan and A. Cosgun, "Placing objects on table is preferred over direct handovers when users are occupied," *Sensors*, vol. 25, no. 7, p. 2140, 2025.
- [30] Y. S. Choi, T. Chen, A. Jain, C. Anderson, J. D. Glass, and C. C. Kemp, "Hand it over or set it down: A user study of object delivery with an assistive mobile manipulator," in *IEEE RO-MAN*. IEEE, 2009, pp. 736–743.
- [31] T. Fujii, J. H. Lee, and S. Okamoto, "Gesture recognition system for Human-Robot Interaction and its application to robotic service task," in *IMECS*, vol. 1, 2014.
- [32] R. Mon-Williams, G. Li, R. Long, W. Du, and C. G. Lucas, "Embodied large language models enable robots to complete complex tasks in unpredictable environments," *Nature Machine Intelligence*, vol. 7, no. 4, pp. 592–601, 2025.
- [33] J. Song, S. Jeong, B. Jeon, S. Kim, M. Seo, H. Son, and K. Lee, "HABIT: Human-aware behavior and interaction training dataset for robot manipulation," *arXiv preprint arXiv:2606.31682*, 2026.
- [34] K. Zuo, G. Li, B. Lyu, Y. Lu, B. Ma, S. Han, X. Zhou, X. Yuan, C. Zhou, J. Bai *et al.*, "Gaze2act: Gaze-conditioned vision-language-action policies for interactive robot manipulation," *arXiv preprint arXiv:2605.30282*, 2026.
- [35] P. Liu, G. Li, J. Fan, B. Ma, J. Jia, Y. Xiao, and J. Yang, "Give: Grounding human gestures in vision-language-action models," *arXiv preprint arXiv:2606.13435*, 2026.
- [36] C. Li, K. Xiong, Y. Xu, L. Qian, Y. Wang, and W. Zhu, "Gazevla:

Learning human intention for robotic manipulation,” *arXiv preprint arXiv:2604.22615*, 2026.

- [37] W. Wang, R. Li, Z. M. Diekel, Y. Chen, Z. Zhang, and Y. Jia, “Controlling object hand-over in human–robot collaboration via natural wearable sensing,” *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 1, pp. 59–71, 2018.
- [38] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, “Point transformer,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 259–16 268.