Dynamic Human-to-Robot Object Handover with VLM-based Intention Detection and Movement Primitives

Sebastian Rietsch, Lukas Ruf and Tamim Asfour

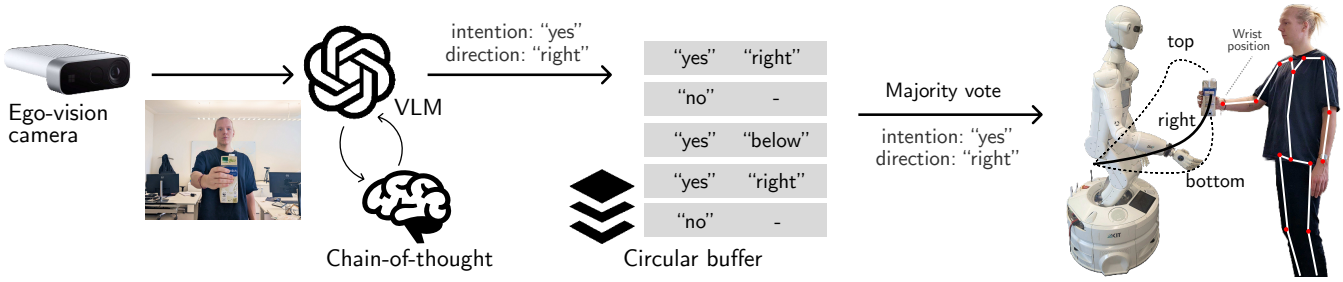

Fig. 1: Overview of the method pipeline. Starting from an ego-vision RGB image, the VLM is queried for handover intention and approach direction using a chain-of-thought prompt. Predictions are aggregated in a rolling buffer to filter noise, where a majority vote determines the final VMP selection.

***Abstract*— This work presents an initial exploration of using Vision-Language Models (VLMs) for dynamic Human-to-Robot (H2R) handovers, integrating VLM-based intention detection with Via-Point Movement Primitives (VMPs) for adaptive motion generation. By employing a structured chain-of-thought prompt and a majority vote over a prediction circular buffer, the system achieves 95.1% handover intention detection accuracy on the ARMAR-6 robot without task-specific training. Preliminary results suggest the approach can react dynamically to changing human behaviors and grasp strategies, though our evaluation reveals current challenges that must be addressed before practical deployment.**

I. Introduction

Human-Robot Interaction (HRI) is a central pillar of modern robotics research, particularly in collaborative environments where robots and humans work in close proximity. Within HRI, the handover of objects, specifically human-to-robot (H2R) handover, represents a complex challenge requiring precise timing, safety, and the ability to anticipate human cues and changes in behavior.

Current approaches to H2R handover typically fall into two categories: heuristic-based methods that rely on predefined rules and sensor thresholds [1], and learning-based approaches that train end-to-end models on large demonstration datasets [2]. While heuristic methods offer reliability, they lack flexibility when encountering novel objects or unconventional human behaviors. Learning-based approaches can achieve better generalization but require extensive data collection, which is costly and time-consuming.

The research leading to these results has received funding from the European Union's Horizon Europe programme under grant agreement No. 101070596 (euROBIN) and by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

The authors are with the Institute for Anthropomatics and Robotics, High Performance Humanoid Technologies Lab (H2T), at the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. {sebastian.rietsch, asfour}@kit.edu

Vision-Language Models (VLMs) have emerged as powerful tools for bridging visual perception and semantic reasoning, offering zero-shot generalization capabilities that extend far beyond traditional sensor-based heuristics. Recent work has demonstrated their potential for robot-to-human (R2H) handovers, leveraging semantic and geometric reasoning to determine suitable grasp regions [3], [4]. However, the specific application of VLMs for real-time intention detection within the dynamic constraints of H2R handovers remains underexplored.

In this work, we explore leveraging VLMs to detect handover intentions and estimate approach directions from RGB images without task-specific training. We combine this with Via-Point Movement Primitives (VMPs) [5], which offer smooth trajectory generation with real-time goal adaptation.

II. Method

As illustrated in Fig. 1, our method consists of two main components: a VLM-based perception module that infers human intention and grasp approach from ego-centric vision, and a VMP-based handover motion generation strategy.

A. Problem Formulation

We consider a bimanual humanoid robot that observes the environment through ego-centric RGB images and 3D human skeletal poses estimated via the Azure Kinect Body Tracking SDK [6]. We formulate dynamic H2R handover as three subproblems.

Intention detection classifies whether the human is actively offering an object, mapping observations to a discrete state $y \in \{\text{yes}, \text{no}\}$, where "yes" indicates a clear offering gesture and "no" encompasses all other states.

For *approach direction detection*, we discretize the space of possible grasp approaches into cardinal directions $\mathscr{D} =$

$\{\text{top}, \text{bottom}, \text{left}, \text{right}\}$. The goal is to infer the optimal direction $d^* \in \mathscr{D}$ that avoids collision with the human hand and, preferably, accounts for object-specific grasping constraints.

Finally, *trajectory adaptation* addresses the dynamic nature of human behavior. This includes aborting and returning to a home pose if intention changes from "yes" to "no", and modulating the trajectory in real-time to track the object position (approximated by the human wrist).

B. Intent and Grasp Direction Detection

To overcome the stochastic nature of VLMs, we propose a structured chain-of-thought [7] prompt that guides the VLM through a two-step reasoning pipeline.

1) Handover Intention Classification: The model first determines if a handover is occurring based on specific visual cues. A positive classification requires an "active offering gesture" (e.g., arm extension, gaze directed at the robot). The model is explicitly instructed to distinguish this from passive states like holding, carrying, or resting with an object.

To mitigate prediction noise, the VLM is queried continuously at a predefined rate. We store a buffer of past results and apply a majority voting scheme to determine the active intention for the current timestep.

2) Approach Direction Reasoning: If a handover is detected, the VLM infers the optimal robot approach direction through a single chain-of-thought prompt that guides it to: (1) trace the holding arm back to the shoulder to identify the anatomical hand (left vs. right), counteracting common VLM mirroring errors; (2) identify which parts of the object are occluded by the human hand; and (3) select an approach direction that avoids collision with the hand. Notably, the current prompt does not explicitly instruct the VLM to consider object-specific constraints such as fragility, which explains some failure cases observed in our evaluation.

C. Trajectory Adaptation and Grasping

We leverage VMPs defined over 6D end-effector poses for motion execution. A VMP generates a trajectory $y(x)$ parameterized by a canonical value $x \in [0,1]$, where $x = 1$ corresponds to the start pose and $x = 0$ to the goal pose. For each arm and approach direction, a distinct VMP is learned from multiple kinesthetic demonstrations.

To execute the reaching motion, we initialize the start pose $y(1)$ at the robot's current end-effector state. The goal position $p_{goal}^{(t)}$ is dynamically updated to the human's wrist position $p_{wrist}^{(t)} \in \mathbb{R}^3$, preserving the demonstrated goal orientation $R_{demo} \in SO(3)$, but applying a translational offset specific to the approach direction (e.g., an upward offset for a top grasp). We progress x from 1 to 0 over a predefined time frame and track the sampled poses $y(x)$ with an impedance controller for human safety.

Our approach addresses three aspects of dynamic H2R handovers:

1) Adaptation to a moving object: To avoid sudden jumps in the target pose when the human moves quickly, we smooth the goal position update using an exponential moving average (EMA): $p_{goal}^{(t)} = (1-\alpha)\, p_{goal}^{(t-1)} + \alpha\, p_{wrist}^{(t)}$, where α is a smoothing factor.

2) Switching approach direction: When the approach direction changes, we interpolate from the current end-effector pose onto the new VMP trajectory using EMA for position and Spherical Linear Interpolation (SLERP) for orientation. The new VMP retains the original start pose and adopts the canonical value from the previous VMP.

3) Intention change or switching arm: To retract the arm when intention switches from "yes" to "no", or when the executing arm changes, we reverse the canonical value progression, sampling poses from the VMP backward toward $x = 1$.

To finalize the grasp, we close the hand upon detecting a force spike in the wrist-mounted force-torque sensor, allowing the human to initiate the final transfer by gently pushing the object into the robot's hand.

III. Preliminary Results

We evaluated our approach on the ARMAR-6 robot. The evaluation was threefold: assessing detection accuracy, dynamic adaptability, and object generalization.

In standard handover scenarios with varying objects (cup, milk carton, bottle), the system achieved a 95.1% accuracy in detecting positive handover intentions and 98.4% accuracy in correctly distinguishing non-handover activities (e.g., drinking, placing on a table). The system correctly inferred the approach direction in 91.67% of cases, where correctness was determined by whether the selected direction avoided the human hand.

To evaluate adaptability, we introduced real-time perturbations. The robot successfully transitioned between grasps with a median reaction time of 4.55 seconds. However, the total interaction time averaged approximately 13 seconds, primarily bottlenecked by the VLM inference latency. Finally, tests on 10 diverse objects (e.g., cup, bottle, flower pot, hammer) demonstrated that the approach generalizes across different object geometries without task-specific training. However, the VLM did not consistently reason about object-specific constraints; without explicit prompting, it often selected directions that avoided the hand but were semantically unsafe for the object (e.g., grasping a flower from above).

IV. Conclusion

This work demonstrates that the combination of VLMs and VMPs offers a flexible framework for dynamic human-to-robot handovers. While VLMs prove effective at classifying human handover intentions, their outputs are not perfectly reliable, motivating our circular buffer and majority voting approach to filter prediction noise. Our evaluation also reveals that current VLM prompts require refinement to better account for object-specific grasping constraints (e.g., avoiding sharp edges, fragile parts), which were not consistently considered in our initial approach. Additionally, inference latencies remain a bottleneck for natural interaction. Future work will focus on systematic comparison with heuristic-based baselines to quantify the specific benefits of VLM-based reasoning.

REFERENCES

[1] T. Asfour, M. Waechter, L. Kaul, S. Rader, P. Weiner, S. Ottenhaus, R. Grimm, Y. Zhou, M. Grotz, and F. Paus, "Armar-6: A high-performance humanoid for human-robot collaboration in real-world scenarios," *IEEE Robotics & Automation Magazine*, vol. 26, no. 4, pp. 108–121, 2019.

[2] Z. Wang, J. Chen, Z. Chen, P. Xie, R. Chen, and L. Yi, "Genh2r: learning generalizable human-to-robot handover via scalable simulation demonstration and imitation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 362–16 372.

[3] J. Liu, W. Dong, J. Wang, and M. Q.-H. Meng, "Leveraging semantic and geometric information for zero-shot robot-to-human handover," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 16 340–16 346.

[4] A. Tulbure, R. Zurbrügg, T. Grigat, and M. Hutter, "Llm-handover: Exploiting llms for task-oriented robot-human handovers," *IEEE Robotics and Automation Letters*, 2025.

[5] Y. Zhou, J. Gao, and T. Asfour, "Learning via-point movement primitives with inter- and extrapolation capabilities," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4301–4308.

[6] Microsoft, "Azure Kinect Body Tracking SDK," 2021. [Online]. Available: https://learn.microsoft.com/en-us/azure/kinect-dk/body-sdk-download

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *CoRR*, vol. abs/2201.11903, 2022. [Online]. Available: https://arxiv.org/abs/2201.11903