

A Generative System for Robot-to-Human Handovers: from Intent Inference to Spatial Configuration Imagery

Hanxin Zhang¹, Abdulqader Dhafer¹, Zhou Daniel Hao^{1*} and Hongbiao Dong²

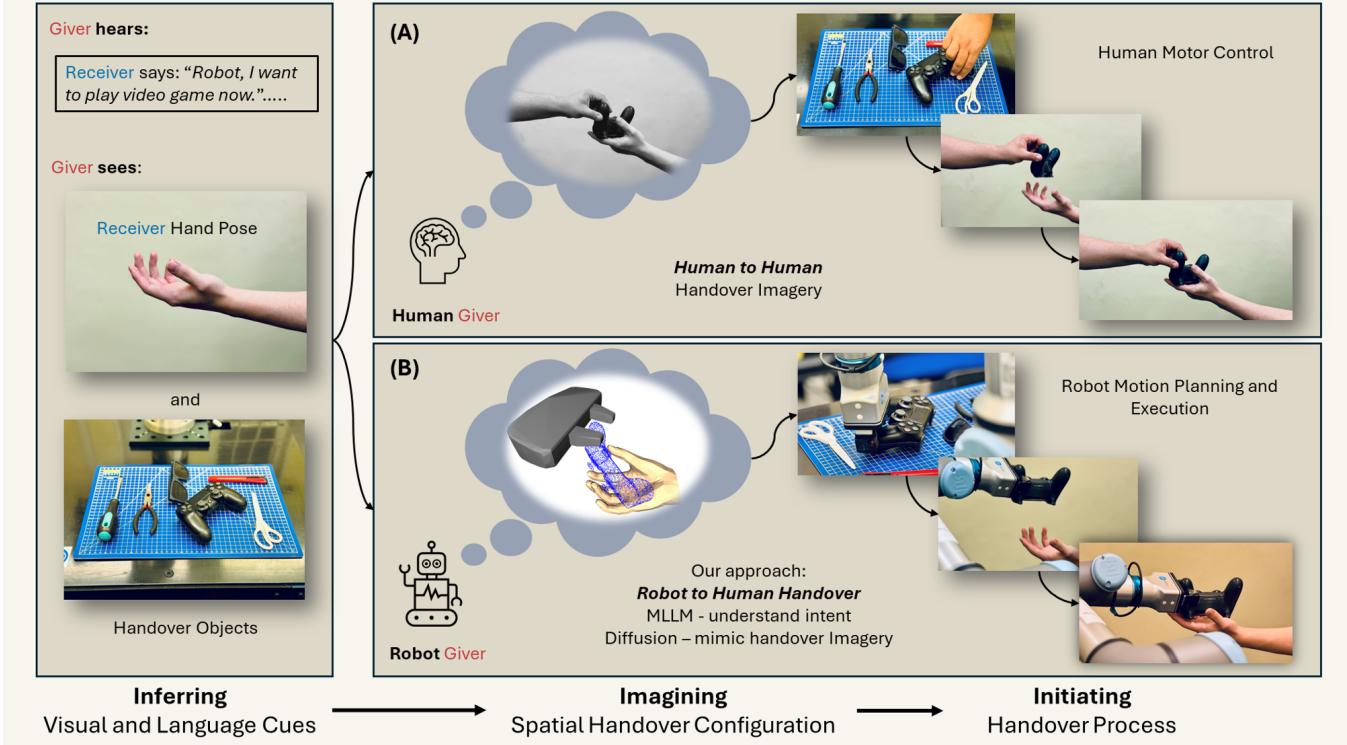


Fig. 1: During a handover, a human coworker effectively understands the receiver’s intent to pick up the appropriate tool, selects the correct grasping point, and envisions the final handover gesture, thereby delivering a satisfying transfer. Our approach emulates this process by using language-vision models and diffusion models to infer the receiver’s intent and mimic the motor imagery process, achieving a human-like, effective robot-to-human handover.

Abstract—We propose a novel system for robot-to-human object handover that emulates human coworker interactions. Unlike most existing studies that focus primarily on grasping strategies and motion planning, our system focus on 1) inferring human handover intents, 2) imagining spatial handover configuration. The first one integrates multimodal perception—combining visual and verbal cues—to infer human intent. The second one using a diffusion-based model to generate the handover configuration, involving the spacial relationship among robot’s gripper, the object, and the human hand, thereby mimicking the cognitive process of motor imagery. Experimental results demonstrate that our approach effectively interprets human cues and achieves fluent, human-like handovers, offering a promising solution for collaborative robotics. Code, videos,

and data are available at: <https://i3handover.github.io>.

I. INTRODUCTION

Object handover between humans and robots is a critical application in human–robot interaction that demands seamless collaboration. In robotic handovers, one agent (the giver) must transfer an object to another (the receiver) through a coordinated process.

In natural human-to-human handovers as shown in Fig. 1(A), the giver interprets complex multimodal cues—such as the receiver’s extended hand, the spatial arrangement of nearby objects, and verbal instructions—to infer the appropriate moment and manner for transferring an object. Cognitive and neuroscientific studies have shown that humans routinely plan their actions by mentally simulating the spatial handover configuration [1], [2], also known as motor imagery. That is, they anticipate how the object will be handed over, considering the spatial relationship of the

¹H. Zhang, A. Dhafer, and Z. D. Hao are with DANiLab, School of Computing and Mathematical Sciences, University of Leicester, Leicester, UK (corresponding to hz273, d.hao@leicester.ac.uk)

²H. Dong is with School of Engineering, University of Leicester, Leicester, UK (hd38@leicester.ac.uk)

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

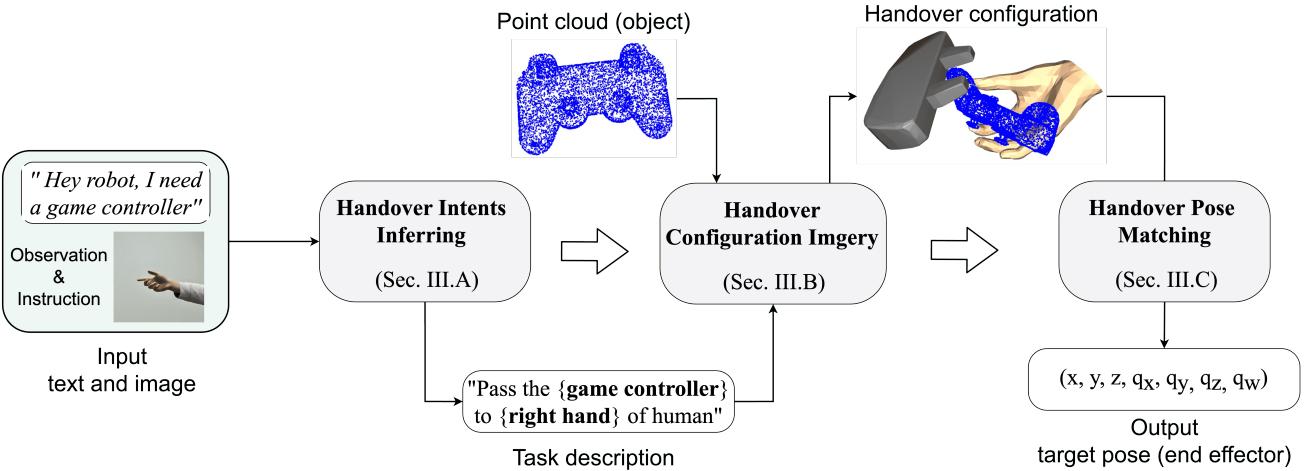


Fig. 2: Our approach consists two stages: handover intents inferring and handover configuration imagery. The user conveys handover intent through text input and a receiving hand image. At first stage, robot infers a task description (e.g., “Pass the game controller to the right hand”) as an input to the next stage. With the point cloud of object in the task description, the robot generates a receiving hand pose, and estimates several potential grasping angles for gripper, then determines the optimal handover pose.

giver’s hand, the object, and the receiver’s hand.

Emulating such human-like behaviour in robot-to-human handovers could enable robots to more effectively interpret multimodal cues and generate appropriate grasp configurations. Such capability is vital for deploying collaborative robots (co-bots) in dynamic environments such as factories, laboratories, or any setting where humans and robots work side by side. If robots could mimic the human strategy of motor imagery based on visual and auditory input, they would achieve a higher level of natural and safe object exchange. Our research aims to achieve this, as shown in Fig. 1(B).

However, replicating human handover processes in robotic systems presents several complex challenges. Firstly, robots struggle with comprehending complex tasks, requiring more advanced reasoning and perception capabilities. In handover, human intents contain not only explicit expressions (“I need a game controller.”), but also includes implicit expressions (“I want to play games now.”). Likewise, the human receiving hand serves as a crucial intent signal. Secondly, generating a suitable spatial handover configuration remains a challenging task. This study focuses on how to generate target positions about end-effector, object and human hand through text prompt. To overcome the challenges above, we propose to decompose the handover task into two sub-tasks: handover intents recognition and handover configuration generation. The texts of handover gesture and object name are predicted at first, then the target positions and handover poses are generated based on the text prompt, then robotic arm is guided to complete the motion planning.

We use multi-modal large language models (MLLMs) for handover intents recognition. MLLMs exhibit advanced capabilities, enabling them to integrate gesture embeddings and linguistic descriptions in a more semantically coherent and contextually appropriate way. MLLMs reason about un-

ambiguous or implicit language and hand images to generate explicitly templated output texts (“Pass the game controller to left hand of human”).

We integrate a diffusion-based method to generate handover poses, considering the spatial relationships among the human hand, robotic gripper, and object simultaneously. The idea is inspired by the article [3] in cognitive science, which emphasizes that humans mentally simulate actions before execution. Generative diffusion models use stepwise de-noising to accurately capture complex geometries making them ideal for point-cloud and mesh generation. Their incremental approach also enables flexible conditional generation by integrating conditional information at each step.

Key contributions of our work are as follows:

- We develop a handover intent recognition method using MLLMs to interpret both verbal instructions and visual gestural cues. This is, to our knowledge, one of the few studies exploring an LLM-based approach to infer human intent in object handover scenarios.
- We introduce a diffusion-based generative model for 3D handover configuration, conditioned on text prompts derived from the recognised intent. The advantage of this approach is to generate the spatial relationship among the robotic gripper, the object, and the human hand.
- We propose a matching approach to align the generated hand with the real-world hand, preventing contact between the hand and the robotic gripper to ensure the safe handover.

II. RELATED WORKS

Robot-to-human, human-to-robot, and robot-to-robot handovers are the three primary research directions identified in a recent survey [4]. Our study focuses on robot-to-human handover, where the robot acts as the giver. The

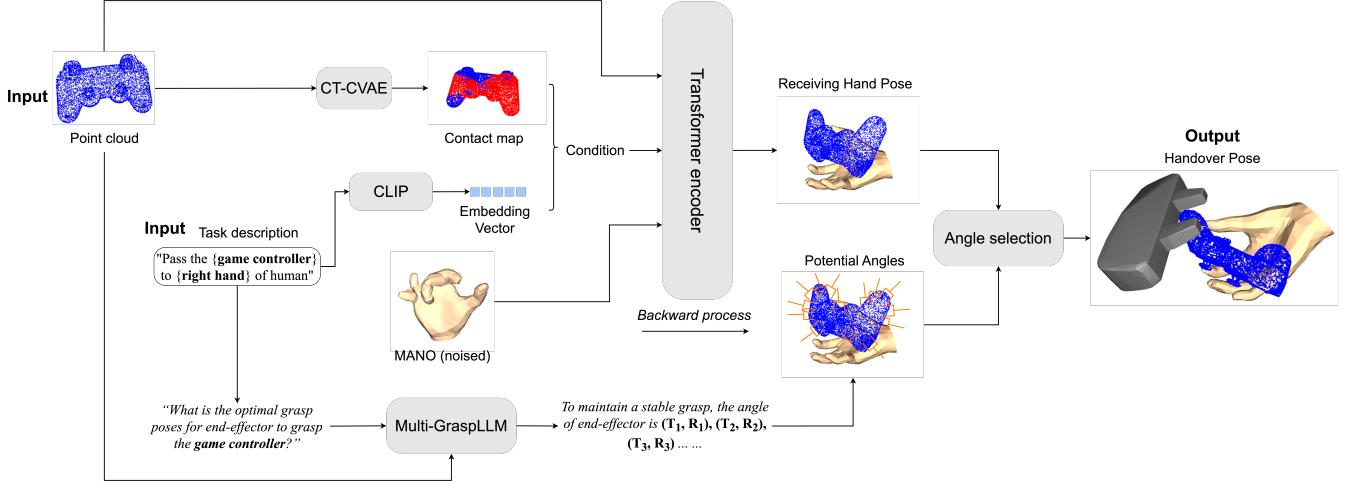


Fig. 3: Flowchart of generating handover configuration, integrating a text-guided diffusion model and an LLM. Both models take 3D object point cloud and a task description as inputs. The CT-CVAE generates the contact map, while CLIP encodes the textual input into an embedding vector, providing conditional guidance for the diffusion model to predict the receiving hand pose. Simultaneously, Multi-GraspLLM generates multiple candidate grasp angles, which are then refined by the angle selection strategy to determine the optimal one. The final output is the most suitable spatial handover configuration among the object, receiving hand, and robotic gripper.

primary objective is to enable the robot to generate a feasible handover configuration based on the recognized handover intent.

Multi-modal Large Language Models (MLLMs) represent a state-of-the-art approach for understanding human handover intent by integrating language and vision. In human-robot handover, there are emerging studies [5], [6] that have attempted to use MLLMs as a brain of robots to understand complex environments and human intentions. These following works have made great progress in grasp detection. FLarG [7] is a language-guided grasping method based on CLIP [8] that provides object positions for robots during handover. However, it only considers language input, ignoring other modalities. The GraspCLIP [9] is a task-oriented grasping prediction method that combines visual and linguistic information for object localization and robotic grasping configuration prediction. RefTR [10] combines BERT and ResNet for cross-modal fusion of vision and instructions, focusing on improving success rate of robotic grasping. Although existing MLLM-based studies consider object-related language and visual information, they overlook the receiving hand, which is crucial for understanding handover intent. Our work is inspired by [11], integrating hand and language as inputs to infer which item a human needs and which hand they intend to receive it with.

Generating a feasible 3D handover configuration requires accounting for the spatial relationship among the hand, object, and robotic gripper, including their 6D poses. Diffusion models can generate 3D objects from task descriptions, enabling the creation of grasp configurations for robot handovers. However, existing studies generate human grasp poses and robotic gripper poses separately, which may result in spatial conflicts or excessive distance, making handover impractical and posing safety risks in real-world scenarios.

In recent years, these datasets [12], [13], [14], [15], [16] have provided human grab poses and gripper 6D configurations for object grasping, serving as essential resources for handover study. [17] proposed a contact map constraint-based grasping generation method for guided generation of human grasp poses from contact maps. DiffH2O [18] is a framework based on diffusion models to generate accurate hand-object interaction sequences from textual descriptions and object geometric information. [19] proposed GeneOH-Diffusion, a three-stage denoising framework, which unites the features of space, time, and hand-object relationship to improve the generation quality. In addition to generate grasp gestures of human, another direction is to generate grasp poses of robotic gripper on objects. GraspLDM [20] is a diffusion-based 6 DOF grasping generation framework, which improves the grasping quality by introducing a diffusion model in the variational auto-encoder potential space. [21] presents a handover system named ContactHandover, which optimizes the robotic grasping strategy by human contact point prediction to determine the optimal delivery position and orientation. Existing studies primarily focus on generation of grasping poses of the robotic gripper and the human hand. We emphasize generating the spatial handover configuration rather than individual poses.

III. METHODOLOGY

Our method is structured into three key stages: handover intents inferring, handover configuration imagery, and handover pose matching. The robot first interprets the human handover intent, infers a task description then it generates the corresponding configuration, which is subsequently aligned with the actual receiving hand. The framework is illustrated in Fig. 2. We implemented our handover system on a robotic arm using the proposed method, incorporating insights from

prior works [4], [22], [23]. This system is utilized to drive the real robotic arm to complete the object handover.

A. Handover intents inferring

For handover intent understanding, we choose Gemma 2 [24] as our robot’s ‘brain’; its inputs comprise two modalities: vision and language. Before feeding them into MLLMs, we preprocess each modality separately. For language input, we employ Whisper [25], a speech recognition model, for real-time speech-to-text conversion, using “Hey robot” as the activation trigger. The converted text, denoted as T , serves as our input. Once the voice signal is recognized, an RGB-D camera is triggered to capture an image of the receiving hand, represented as $I \in \mathbb{R}^{H \times W \times G}$. The processed text-image pair (T, I) is then passed into the MLLMs, which infer the item name and the type of receiving hand.

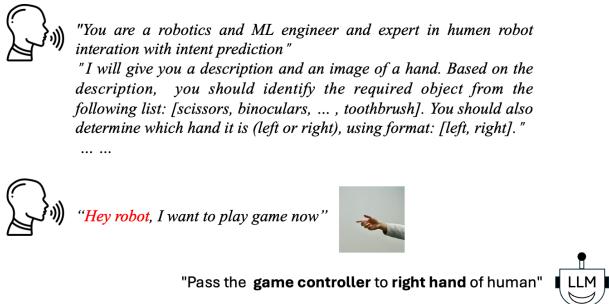


Fig. 4: MLLM interpretation of intent. The LLM processes textual and visual inputs through a structured prompt template to generate a task description.

We also designed a structured dialogue template, as shown in Fig. 4. The model receives a standardized textual description and an image of the receiving hand as input. It will recognize the tool name and hand type, then generates a templated output, such as “*Pass the game controller to right hand of human*”. To ensure consistency and clarity, we format the input as a system-user dialogue, explicitly guiding the model to generate a structured response strictly following the template, containing only the tool name and hand type.

B. Handover configuration imagery

Our approach consists of two generative modules. First, we use the text-guided diffusion model named Text2HOI [26] to generate receiving hand poses from task description T and the 3D point cloud of object x_{obj} . Second, for generating the gripper’s grasping angles, we use Multi-GraspLLM [16], which can predict potential grasp angles based on the same inputs. The final output of our system is the imagined handover configuration. The overall pipeline is illustrated in Fig. 3.

We defined the imagined hand configuration for receiving object as $\{x_{lhand}, x_{rhand}, x_{obj}\}$, where x_{lhand} and x_{rhand} represent the left and right hands, respectively. Meanwhile, x_{obj} denotes the 3D point cloud vertices of the object. We use

MANO [27] as our hand model. Its surface mesh can be fully deformed and articulated using a standard linear blend skinning function, as given by:

$$\mathcal{M}(\theta, \beta) = \mathcal{W}(\mathcal{T}(\beta, \theta), \mathcal{J}(\beta), \theta, \omega)$$

where $\mathcal{M}(\theta, \beta) \in \mathbb{R}^{773 \times 3}$ is hand mesh surface, $\beta \in \mathbb{R}^{10}$ and $\theta \in \mathbb{R}^{16 \times 3}$ are shape parameters and pose parameters, ω is a constant

The representation of the hand for the diffusion model can be expressed as: $x_{lhand} = \{t_l, \theta_l\}$ and $x_{rhand} = \{t_r, \theta_r\}$. In this symbol, $t_h, t_h \in \mathbb{R}^3$ stand for 3D translation parameters, representing the position of the hand. And $\theta_l, \theta_r \in \mathbb{R}^{16 \times 6}$ stand for the hand pose parameters, encoded in 6D rotation representation. For visualization, the meshes of left and right hand are generated from x_{lhand} and x_{rhand} by feeding them to the MANO layer to output the hand vertices $\mathbf{V}_{lhand}, \mathbf{V}_{rhand} \in \mathbb{R}^{V \times 3}$, and hand joints $\mathbf{J}_{lhand}, \mathbf{J}_{rhand} \in \mathbb{R}^{J \times 3}$ in global space, where $V = 778$ and $J = 21$.

Text2HOI primarily consists of a transformer f^{THOI} , which is used to denoise the MANO hand representation. The remaining modules include CT-CVAE, which generates a contact map based on the object’s point cloud, and CLIP, which transforms the task description into an embedding vector. These two components serve as conditional inputs to the transformer, guiding it in generating the pose based on the task description. The final output of the Text2HOI is the receiving hand pose \hat{x}_{hand} . Multi-GraspLLM [16] incorporates a point cloud encoder, which processes the object point cloud into an embedding vector. The object name from the task description T is used as part of the prompt template, such as “*How can 2-fingers parallel gripper grasp the game-controller?*”. The LLM then generates the grasping pose parameter as the homogeneous transformation matrix $\mathbf{H} \in \mathbb{R}^{4 \times 4}$, which encodes both position and orientation of the end-effector. It consists of the rotation matrix $\mathbf{r} \in \mathbb{R}^{3 \times 3}$ and the translation vector $\mathbf{t} \in \mathbb{R}^3$.

In order to avoid gripper contact with the human hand, we designed a strategy for selecting the optimal grasp angle. The gripper should ideally be positioned on the opposite side of the object relative to the hand. Therefore, we define the hand’s direction vector from wrist to the tip of middle finger, which can be computed using key points from the MANO. The gripper’s direction vector \mathbf{v}_g and the hand’s direction vector \mathbf{v}_h are aligned to achieve an angle close to 180 degrees, while the distance between their centre points \mathbf{p}_g and \mathbf{p}_h is maximized. The optimization objective f is formulated in the following equation:

$$\min f = \sum_{i=1}^n \left| \frac{\mathbf{v}_g \cdot \mathbf{v}_h}{\|\mathbf{v}_g\| \|\mathbf{v}_h\|} \right| - |\mathbf{p}_g - \mathbf{p}_h|$$

where n represents the number of potential grasp angles. The first term denotes the angle between their vectors, while the second term measures the distance between them.

C. Handover pose matching

The generated handover configuration must match the actual receiving hand to locate the end-effector in the world

frame, as shown in Fig. 5. We first match the hand and then compute the robotic gripper's 6D pose in the world coordinate frame. The geometric centre of the hand c serves as the coordinate origin and is computed from the vertices of the point cloud. The hand's direction vector \vec{d} defines the positive x-axis, while the palm plane is formed by the x-axis and y-axis. The palm normal vector \vec{p} points outward from the palm and defines the positive z-axis. The imagined hand configuration is represented by $(c_1, \vec{d}_1, \vec{p}_1)$, while the real hand pose is denoted as $(c_2, \vec{d}_2, \vec{p}_2)$.

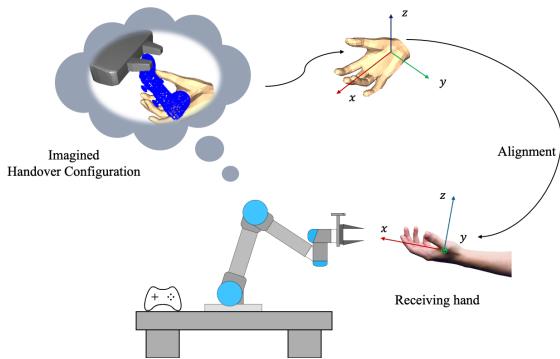


Fig. 5: Hand coordinate matching. The robot generates an imagined handover configuration. This imagined receiving hand is matched with the actual receiving hand to ensure proper spatial consistency.

The matching process ensures proper alignment by enforcing three key conditions. First, the coordinate origins of the them must coincide. Second, the direction of imagined hand must be aligned with the real hand, ensuring that both vectors are normalized to unit length and their dot product is maximized. Third, the angle between palm normal vectors must be minimized. The homogeneous transformation matrix H is constructed as

$$H = \begin{bmatrix} R_2 R_1^\top & c_2 - R_2 R_1^\top c_1 \\ 0^\top & 1 \end{bmatrix}$$

where the rotation matrix $R_1, R_2 \in SO(3)$ are constructed using the direction and normal vectors of the imagined and real hand poses, respectively, and are defined as:

$$\begin{aligned} R_1 &= [a_1 \quad (p_1 \times a_1) \quad p_1], \\ R_2 &= [a_2 \quad (p_2 \times a_2) \quad p_2] \end{aligned}$$

The spatial relationship between the end-effector and hand is directly generated through our method. The end-effector's final position \hat{p} and quaternion \hat{q} can be determined by transforming its initial position p_0 and quaternion q_0 using H as given by the following equation

$$\begin{aligned} \hat{p} &= R_2 R_1^\top p_0 + (c_2 - R_2 R_1^\top c_1), \\ \hat{q} &= q_0 \otimes q_{R_2 R_1^\top}. \end{aligned}$$

where \otimes represents quaternion multiplication. The term $q_{R_2 R_1^\top}$ represents the quaternion equivalent of the rotation matrix $R_2 R_1^\top$.

IV. IMPLEMENTATION AND EVALUATION

The experiments used 6 degrees-of-freedom UR5e robotic arm, with a maximum range of motion of 850 mm, and a maximum payload of 5 kg. The end-effector is equipped with a two-finger parallel gripper 2FG7. The gripper can pick up objects with a maximum width of 74 mm. Isaac Sim is used as simulation environment. An Intel RealSense D435i RGB-D cameras were installed directly above the robotic arm to obtain the top and side view. The experiments run on a desktop computer with AMD Ryzen 9 7900X3D 12-Core CPU, 32 GB system memory and one NVIDIA GeForce RTX 4090. The communication between the computer and robotic arm is done through ROS2 humble.

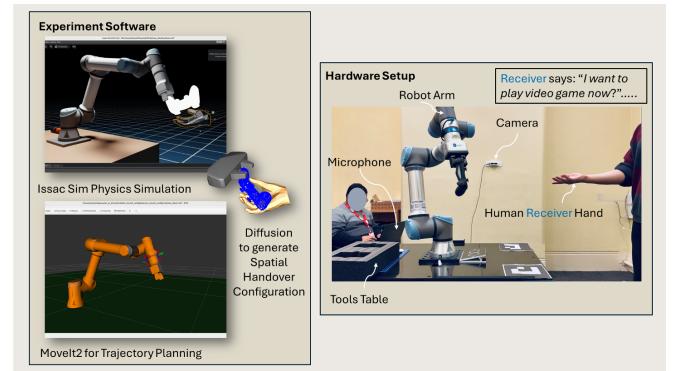


Fig. 6: Experimental setup. The tests were conducted in both simulation and the real hardware.

A. Inferring test of intention

Three fuzzy levels of handover intent for evaluating MLLMs' ability in interpreting from both language and images. We selected eight mainstream open-source models, including LLaMA, Gemma 2, and others. The evaluation utilized 30 hand images with diverse receiving poses and 16 objects from Multi-GraspSet [16]. For each object, 30 task descriptions were generated across three levels of ambiguity. The model is required to select the appropriate tool from the given options. For the evaluation criteria, the model is considered to have passed the test only if its output includes both the correct hand type ('left', 'right') and the most appropriate object name (like 'screwdriver') from the given set of tools.

We defined the angles for the receiving hand based on the three degrees of freedom of the human wrist [28]. For an unaffected human wrist, the maximal ranges of each degrees of freedom fall within the bounds of $76^\circ/85^\circ$, $75^\circ/75^\circ$, and $20^\circ/45^\circ$ for pronation/supination, flexion/extension, and radial/ulnar deviation, respectively. In addition, hand pose for grasping an object can be classified into three types based on grasping stability [29]: open, precise grip, and power grip. Fig. 8 displays all receiving hand poses, but we selected only 9 to account for duplicates.

As shown in Fig. 7, we use the average accuracy to evaluate the model. A successful pass is considered only if both the object name and hand type are correctly identified.

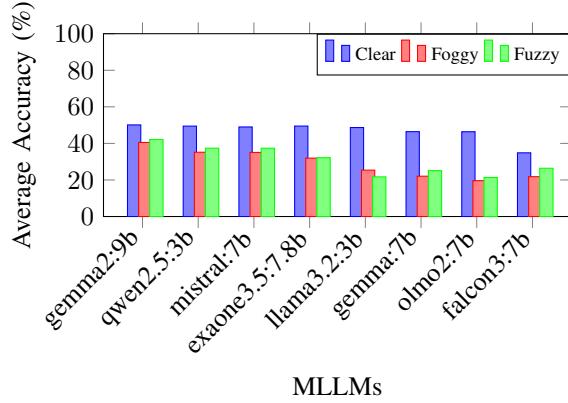


Fig. 7: Accuracy of inference across different MLLMs (ordered by highest average). The diagram presents the evaluation based on three classifications: (1) **Clear** represents direct descriptions and naming of the object. (2) **Foggy** descriptions focus on describing the task, the object, and similar phrases, such as "screw" and "screwdriver." (3) **Fuzzy** tasks primarily describe the task itself and provide a general, simplified description of the scenario.

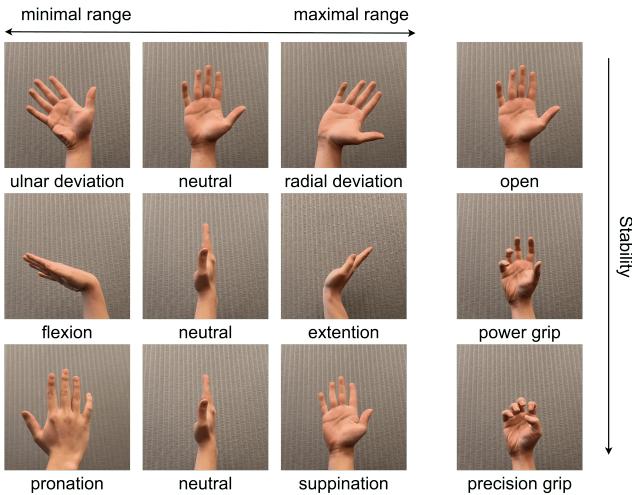


Fig. 8: Receiving hand pose. The receiving hand can be classified based on wrist joint degrees of freedom and grasp stability. The nine images on the left illustrate various wrist flexion levels, progressing from minimal to maximal bending. On the right, the three images represent different grasping styles, arranged from least to most stable.

The results indicate that Gemma2 performed best in fuzzy test, with accuracy of 44.24%. However, it also show that the parameter size of model does not necessarily yield better performance. Qwen2.5 achieved a high accuracy of 40.67%, despite using only one-third of the parameters of Gemma2. Clear level of task descriptions resulted in the highest average accuracy, while the foggy level was lower than at the fuzzy level. It shows that even without knowing the object name, MLLMs can infer the required item based on the scene and its intended use. A complete scene description better facilitates the inference, especially when no explicit object

name is provided as input.

Although, MLLMs exhibit inference ability, the overall accuracy remains low. Most failures due to two reasons. First, the side views of the hand significantly affected recognition, which caused most failures. Second, when multiple tools could be used for the same scene, leading the model to make incorrect judgments.

B. Visualization of robotic imagery

We applied the proposed generation method to 16 objects, all of which are daily household items. These objects vary in shape, size, and functionality, including tools, kitchenware, and electronic devices, ensuring a diverse evaluation of our approach.

The visualization of generated handover configuration is shown in Fig. 9. Generated receiving hands appear natural for most objects, aligning well with real receiving poses of human. In all results, robotic gripper rarely makes contact with the human hand, ensuring a safer handover. These generated configuration can provide reliable guidance for the robotic arm by establishing a well-aligned spatial relationship between the human hand, gripper, and object.

C. Execution testing

We built the handover system on UR5e robotic arm and mapped it to a simulation environment, as shown in Fig. 6. Five objects were selected for handover, and our approach was validated in both virtual and real environments. 5 objects with known point clouds are chosen for robotic arm to transfer to the human hand. The testing evaluates whether the robotic arm can accurately deliver the object to the human hand by following the generated configuration.

TABLE I: Success rate and time of 30 experiments per objects.

Object	Success Rate	Total Time (s)	imagination Time (s)	Execution Time (s)
scissors	0.80	10.21	3.05	7.16
game controller	0.87	9.26	3.06	6.20
pincer	0.90	9.78	3.29	6.49
knife	0.93	9.43	3.17	6.26
eyeglasses	0.80	9.53	2.97	6.56

For object handover, the robot execution success rate and time are shown in the table I. As can be seen from the data distribution in the table, the success rate of our system does not depend on object. The time spent on imagination accounts for approximately one-third of the total time. More detail of video is available on <https://i3handover.github.io>.

Failure cases reveal that when the receiving hand is positioned at an unfavourable angle, it can cause kinematic singularities or suboptimal planning. There are some engineering challenges in our study, including issues in gripper actuation and incomplete ROS2 packages. In the future, we aim to address these problems.

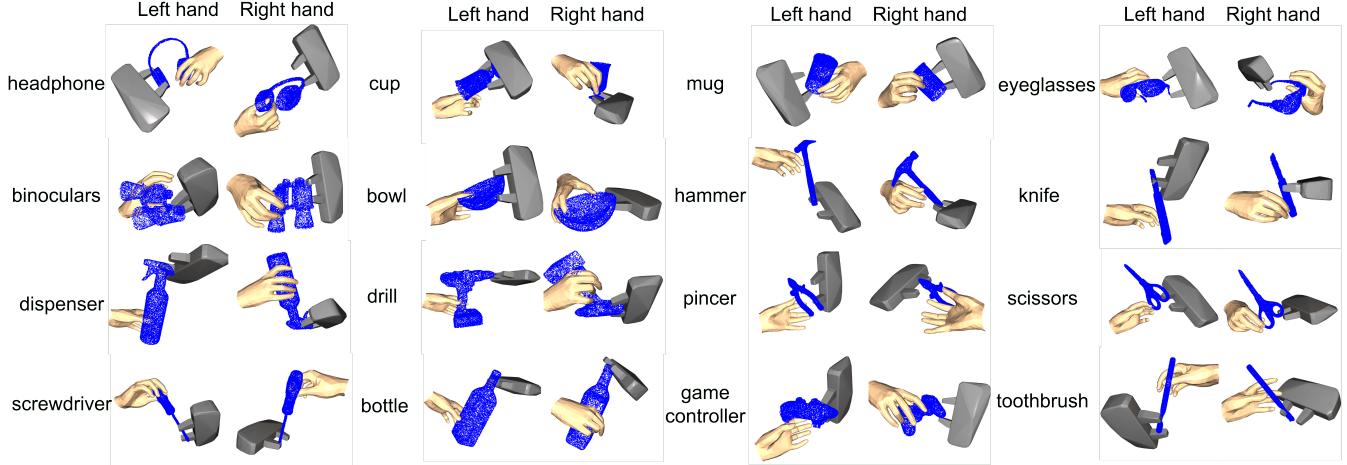


Fig. 9: Handover configuration visualization. Sixteen daily objects with their generated handover configurations, including both left and right receiving hands.

V. LIMITATION AND FUTURE DIRECTION

Although we propose a novel and promising system, certain limitations should be acknowledged and addressed in future work.

- We consider only the primary intent, without accounting for subtle or implicit human cues such as eye gaze, micro-expressions or muscle tension. These cues are also crucial indicators of human intent.
- Our system does not understand tool functionality or how different regions of object affect human during handovers. This ability is also essential for ensuring safe and reliable handovers.
- Our system assumes a static receiving pose and does not account for dynamic adjustments, such as the receiver switching hands or changing their gesture during the handover. Addressing these scenarios would require real-time adaptation, enabling the robot to reorient the object accordingly.

VI. CONCLUSIONS

We propose a novel robot-to-human object handover system that simulates human collaborative interactions. This system integrates multimodal large language models and diffusion models to infer human handover intent, and generate reasonable handover configurations. Experimental results demonstrate that our system is able to interpret handover intent at varying levels of ambiguity. Additionally, the generated configurations maintain a natural and spatially coherent relationship. Our approach demonstrates the ability to understand complex human cues and facilitates a safe and fluid handover process. This work presents promising solution for human-robot collaboration. In the future, we aim to improve execution stability, safety, and task generalization within the system.

REFERENCES

- [1] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, “An internal model for sensorimotor integration,” *Science*, vol. 269, no. 5232, pp. 1880–1882, 1995.
- [2] G. Rizzolatti and C. Sinigaglia, “The mirror mechanism: a basic principle of brain function,” *Nature Reviews Neuroscience*, vol. 17, pp. 757–765, 10 2016.
- [3] M. Jeannerod, “The representing brain: Neural correlates of motor intention and imagery,” *Behavioral and Brain sciences*, vol. 17, no. 2, pp. 187–202, 1994.
- [4] H. Duan, Y. Yang, D. Li, and P. Wang, “Human-robot object handover: Recent progress and future direction,” *Biomimetic Intelligence and Robotics*, p. 100145, 2024.
- [5] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, “Progrompt: Generating situated robot task plans using large language models,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530, IEEE, 2023.
- [6] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, “Code as policies: Language model programs for embodied control,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500, IEEE, 2023.
- [7] Q. Sun, H. Lin, Y. Fu, Y. Fu, and X. Xue, “Language guided robotic grasping with fine-grained instructions,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1319–1326, IEEE, 2023.
- [8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.
- [9] C. Tang, D. Huang, L. Meng, W. Liu, and H. Zhang, “Task-oriented grasp prediction with visual-language inputs,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4881–4888, IEEE, 2023.
- [10] Y. Lu, Y. Fan, B. Deng, F. Liu, Y. Li, and S. Wang, “Vi-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 976–983, IEEE, 2023.
- [11] Y. Li, X. Chen, H. Zhao, J. Gong, G. Zhou, F. Rossano, and Y. Zhu, “Understanding embodied reference with touch-line transformer,” in *ICLR*, 2023.
- [12] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, “Grab: A dataset of whole-body human grasping of objects,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 581–600, Springer, 2020.
- [13] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays, “Contactpose: A dataset of grasps with object contact and hand pose,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pp. 361–378, Springer, 2020.
- [14] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, “H2o: Two hands manipulating objects for first person interaction recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10138–10148, 2021.

- [15] L. F. Casas, N. Khargonkar, B. Prabhakaran, and Y. Xiang, “Multi-grippergrasp: A dataset for robotic grasping from parallel jaw grippers to dexterous hands,” *arXiv preprint arXiv:2403.09841*, 2024.
- [16] H. Li, W. Mao, W. Deng, C. Meng, H. Fan, T. Wang, P. Tan, H. Wang, and X. Deng, “Multi-graspllm: A multimodal llm for multi-hand semantic guided grasp generation,” *arXiv preprint arXiv:2412.08468*, 2024.
- [17] H. Li, X. Lin, Y. Zhou, X. Li, Y. Huo, J. Chen, and Q. Ye, “Contact2grasp: 3d grasp synthesis via hand-object contact constraint,” *arXiv preprint arXiv:2210.09245*, 2022.
- [18] S. Christen, S. Hampali, F. Sener, E. Remelli, T. Hodan, E. Sauser, S. Ma, and B. Tekin, “Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions,” in *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.
- [19] X. Liu and L. Yi, “Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion,” *arXiv preprint arXiv:2402.14810*, 2024.
- [20] K. R. Barad, A. Orsula, A. Richard, J. Dentler, M. Olivares-Mendez, and C. Martinez, “Graspldm: Generative 6-dof grasp synthesis using latent diffusion models,” *IEEE Access*, 2024.
- [21] Z. Wang, Z. Liu, N. Ouporov, and S. Song, “Contacthandover: Contact-guided robot-to-human object handover,” *arXiv preprint arXiv:2404.01402*, 2024.
- [22] C. Meng, T. Zhang, and T. Iun Lam, “Fast and comfortable interactive robot-to-human object handover,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3701–3706, IEEE, 2022.
- [23] Z. Wang, Z. Liu, N. Ouporov, and S. Song, “Contacthandover: Contact-guided robot-to-human object handover,” *arXiv preprint arXiv:2404.01402*, 2024.
- [24] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Huszenot, T. Mesnard, B. Shahriari, A. Ramé, *et al.*, “Gemma 2: Improving open language models at a practical size,” *arXiv preprint arXiv:2408.00118*, 2024.
- [25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.
- [26] J. Cha, J. Kim, J. S. Yoon, and S. Baek, “Text2hoi: Text-guided 3d motion generation for hand-object interaction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585, 2024.
- [27] J. Romero, D. Tzionas, and M. J. Black, “Embodied hands: Modeling and capturing hands and bodies together,” *arXiv preprint arXiv:2201.02610*, 2022.
- [28] N. M. Bajaj, A. J. Spiers, and A. M. Dollar, “State of the art in prosthetic wrists: Commercial and research devices,” in *2015 IEEE International Conference on Rehabilitation Robotics (ICORR)*, pp. 331–338, IEEE, 2015.
- [29] M. R. Cutkosky and R. D. Howe, “Human grasp choice and robotic grasp analysis,” *Dextrous robot hands*, pp. 5–31, 1990.