RHINO: Learning Real-Time Humanoid-Human-Object Interaction from Human Demonstrations

Jingxiao Chen*, Xinyao Li*, Jiahang Cao*, Zhengbang Zhu, Wentao Dong, Minghuan Liu†, Ying Wen, Yong Yu, Liqing Zhang, Weinan Zhang‡

Shanghai Jiao Tong University

timemachine@sjtu.edu.cn, wnzhang@sjtu.edu.cn

Abstract

Humanoid robots have shown success in locomotion and manipulation. Despite these basic abilities, humanoids are still required to quickly understand human instructions and react based on human interaction signals to become valuable assistants in human daily life. Unfortunately, most existing works only focus on multi-stage interactions, treating each task separately, and neglecting real-time feedback. In this work, we aim to empower humanoid robots with real-time reaction abilities to achieve various tasks, allowing human to interrupt robots at any time, and making robots respond to humans immediately. To support such abilities, we propose a general humanoid-human-object interaction framework, named RHINO*, i.e., Real-time Humanoid-human Interaction and Object manipulation.* RHINO *provides a unified view of human intent prediction, interactive motion, instruction-based manipulation, and safety concerns, over multiple human signal modalities.* RHINO *is a hierarchical learning framework that enables humanoids to acquire interaction skills from human-human-object demonstration and teleoperation data, while generalizing across diverse human appearances. In particular, it decouples the interaction process into two levels: 1) a high-level planner inferring human intents from real-time human behaviors; and 2) a low-level controller achieving expressive interaction behaviors and object manipulation skills based on the predicted intents. We evaluate our framework with human studies and quantitative experiments on a real humanoid robot and demonstrate its effectiveness and robustness in various scenarios.*

*These authors contributed equally.
†Project leader.
‡Corresponding author.

1. Introduction

Humanoid robots are increasingly being explored to perform tasks in diverse environments [4, 17, 19]. Their human-like morphology provides a potential for acting with human-like dexterity, making them ideal for general-purpose daily-life human assistants. Considering how we as humans react to our friends, a practically helpful humanoid assistant should possess three fundamental capabilities: 1) skill proficiency, equipped with diverse and essential skills to achieve various tasks; 2) intent recognition, capable of discerning human intents, from either motion or language; and 3) instant feedback, able to respond in real-time with feasible actions.

Nonetheless, most studies on human-robot interaction only focus on only one or two of these aspects. For instance, a significant body of work on human-robot interaction focuses on object handover [32, 34], or interactive motion generation [7, 21, 22, 25, 29], lacking the ability to switch between different tasks in real-time. Some others focus on recognizing human intents [12–14, 24, 31], which simplify the diversity of reaction and treat the interaction as an alternated two-stage process. The robot cannot be interrupted once a task is in progress, and further human commands can only be executed after the completion of the robot's current task. Many recent works have attempted to combine the ability of general foundation models to enable robots to understand the complexity of human interactions [33, 38], but they often suffer from high latency and are not suitable for real-time interaction tasks. These limitations hinder robots from rapid interventions and robust, multi-step interactions in human-centered tasks. Therefore, a framework that masters human-robot interaction with real-time intent recognition and various skills is urgently needed to tackle the above challenges.

To achieve this goal, we propose RHINO, a hierarchical learning framework for Reactive Humanoid-human INteraction and Object Manipulation. RHINO decouples

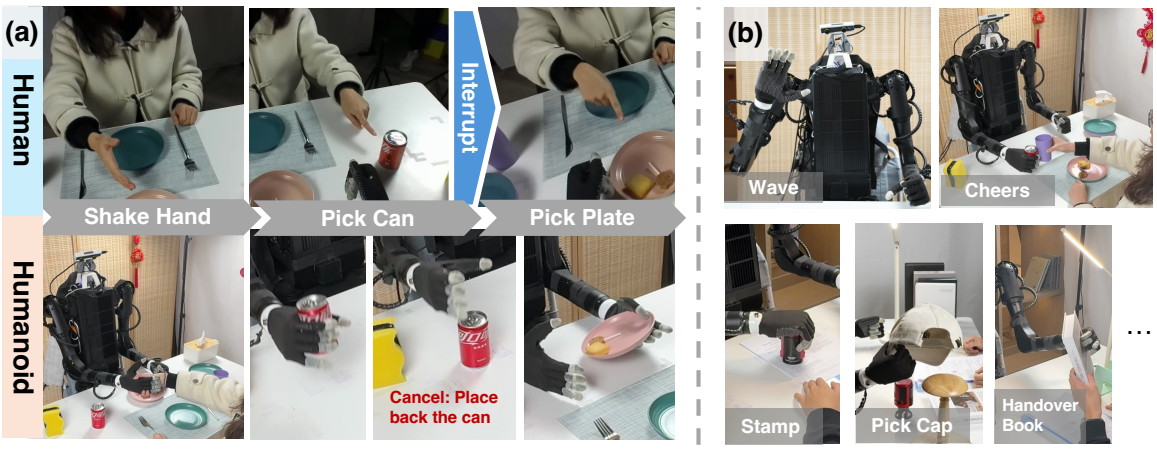


Figure 1. **RHINO has the capabilities of real-time interaction on diverse tasks.** (a) RHINO enables real-time humanoid-human-object interaction, allowing seamless task interruption and dynamic switching during operation. (b) The system demonstrates diverse capabilities, including waving, cheering, stamping, object pickup, handovers, and more.

the interaction process into two levels. High-level planner infers human intents from real-time human behaviors and plans to execute the corresponding skills. Low-level controller achieves reactive motion behaviors and object manipulation skills based on high-level signals. The high-level planner updates at high frequency, and the low-level controller is designed to be interruptible, enabling it to react to high-level commands at any time. The design of RHINO aims to ensure the scalability and robustness of the framework. To ensure scalability, we design a pipeline to learn interactions from human-object-human demonstrations, which can be easily extended to different tasks and scenarios. To ensure robustness across diverse human appearances, our approach decouples the human-centered interaction ability from the object-centered manipulation ability and designs different input modalities for each. We implement RHINO on a real humanoid robot and demonstrate its effectiveness and robustness in various scenarios (see Figure 1) with human studies and quantitative experiments. Although this work only focuses on the upper body of a humanoid, it has the potential to be extended to whole-body interaction with a unified humanoid controller.

Our main contributions are in the following aspects:

- We introduce the first real-time humanoid interaction framework capable of learning from human demonstrations, enabling dynamic task-switching and immediate responses to human instructions.
- We design a pipeline with two key innovations: 1) decoupling human-centered interaction from object-centered manipulation for *robustness* to diverse human appearances, and 2) learning directly from human-human demonstrations, which enables *low-cost scaling* to new tasks.
- We implement and validate RHINO on the Unitree H1 humanoid robot and demonstrate its effectiveness and robustness in 2 scenarios with over 20 tasks, and open-source the code and datasets to facilitate future research.

2. Related Works

Recent progress in building a human assistant robot can be divided into three categories: 1) human intent prediction, 2) basic skills, and 3) unified interaction framework. Table 1 compares RHINO and representative related works.

2.1. Human Intent Prediction

Humanoid robots need to estimate the human physical and mental states to provide appropriate assistance [35]. More specifically, many signals can be used to infer human intents, such as whole-body motion [34, 40], forces [4], gaze [12, 31], and language [33]. Object information in the environment also plays an important role in predicting human intent [14, 24] by combining it with human motion. In most works on human intent recognition, the robot first predicts the human intent and then executes the task. This two-stage design neglects the real-time reaction ability of the robot. Our work aims to react to human signals in real time, with interruptible downstream skills.

2.2. Robot Skills

Interactive motion synthesis. In human-robot interaction (HRI), learning to generate interactive and expressive motions, such as shaking hands and waving, are fundamental skills. Recent works [21, 39] collect multi-human motion data, capturing real-time interaction and reaction between humans. The human-like morphology of humanoid robots provides a unique opportunity to learn natural motion from retargeted human motion data [15]. Human motion data can be collected from motion capture systems or network videos. Compared to collecting robot motion data, it has a

Table 1. Comparison of RHINO with Prior Works in Human-Robot Interaction. Selected representative works are shown for each category to ensure clarity.

Representative Work	Category	Data Source	Manipulation Ability	Interaction Ability
EgoPAT3Dv2 [14]; HOI4ABOT [24]	Intent Prediction	Collected Images	✗	✗
InterGen [21]	Motion Synthesis	Human-human Interaction	✗	✓
Co-GAIL [36]; HandoverSim [9]	Sim-based Interaction	Simulation	✓	(specific task)
OpenVLA [20]; SayCan [5]	Vision-Language Models	Internet-scale Images and Text; Teleoperation	✓	(text-only)
RHINO (Ours)	Humanoid-Object-Human Interaction	Human-Human Interaction; Teleoperation	✓	✓

lower cost and higher scalability. Based on this, studies encode social scenes [25], simulate reactions [22], or deploy interaction models on robots [29].

Simulation-based Interaction. For tasks with clear goals in interaction, such as handover [9] and collaboration tasks [36], the interaction process can be formulated as a reinforcement learning (RL) problem with a simulation environment. These simulation-based interaction methods can combine manipulation and interaction abilites, but suffer from the sim-to-real gap and can not generalize to general interaction tasks without clear goals, such as waving and shaking hands.

Object manipulation. The ability to manipulate objects is another fundamental skill for a humanoid assistant robot, which requires more precise control of the robot's end-effector. Limited by the dexterity of the robot, especially the degree of freedom of our humanoid robot's arm and hand, imitating learning from real-world teleoperation data [11, 23, 27] is a more practical way to ensure success, compared to learning from human data [30, 37, 43]. Our work learns manipulation skills based on the teleoperation data.

2.3. Unified Interaction Framework

Recent works have attempted to leverage the capacity of general foundation models, such as vison-language models (VLMs), to enable robots to understand human intent in the format of text-based instructions [5, 20, 33]. However, such interaction is often high-latency and not suitable for real-time environments, limiting the potential for immediate response and natural interaction. Asfour et al. [6] designed rules of the real-time human-robot interaction, which is hard to scale up. Cardenas-Perez et al. [8] tries to learn an end-to-end model with 5 different tasks. Limited by the sample efficiency, this end-to-end paradigm makes it difficult to scale to more tasks.

Our framework decouples the interaction process and enables each module to model the interaction with different observation modalities, which is more robust and scalable. We also deploy the framework on a real robot and demonstrate its effectiveness and robustness in more than 20 tasks.

3. Problem Formulation

In this work, we consider the interaction as a leader-follower formulation [35], where the human is the leader and the humanoid robot is the follower. At t time, the leader shows an intent $I_t \in \mathcal{I}$ to ask the follower to perform a skill $K_t \in \mathcal{K}$. $\mathcal{I}$ is a predefined set of human intents and $\mathcal{K}$ is the predefined set of skills. Each skill is a basic ability to complete one simple task, such as picking up a can, brushing a plate, or stamping a file, and the mapping function $f : \mathcal{I} \rightarrow \mathcal{K}$ from intent to skill is a one-to-one mapping.

There are three types of skills that correspond to the leader's intents: interactive motion skills, manipulation skills and idle. The interactive motion skills require the robot to perform expressive and diverse behavior, and the manipulation skills require the robot to interact with objects in the environment precisely. When the human leader does not show any intent, the robot will be in an idle state and maintain its joints in a default state.

Each skill K has a start condition $s_K \in \mathcal{P}$ and an end transition $e_K \in \mathcal{P}$. The start condition shows the required hand occupancy of each hand to start the skill. For example, the skill to cheer with the leader requires the humanoid to hold a can of drink in the right hand. The end transition determines the change of hand occupancy after finishing this skill successfully. For example, for the skill of picking up a can, the start condition and end transition are `[empty, empty]` and `[empty, can]` respectively. A comprehensive description of all skills is shown in Section D.

4. RHINO Framework

The observation of the humanoid robot includes the environment state and the human behavior. To enhance the robustness of human appearances and enable scaling up, our framework decomposes the interaction policy into human-centered and object-centered modules. As Figure 2 (a), we first collect two types of data: human-object-human interaction data and teleoperation data. The human-object-human interaction data is used to train the human-centered

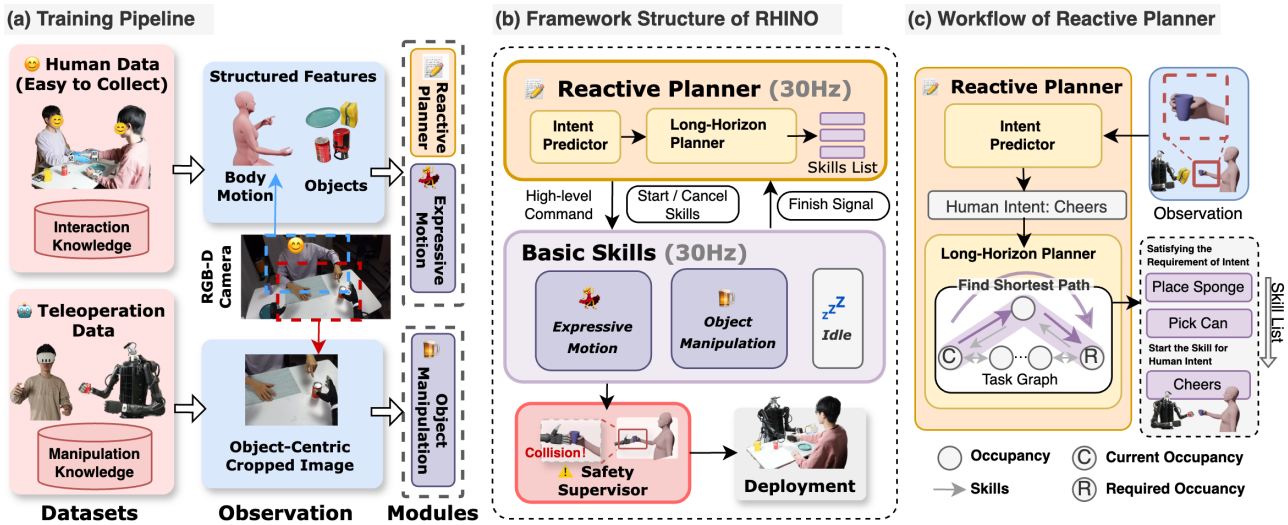


Figure 2. **Overview of RHINO framework.** (a) Training pipeline. Reactive planner and expressive motion modules are trained on human-human data, while the manipulation module is trained on teleoperation data. RHINO extracts human skeleton motion from the image to enhance the generalization of human appearances. (b) Hierarchical framework structure. High-level planner predicts intent and plans skills, while low-level skill follows the high-level signal and generates motor commands. (c) Workflow of the reactive planner. For the example shown, if the intent is cheering when the robot holds a sponge, the planner selects: place sponge → pick can → cheer.

modules, i.e., reactive planner and interactive motion skills, while the teleoperation data is used to train the object-centered modules, i.e,. manipulation skills. The input of human-centered modules is non-image structured data, such as human skeleton motion, and the input of object-centered modules is cropped images of robot-object interaction. Empirical evidence in Section 5 proves that the decoupling design makes RHINO more robust and efficient than a single end-to-end model.

As Figure 2 (b), the high-level planner predicts the leader's intent I from the real-time observation and selects a sequence of skills. It sends start or cancel signals to the low-level controller based on human intention changes or finish signals received from the low-level controller (e.g., starting *cheers* after *pick can* is finished). The low-level controller, composed of motion and manipulation skill modules, generates joint positions based on the leader's intent and the robot's current state. Finally, the safety supervisor monitors the joint positions and stops the robot if a human is too close to it. Figure 3 illustrates the network structure of all submodules.

4.1. Data Collection

Human-object-human interaction data. To learn the interaction between humans and robots, we first collect a dataset of human-object-human interaction [41], where two people perform a series of daily interaction tasks with various objects. In comparison to human-robot interaction data, human-object-human interaction data can be collected without a real robot, which is cheaper to collect and easier to scale to more skills in various of scenarios. The dataset is recorded with a simple motion capture system, and a stereo RGB-D camera in the first-person view of the follower. Motion data is retargeted to the humanoid robot and used by imitation learning algorithms to construct the reactive motion skills. See Sec. B and Sec. C for more details.

Teleoperation data. Manipulation skills, such as picking up a cup, require more precise control of the robot's end-effector. To ensure the success of those skills, we collect demonstrations with a teleoperation system [10], where the human's motion is captured with a VR device, and the robot's joint positions are set by retargeting the human's motion.

4.2. Reactive Planner

The reactive planner is designed to infer the leader's intent, I_t, from real-time observations and plan a skill sequence by finding the shortest path in the task graph. The intent predictor of the planner is a Transformer model, which predicts the leader's intent at a 30 Hz frequency. The input to models is structured features without images, including human body postures, the human's hand and head positions, and the nearest object to the hands, as well as the robot's hand occupancy p_t.

Long-horizon task planning. When the current occupancy is not satisfied with the start condition of a skill $p_t \neq s_K$, the skill is not able to start. In this case, the reactive planner first generates a sequence of skills to satisfy the requirement s_K, then executes the skill $K = f(I_t)$ asked by the leader, as shown in Figure 2 (3). We build a directed graph of occu-

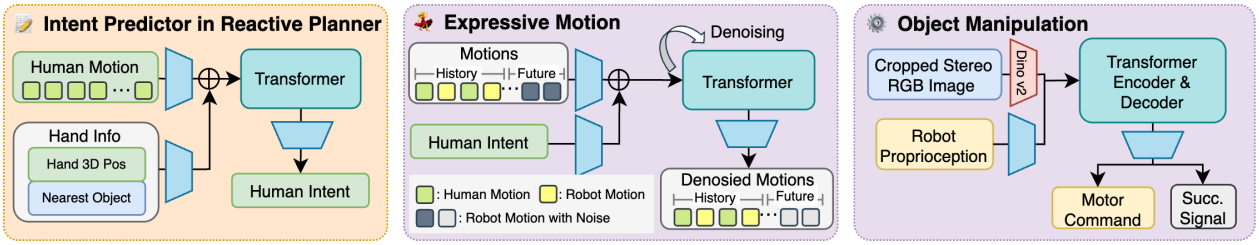


Figure 3. Network architecture of RHINO modules

pancy transitions to find the sequence. The node $n \in \mathcal{P}$ is the hand occupant, and the edge $e \in \mathcal{T}$ is skills. The shortest path from the current occupancy p_t to the start condition s_K, is the skill list to achieve $p'_t = s_K$.

For stability, the reactive planner plans to execute skill $K = f(I)$, only if it consistently predicts an intent I for n_r time steps. After a skill is initiated, the motion skill persists until a change in human intent occurs, while the manipulation skill persists until the execution is judged successful or exceeds a time limit. When a skill is complete, the robot returns to the idle state.

Real-time Interaction. To enable low-latency interaction, the application of most skills can be interrupted by another skill when a different human intent persists for k steps. The motion skills can be easily undone by returning to an idle pose, whereas the interruption of manipulation skills is more complicated, as it requires reversing the object's state. We use a corresponding reverse skill to interrupt each interruptible manipulation skill. For example, the skill of placing the can is a reversal of the skill of picking the can. We show the task graphs and transitions between skills in Section F.

4.3. Interactive Motion Skills

In human-robot interactions, a primary goal is to produce smooth, consistent motions while offering robust, real-time feedback regarding human behavior. To accomplish this, we employ a multi-body motion diffusion model [21] to generate low-level interactive motion skills.

Different from multi-person motion generation, the humanoid and human are heterogeneous and asymmetric in the humanoid-human interaction. We represent the human motion m_t^1 as a 6D rotation vector for each joint, and the humanoid motion m_t^2 as the target of humanoid robot joint positions. Both motions are simplified to arm and hand joints. We also add hand occupancy p_t and human intent I_t as input to the model, to ensure the robot's motion is consistent with the human's intent. Our model predicts the future motion of the humanoid robot $m^2_{t+1:t+10}$ based on the history of human motion $m^1_{t-30:t}$ and humanoid motion $m^2_{t-30:t}$. The model predicts 10 future frames of humanoid motion with a 3 Hz frequency, which generates 30 frames of motion in one second.

4.4. Manipulation Skills

Manipulation skills require precise control of the robot's end-effector, which is difficult to achieve with the retargeted human motion data. As a result, we train an independent Action Chunking Transformer (ACT) [42] model for each low-level manipulation skill, which inference at a real-time frequency of 30Hz. The models are trained on tele-operated demonstrations manually segmented and labeled as atomic manipulation skills such as picking, placing, and cheering.

Robust and safe manipulation. We crop input images to the robot-object interaction area, removing the human body and retaining only hand information, which helps the model focus on the manipulation skills and be robust to human appearance and behavior changes. We also collect in-skill interruption data, where the robot pauses or withdraws its current movement if it collides with the human or the target object is unreachable. Such data enables the robot to exhibit safe behavior and adapt to changes in human behavior or the environment.

Learning terminal conditions. In our multi-skill interactive framework, the model must recognize when a current skill is completed in order to send finish signals back to the high-level planner to move on to the next skill. To implement this, each manipulation policy predicts an additional 0/1 success signal, an indicator of whether the skill is completed, in addition to the joint positions. We add an extra cross-entropy loss to train the 0/1 classification.

4.5. Safety Supervisor

The safety supervisor guarantees safe human-robot interaction by monitoring the distances between the robot and human hands. When human hands are too close to the robot, all joints of the robot are paused to prevent potential harm. Detailed implementation can be found in Section F.

5. Experiment

In experiments, we first demonstrate the necessity of our modular design in Sec. 5.1, comparing it to a simple end-to-end design alternative. The quantitative and human study results show RHINO can scale on various tasks and generalize to changes in human appearances.

To further demonstrate the scalability, we evaluate its

performance in two different scenarios: a dining waiter scenario and an office assistant scenario, also called Scene 1 and Scene 2. The details of the skills in each scenario are shown in Section D. In each scene, the robot should react to the human with its arms, hands, and active head. The results show that all modules in RHINO perform well in the skills. The supplementary video shows demonstration results in two scenarios.

We use Unitree H1 humanoid robot as our real-world experiment platform. Detailed information regarding the setup and deployment of the experiment, including the hardware design, motion and object detection, and our motion capture system, can be found in Section C.

5.1. Framework Performance

1) Framework Structure

We compare RHINO to end-to-end (E2E) ACT baselines [10, 42], with uncropped input images to capture full human motion and environment states, following [8]. Baselines are trained on 1, 3, 5, and 7 skills using 100 slices per skill in a simple scene. We also train improved versions with additional interruption data (E2E-I). However, while as many as n^2 combinations of skills are theoretically needed, we only include 20% for fair comparison, resulting in a rather low performance. In deployment, we test the E2E model in the in-distribution (**I.D.**) scene in which human clothing and object arrangement are the same as the training datasets, and the out-of-distribution (**O.O.D.**) scene where these conditions vary. Figure 5 shows the difference between **I.D.** and **O.O.D.** scenes.

The average success rates of each setting are shown in Figure 4. RHINO outperforms the E2E baselines in all settings, thanks to better prediction of human intent and robustness to O.O.D. data. The high dimensionality and noise of the image input cripple the robustness of the end-to-end policies, leading to more failures under the O.O.D. settings. More detailed results are shown in Section G.

In addition, the E2E framework struggles to scale, with performance dropping as the number of skills increases. In contrast, RHINO maintains stable performance and scales easily—supporting up to 16 skills in our experiments—with minimal effort required to add more via modular training.

2) Human Study

To further evaluate RHINO's effectiveness and robustness in a real-world deployment, we conducted a user study with a total of 21 participants. Participants interacted with each system and ranked them based on general performance and seven detailed abilities. Participants were encouraged to wear personalized clothing to evaluate generalization (see Fig. 6). Detailed settings about the human study can be found in G.5.

RHINO was ranked best overall by 81.0% of participants and outperformed both baselines across all metrics,

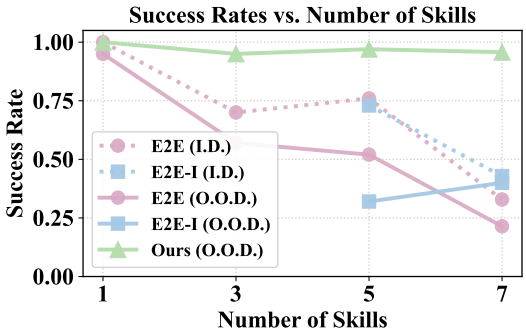


Figure 4. Model success rates with different numbers of skills.

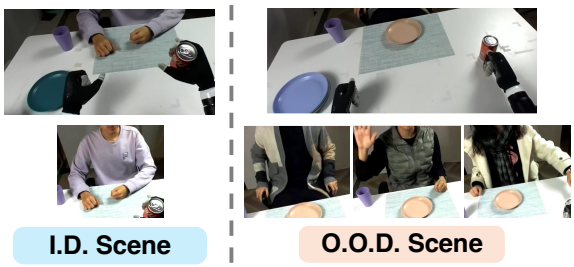


Figure 5. I.D./O.O.D. scene examples (camera view).

Figure 6. Appearances of participants in the human study.

as shown in Fig. 7. We find RHINO generalizes well to human appearances while E2E struggles with multi-task execution and unseen appearances. For example, E2E is likely to keep waving or be confused when picking and placing objects with unseen human appearances. In the supplementary video, we show RHINO can even generalize to interact with a humanoid robot.

5.2. Performance of Sub-Modules

To demonstrate scalability, we deploy RHINO across 20 tasks in two scenes, evaluating sub-module performance. See the supplementary video for examples and Section G for more details.

Human intent prediction. We assess our human intent prediction module using mAP scores, with an 80%-20% train-valid split and additional human-robot interaction data

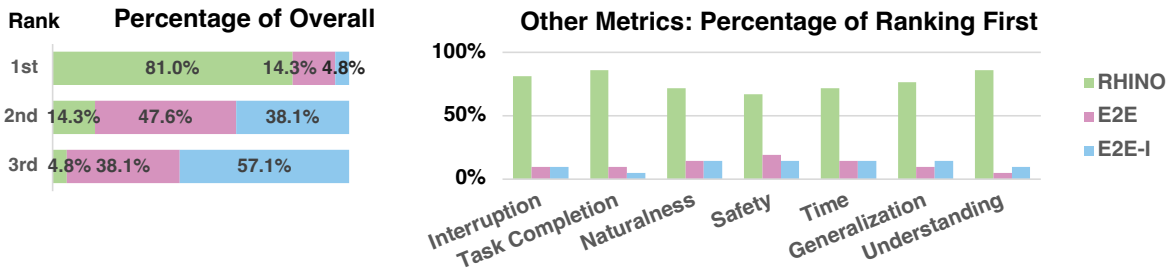


Figure 7. Results of the human study. Left: Overall ranking distribution across the three systems. Right: Percentage of participants who ranked each system first on individual metrics.

Table 2. Performance of the interaction planner.

Method	mAP ↑				Inference Frequency ↑
	Validation Data		Test Data		
	Scene 1	Scene 2	Scene 1	Scene 2	
Qwen2	-	-	0.213	0.167	**30 Hz**
Finetuned Qwen2	0.284	0.322	0.228	0.159	**30 Hz**
ChatGPT-4o-mini	-	-	0.573	0.564	≈ 0.46 Hz
Ours	0.982	**1.0**	**0.787**	**0.643**	**30 Hz**
Ours w/o HD	**0.999**	0.925	0.729	0.587	**30 Hz**

Table 3. Performance of motion generation.

Method	FID↓	JPE(mm)↓	Diversity	MModality↑
Real	-	-	3.74 ±0.05	-
Zero Velocity	43.22 ±0.01	84.82 ±0.01	2.85 ±0.09	-
InterGen	22.95 ±2.28	173.05 ±8.39	3.52 ±0.19	**0.18 ±0.02**
Ours	**10.67 ±0.01**	**48.79 ±0.00**	**3.68 ±0.06**	0.02 ±0.00
Ours w/o Diff.	38.50 ±0.01	142.85 ±0.02	2.91 ±0.04	-
Ours w/o HM	17.34 ±0.05	60.52 ±0.04	3.59 ±0.08	0.06 ±0.01

as the test set. As is shown in Table 2, our model outperforms all baselines despite a performance drop in deployment. The performance drop in the ablation baseline (Ours w/o HD), which excludes object information, highlights the critical role of hand details in intent recognition. The VLMs, Qwen2-VL-2B-Instruct, underperform even after fine-tuning, probably due to a relatively small amount of training data. GPT-4o-mini achieves good results but suffers from slow inference, delaying robot responses.

Motion generation. We evaluate our motion generation module against four baselines on six interaction motions. Baselines include a zero-velocity baseline, InterGen [21], and two ablations: without diffusion (**w/o Diff.**) and without human motion input (**w/o HM**). See Section G for details. As is shown in Table 3, RHINO achieves the best FID, JPE, and diversity scores, demonstrating high motion quality and variability. Removing diffusion or human motion input both degrade FID and JPE, highlighting the importance of conditioning on human motion in generating high-quality reactions.

Objects manipulation. We compare our manipulation module with human teleoperation in Table 4. RHINO matches or surpasses human performance on simpler tasks (e.g., *can*, *tissue*), likely due to training on successful demonstrations only. Performance slightly drops on fine-grained tasks (e.g., *cap*, *stamp*) due to hardware limits such as insufficient DoFs and lack of haptic sensing. The module is slightly slower than humans due to conservative progress prediction and the robot's safe posture initialization. To test robustness, we evaluate the effect of in-skill interruption data during training. As shown in Table 5, with extra 20% interruption data, the success rate in handling interruptions (e.g., withdrawing from *picking a can* when a human interferes) reaches 85%.

6. Limitations

6.1. Analysis of System Failure

As a framework of multiple modules, the failure of the system could be caused by various reasons.

Error and limitation of sensors. Most of the perception of our implementation of RHINO is based on one RGB-D camera. However, the estimation of 3D position often shifts with time and is missing when the estimated object is occluded by other objects or the robot arms. The cumulative error of the sensors leads to a misunderstanding of human intent and incorrect judgment by the safety supervisor.

Stability of hardware. The zero position of the robot arm may have shifted in a small range, which leads to incorrect proprioception. Also, the robot's electronics age over time, which causes errors that require precise control of the robot's end-effector.

Failure of Model Generalization. Due to the limited data collection, the model may fail to generalize to extreme out-of-distribution scenarios, although we mitigate this issue by cropping the image to the region of interest for manipulation skills and using extracted information rather than raw images for human intent prediction. Some unseen human clothing or unexpected object arrangement may lead to the failure of the manipulation. Also, non-standard sitting pos-

Table 4. Performance of manipulation across objects.

Metrics	Scene 1				Scene 2			
	Can	*Plate*	*Sponge*	Tissue	*Cap*	*Book*	*Stamp*	*Lamp*
Succ. Rate	1.00	0.96	0.90	0.95	0.93	0.95	0.93	1.00
Avg. Time	9.41	29.59	23.69	9.43	16.14	10.81	15.17	5.06
Succ. Rate (Human)	0.97	0.98	0.99	0.91	0.91	0.93	0.92	0.96
Avg. Time (Human)	10.42	25.77	17.04	9.54	18.98	10.21	11.84	3.53

Table 5. Performance of manipulation with different ratios of interruption data.

Ratio of data with interruption	Pick Can	Stamp the Paper	Place Plate to Stack
1%	0.00	0.00	0.00
10%	0.05	0.15	0.30
20%	0.85	0.60	0.90

ture or body shape can also have an impact on the prediction of human posture and intent, which leads to a misunderstanding of human intent. Fortunately, human leaders can intervene to correct the robot's behavior in RHINO, which helps prevent a complete breakdown of the system.

6.2. System Limitations

Despite promising results, several limitations remain. Firstly, while RHINO is designed to be scalable, the current implementation is constrained by the availability of high-quality training data. The generalization of the system across a broader range of tasks and environments is still a challenge, as it heavily relies on human demonstrations and teleoperation data, which is time-consuming to collect. Future work will focus on utilizing existing datasets and simulation environments to improve the scalability and generalization of the framework.

Additionally, the current implementation of RHINO is limited to the upper body at a fixed workspace, but a humanoid assistant should have locomotion and navigation abilities in a dynamic environment, and react with whole-body behaviors. Future work should integrate a whole-body controller to extend the framework to whole-body interaction for humanoid robots, and more general tasks with varying levels of human intervention.

7. Conclusion

We propose RHINO, a hierarchical framework that enables humanoid robots to perform real-time, adaptive humanoid-human-object interactions. By decoupling high-level planning and low-level control, RHINO supports fast intent recognition, task interruption, and a wide range of skills from manipulation to expressive motions. Deployed on a real robot, it demonstrates strong flexibility, safety, and responsiveness in dynamic environments—marking a step toward autonomous, human-integrated humanoid systems for daily assistance, disaster response, and industrial tasks.

References

[1] The Dexterous Hands. https://inspire-robots.store/collections/the-dexterous-hands. 11

[2] DYNAMIXEL XL330-M288-T. https://www.robotis.us/dynamixel-xl330-m288-t/. 11

[3] ZED Mini Stereo Camera | Stereolabs. https://www.stereolabs.com/store/products/zed-mini. 11

[4] Don Joven Agravante, Andrea Cherubini, Alexander Sherikov, Pierre-Brice Wieber, and Abderrahmane Kheddar. Human-humanoid collaborative carrying. *IEEE Transactions on Robotics*, 35(4):833–846, 2019. 1, 2

[5] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 3

[6] Tamim Asfour, Fabian Paus, Mirko Waechter, Lukas Kaul, Samuel Rader, Pascal Weiner, Simon Ottenhaus, Raphael Grimm, You Zhou, and Markus Grotz. ARMAR-6: A High-Performance Humanoid for Human-Robot Collaboration in Real-World Scenarios. *IEEE Robotics & Automation Magazine*, 26(4):108–121, 2019. 3

[7] Judith Bütepage, Ali Ghadirzadeh, Özge Öztimur Karadag, Mårten Björkman, and Danica Kragic. Imitating by generating: Deep generative models for imitation of interactive tasks, 2019. 1

[8] Carlos Cardenas-Perez, Giulio Romualdi, Mohamed Elobaid, Stefano Dafarra, Giuseppe L'Erario, Silvio Traversaro, Pietro Morerio, Alessio Del Bue, and Daniele Pucci. Xbg: End-to-end imitation learning for autonomous behaviour in human-robot interaction and collaboration. *IEEE Robotics and Automation Letters*, 2024. 3, 6

[9] Yu-Wei Chao, Chris Paxton, Yu Xiang, Wei Yang, Balakumar Sundaralingam, Tao Chen, Adithyavairavan Murali, Maya Cakmak, and Dieter Fox. Handoversim: A simulation framework and benchmark for human-to-robot object handovers. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6941–6947. IEEE, 2022. 3

[10] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024. 4, 6, 11

[11] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023. 3

[12] Nuno Ferreira Duarte, Jovica Tasevski, Moreno Coco, Mirko Raković, Aude Billard, and José Santos-Victor. Action Anticipation: Reading the Intentions of Humans and Robots. *IEEE Robotics and Automation Letters*, 3(4):4132–4139, 2018. 1, 2

[13] Sadman Sakib Enan, Michael Fulton, and Junaed Sattar. Robotic Detection of a Human-Comprehensible Gestural

Language for Underwater Multi-Human-Robot Collaboration, 2022.

[14] Irving Fang, Yuzhong Chen, Yifan Wang, Jianghan Zhang, Qiushi Zhang, Jiali Xu, Xibo He, Weibo Gao, Hao Su, Yiming Li, and Chen Feng. EgoPAT3Dv2: Predicting 3D Action Target from 2D Egocentric Vision for Human-Robot Interaction, 2024. 1, 2, 3

[15] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid Shadowing and Imitation from Humans, 2024. 2

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 18

[17] Matthew Johnson, Brandon Shrewsbury, Sylvain Bertrand, Tingfan Wu, Daniel Duran, Marshall Floyd, Peter Abeles, Douglas Stephen, Nathan Mertins, Alex Lesman, John Carff, William Rifenburgh, Pushyami Kaveti, Wessel Straatman, Jesper Smith, Maarten Griffioen, Brooke Layton, Tomas de Boer, Twan Koolen, Peter Neuhaus, and Jerry Pratt. Team ihmc's lessons learned from the darpa robotics challenge trials. *Journal of Field Robotics*, 32(2):192–208, 2015. 1

[18] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 11

[19] Abderrahmane Kheddar, Stephane Caron, Pierre Gergondet, Andrew Comport, Arnaud Tanguy, Christian Ott, Bernd Henze, George Mesesan, Johannes Englsberger, Máximo A. Roa, Pierre-Brice Wieber, François Chaumette, Fabien Spindler, Giuseppe Oriolo, Leonardo Lanari, Adrien Escande, Kevin Chappellet, Fumio Kanehiro, and Patrice Rabaté. Humanoid robots in aircraft manufacturing: The airbus use cases. *IEEE Robotics & Automation Magazine*, 26(4): 30–45, 2019. 1

[20] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model, 2024. arXiv:2406.09246 [cs]. 3

[21] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pages 1–21, 2024. 1, 2, 3, 5, 7, 18

[22] Yunze Liu, Changxi Chen, Chenjing Ding, and Li Yi. PhysReaction: Physically Plausible Real-Time Humanoid Reaction Synthesis via Forward Dynamics Guided 4D Imitation, 2024. 1, 3

[23] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021. 3

[24] Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. HOI4ABOT: Human-Object Interaction Anticipation for Human Intention Reading Collaborative roBOTs. *arXiv preprint arXiv:2309.16524*, 2023. 1, 2, 3

[25] Esteve Valls Mascaro, Yashuai Yan, and Dongheui Lee. Robot Interaction Behavior Generation based on Social Motion Forecasting for Human-Robot Interaction, 2024. 1, 3

[26] Edwin Olson. AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3400–3407. IEEE, 2011. 11

[27] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. *arXiv preprint arXiv:2112.01511*, 2021. 3

[28] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 11

[29] Vignesh Prasad, Alap Kshirsagar, Dorothea Koert, Ruth Stock-Homburg, Jan Peters, and Georgia Chalvatzaki. MoVEInt: Mixture of Variational Experts for Learning Human-Robot Interactions from Demonstrations. *IEEE Robotics and Automation Letters*, 9(7):6043–6050, 2024. 1, 3

[30] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system, 2024. 3, 11

[31] Lisa Scherf, Lisa Alina Gasche, Eya Chemangui, and Dorothea Koert. Are You Sure? - Multi-Modal Human Decision Uncertainty Detection in Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 621–629, Boulder CO USA, 2024. ACM. 1, 2

[32] Kyle Wayne Strabala, Min Kyung Lee, Anca Diana Dragan, Jodi Lee Forlizzi, Siddhartha Srinivasa, Maya Cakmak, and Vincenzo Micelli. Towards Seamless Human-Robot Handovers. *Journal of Human-Robot Interaction*, 2(1):112–132, 2013. 1

[33] Daniel Tanneberg, Felix Ocker, Stephan Hasler, Joerg Deigmoeller, Anna Belardinelli, Chao Wang, Heiko Wersing, Bernhard Sendhoff, and Michael Gienger. To help or not to help: Llm-based attentive support for human-robot group interactions. *arXiv preprint arXiv:2403.12533*, 2024. 1, 2, 3

[34] Andreea Tulbure, Firas Abi-Farraj, and Marco Hutter. Fast Perception for Human-Robot Handovers with Legged Manipulators. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 734–742, Boulder CO USA, 2024. ACM. 1, 2

[35] Lorenzo Vianello, Luigi Penco, Waldez Gomes, Yang You, Salvatore Maria Anzalone, Pauline Maurice, Vincent Thomas, and Serena Ivaldi. Human-Humanoid Interaction and Cooperation: A Review. *Current Robotics Reports*, 2 (4):441–454, 2021. 2, 3

[36] Chen Wang, Claudia Pérez-D'Arpino, Danfei Xu, Li Fei-Fei, C Karen Liu, and Silvio Savarese. Co-gail: Learning diverse strategies for human-robot collaboration. *arXiv preprint arXiv:2108.06038*, 2021. 3

[37] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023. 3

[38] Zhen Wu, Jiaman Li, and C. Karen Liu. Human-Object Interaction from Human-Level Instructions, 2024. 1

[39] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-X: Towards Versatile Human-Human Interaction Analysis, 2023. 2

[40] Wei Yang, Chris Paxton, Arsalan Mousavian, Yu-Wei Chao, Maya Cakmak, and Dieter Fox. Reactive Human-to-Robot Handovers of Arbitrary Objects, 2021. 2

[41] Chengwen Zhang, Yun Liu, Ruofan Xing, Bingda Tang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. *arXiv preprint arXiv:2406.19353*, 2024. 4

[42] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 5, 6, 16

[43] Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024. 3

A. Supplementary Video

We provide a supplemental video to showcase the performance and versatility of our system. The video includes: (1) demonstrations of real-time interaction with all skills as described in Section 5.2, presented within both dining and office scenarios; ; (2) a demonstration of long-horizon planning capabilities by the planner; (3) presentations highlighting the system's safety features; (4) demonstrations of in-task interruptions, showing the robustness of the skill; (5) comparisons RHINO with an end-to-end model under out-of-distribution (O.O.D.) conditions; and (6) an illustrative example of the RHINO to generalize interactions with a Unitree G1 humanoid robot.

B. Dataset

Figure 8 shows information collected in the interaction and teleoperation dataset and the requirements of real robot deployment. More details are described in Section F.

For the interaction data, we label each frame t in the interaction data with the leader's intent I_t and the follower's skill $K^{(t)}$, which are represented as ID integers of intentions and skills. We add additional labels for the occupancy $p \in \mathcal{P}$ of the robot, indicating whether the end effector (i.e., the hands) is empty or interacting with an object, with distinct labels assigned to different objects. The motion capture system that collects the follower's behavior is described below.

For manipulation data, the teleoperation system records the control commands, the robot's proprioception, and stereo videos from a camera on the robot's head to perceive the environment. We also label the frames where the skill showcase is completed successfully, referred to as the success signal, which is used to learn the finish condition of the manipulation skills.

C. Real World Setup

Deployment hardware. We implement RHINO on a Unitree H1 humanoid robot with an active head, illustrated in Figure 10. Following Cheng et al. [10], we assembled two DYNAMIXEL XL330-M288-T motors [2] with 3D printed gimble parts and a ZED Mini stereo camera [3] for two-DoF (yaw and pitch) active sensing. Each arm of H1 has 5 DoFs and a 6-DoF end-effector from [1], and other DoFs on the robot are not used.

Motion capture system. We use AprilTag [26] and two cameras to build a simple motion capture system. We place four AprilTag markers on the four corners of the workspace table to locate the cameras. Each human in the workspace wears two 3D-printed wristbands with four AprilTag markers on each of them, illustrated in Figure 9. The cameras are calibrated and located using the OpenCV library and capture the human wristband's position in real time. Each wristband has an additional IMU sensor to capture the orientation of the human's wrist. To reduce the noise in the collected data, we use a Kalman filter to smooth the data.

Motion Detection and Object Detection The human body posture is detected from a body detection model in ZED API with 38 joints and represented as 6D rotation vectors of joints. We capture the leader (human)'s intent and the state of objects from the RGB-D images. For hand detection, we use HaMeR[28] to obtain the human hand motion and then retarget[30] it to the 6-DoF robot hand. The visualization results are shown in Figure 11. We record the information of the nearest objects to the hands, including the object ID, distance to the hand, and the IOU of the object bounding box and the hand bounding box. The 3D bounding box of objects and hands is detected from a stereo RGB-D camera, ZED-mini, with a fine-tuned YoloV11 [18] model and a body detection model in ZED API. The RGB-D stereo camera enables the retrieval of the 3D position of any 2D pixel in the image, represented in world coordinates. The orientation of the humanoid's body determines the X and Y axes, while gravity defines the Z axis.

D. Skill Descriptions

In this section, we describe the skills that we deploy on the humanoid robot. The details include the description, success condition, reverse skill (if it exists), and the human intent related to the skill. Note that the intent is inferred mainly from human behavior, hand positions, and the relative location of objects. The latter two can be concluded trivially from the start condition and end transition, and are shown by skill in Table 6. We concern the intent mainly with the human body motion in the narration as follows:

1) Scenario 1: Humanoid as a Dining Waiter

In this scenario, the human leader and the robot sit face-to-face at the side of a dining table. There are plates with food, a Coke can, a tissue box, and a sponge on the table. In the following skill descriptions, the humanoid robot takes the role of a helpful waiter and serves the leader (human) a meal with these objects. 10 skills related to 4 objects are listed below:

- *Pick Can*: The robot picks up a can with its right arm from the table. The success condition is that the can is lifted off the table. The interruption data includes the human taking the can away or the human putting their hand on the can. *Place Can* is the corresponding reverse skill. The leader (human) shows the intent by pointing to the can when the right robot arm is empty.
- *Place Can*: The robot places the can back on the table with its right arm. The success condition is that the can is placed on the table and the robot's hand is lifted off the can. The intent of this skill is shown by the leader (human) pointing to the place on the table where the can was placed before.

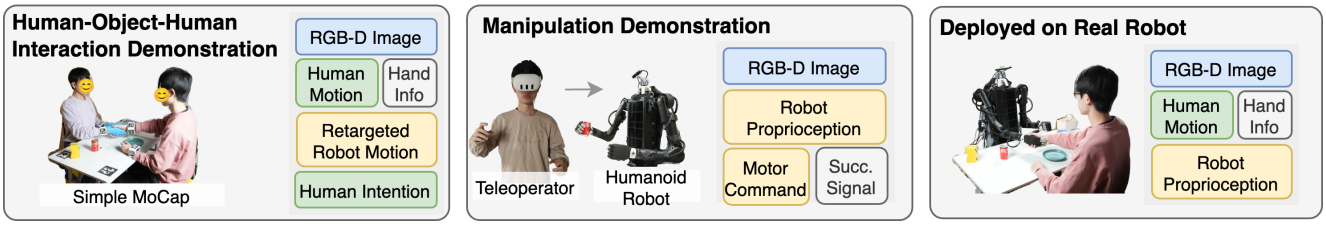


Figure 8. **Information collected in the dataset and the requirement of real robot deployment.** The left box shows our Human-Object-Human Interaction Demonstration Data, which contains human motion and hand position information collected from RGB-D images and a simple motion capture system. The box in the middle illustrates our Manipulation Demonstration Data, which is collected through teleoperation. The right box shows the information required in real robot deployment.

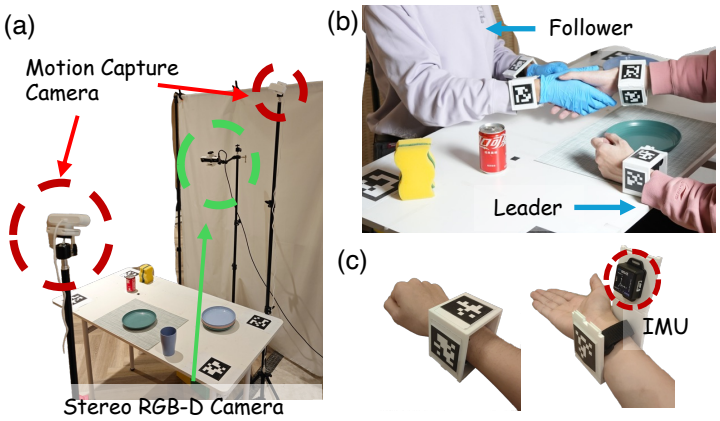


Figure 9. **Setup of motion capture system.** a) The two Motion Capture Cameras are used to detect the ArUco markers. b) Follower and Leader both wear wristbands with ArUco markers for hand position detection. c) We 3D-print our wristbands with 4 ArUco markers on 4 surfaces and embed an IMU beneath the upper surface.

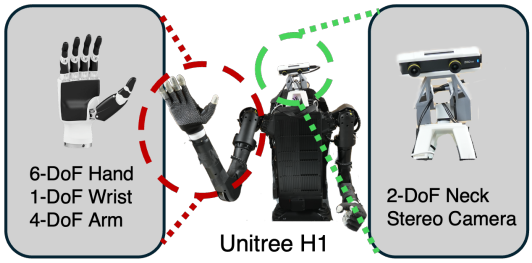


Figure 10. The humanoid robot platform.

- *Get Plate from Human*: The robot fetches a plate from the hand of the human with its left arm. The success condition is the plate in the dexterous hand when the human loosens the grip of the plate. The leader (human) shows the intent by handing a plate forward.
- *Place Plate to Stack*: The robot stacks the plate in its right hand onto a pile of plates on the table. The success condition is that the plate is settled on top of the plate pile without slipping. The intent is given by the leader (human) pointing to the stack. Interruption data, in which the human touches the plate to stop the motion, is added to the collected dataset.
- *Pick Place from Table*: The robot lifts a plate from the table by both arms and holds it in the left hand. Application of this skill succeeds if there is no slippage in the motion until the plate is held. The leader (human) points to the plate to show the intent.
- *Handover Plate*: The robot protracts its left arm to pass the plate to the leader (human). The showcase of this skill ends when the plate is put into the human hand. The leader (human) simply stretches the right hand out to show the intent. It is the reverse skill of *Get Plate from Human*.
- *Pick Sponge*: The robot picks up the sponge with its right arm. The sponge is placed beside the can. When it is lifted off the table, the skill showcase succeeds. The leader (human) shows intent by mimicking washing. Data where the human snatches the sponge before the robot reaches it adds to the dataset. In case this happens during deployment, the robot withdraws its hand to the idle state.
- *Brush with Sponge*: It is a complex skill using both arms. The start condition is a plate in the left hand and a sponge in the right hand when the leader (human) makes the washing gesture (same as that in *Pick Sponge*) again. To apply this skill, the robot moves the sponge close to the plate and rubs the sponge on the plate to brush it. The success condition is that the robot keeps the periodic brushing motion for over 10 seconds.
- *Place Sponge*: The reverse skill of *Pick Sponge*. The robot puts the sponge in the right hand back onto the table

Figure 11. **Visualization results of human body motion detection, object detection, and human hand detection.** The left image demonstrates the human motion detection (only upper body is used) and object detection (in Dining scenario). The right image demonstrates the human hand detection.

to complete the skill demonstration. The intent is shown by the leader (human) pointing to the place on the table where the sponge was placed before (similar to skill *Place Can*).

- *Pick a Piece of Tissue*: The leader (human) points to the tissue box to express the intent. Then the robot uses its left hand to pull a tissue from the tissue box placed on the table corner and gives it to the leader (human). The skill showcase succeeds when the leader (human) receives the tissue.

2) Scenario 2: Humanoid as an Office Assistant

In this scenario, the leader (human) sits across the humanoid robot at an office table. The robot transforms into an office assistant and deals with complicated cases such as stamping paper for approval, settling a baseball cap on the rack, picking up and handing a book over, and reacting properly if the human takes a nap. There are 7 skills related to 4 objects in this scenario.

- *Settle Cap*: The robot gets a cap from the human leader's hand and settles it on a hat rack with its right arm. The skill showcase begins with the human holding the cap with both hands and ends with the robot pulling its hand back from the hat rack.
- *Handover Cap*: The robot takes the cap off the hat rack and sends it to the leader (human). It is the reverse skill of *Settle Cap*. The related intent is inferred when seeing the human pointing to the rack. The success condition is that the human has received the cap.
- *Pick Book*: The robot picks a book from the shelf and hands it over. The skill begins with the human gesturing toward the book. When the human takes the book, this skill is completed successfully.
- *Pick Stamp*: The robot picks up the stamp on the table with its right hand. The skill succeeds when the stamp is lifted near the hand in an idle posture. The leader (human) instructs the execution of this skill by passing along the paper.
- *Stamp the Paper*: It is a delicate operation to make an issue for approval. The robot presses the stamp down onto the paper to mark a sign. This skill is considered successful only if one mark is imprinted. It is noteworthy that printing more than one mark in a single execution means that the model fails to predict the ending, thus being treated as a failure case. The sign of the related intent is the leader (human) pointing at the paper. To make an in-skill interruption, the human covers the paper with a hand to make the robot withdraw its hand if the pressing is not done.
- *Place Stamp*: The robot places the stamp back with its right hand. It is the reverse skill of *Pick Stamp* and is triggered by withdrawing the paper.
- *Turn off/on the Lamp*: Turning on and off the lamp share the same motion, and thus are trained as one skill. When the human slumps over the office desk to take a nap, the robot taps the switch of the lamp to turn it off. And when the human wakes up and lifts the head, the robot performs the same motion to turn on the lamp.

3) Interactive Motion Skills

Some skills are not involved with object operation and, thus, are not trained as manipulation skills. They are noted as motion skills. These skills are considered successful when the robot performs the motion properly, as the human shows the intent and recovers to the idle posture when the intent no longer sustains.

- *Cheers*: The robot reaches out its right hand to touch the bottle held by the right hand of the human. Though holding a Coke can in the right hand during deployment, the robot does not manipulate the object. For this reason, this skill is not trained in a manipulation demonstration model.
- *Wave*: The robot lifts up the right hand and waves the right hand when the leader (human) is waving also.
- *Shake Hands*: The robot stretches its right hand out to touch the hand of the leader (human) with a handshaking posture.
- *Take Photo*: The robot lifts the right hand and makes a V-sign when the human raises the phone to take a photo, and puts the hand down as the human puts away the phone.

Table 6. Description of the skills. Notes: The **Start Condition** or **End Transition** `[A, B]` means that object `A` is in the left hand of the robot and `B` is in the right hand. `empty` for this hand must be empty, `any` for this hand could hold any object or be empty, and − for this object remains unchanged after the skill is completed.

Scenarios	Object	Skill Name	Start Condition	End Transition	Num. of Data	Arm
Scenario 1 Dining Waiter	can	Pick Can	[any, empty]	[-, can]	107	Right
		Place Can	[any, can]	[-, empty]	100	Right
	plate	Get Plate from Human	[empty, any]	[plate, -]	100	Left
		Place Plate to Stack	[plate, any]	[empty, -]	98	Left
		Pick Plate from Table	[empty, empty]	[empty, plate]	115	Dual-Arm
		Handover Plate	[plate, any]	[empty, -]	115	Left
	sponge	Pick Sponge	[any, empty]	[-, sponge]	89	Right
		Brush with Sponge	[plate, sponge]	[-, -]	81	Dual-Arm
		Place Sponge	[any, sponge]	[-, empty]	82	Right
	tissue	Pick a Piece of Tissue	[empty, any]	[-, -]	105	Left
Scenario 2 Office Assistant	cap	Settle Cap	[any, empty]	[-, -]	111	Right
		Handover Cap	[any, empty]	[-, -]	110	Right
	book	Pick Book	[empty, any]	[-, -]	115	Left
	stamp	Pick Stamp	[any, empty]	[-, stamp]	92	Right
		Stamp the Paper	[any, stamp]	[-, -]	87	Right
		Place Stamp	[any, stamp]	[-, empty]	89	Right
	lamp	Turn off/on the Lamp	[empty, any]	[-, -]	85	Left
Expressive Motions	None	Cheers	[any, can]	[-, -]	66	Dual-Arm
		Wave	[any, empty]	[-, -]	39	Dual-Arm
		Shake Hands	[any, empty]	[-, -]	51	Dual-Arm
		Take Photo	[any, empty]	[-, -]	31	Dual-Arm
		Thumb Up	[empty, empty]	[-, -]	22	Dual-Arm
		Spread out Hands	[empty, empty]	[-, -]	26	Dual-Arm

- *Thumb Up*: The robot reaches both hands out with the thumbs up as the human gives it a thumbs-up. Human intent with the left hand, right hand, or both is approved.
- *Spread out Hands*: The robot stretches its arms out to the sides with palms up when the leader (human) spreads its hands out.

E. Prompt for VLMs

Here are the prompts we give to Qwen and GPT-4o-mini in the evaluation of the intent prediction module.

Prompt for the Dining scenario

You are a humanoid robot sitting in front of a human and equipped with a camera slightly tilted downward on your head, providing a first-person perspective. I am assigning you a new task to recognize to human gestures in front of you. Remember, the person is sitting facing you, so be mindful of their gestures. If the person is holding a cup to you and trying to cheer with you, answer 'Cheers'. If the person is giving you a thumbs-up, answer 'Thumbup'. If the person extends their right hand to shake hands with you, answer 'ShakingHand'.If the person is waving to you with the right hand, answer 'Waving'. If the person is taking a photo of you with a cellphone, answer 'Taking Photo'. If the person is spreading out both hands in a gesture of resignation, answer 'Spreading Hands'. If the person is pointing to a Coke can in the middle of the table (on your right side), answer 'Pointing Can'. If the person is pointing to an empty spot on the table with no objects (on your right side), answer 'Pointing Table'. If the person is pointing to a tissue box at the far left of the table, answer 'Pointing TissueBox'. If the person is pointing to a plate in the middle of the table (just in front of you), answer 'Pointing Plate'. If the person is holding out the right hand with the palm open toward you, answer 'Palmup'. If the person is handing you a plate, answer 'Handing Plate'. If the person is clenching their right fist, holding their left hand open and upward, and placing their right hand above the left as if pretending to wash a plate, answer 'Washing'. If the person is pointing at a stack of plates on the left side of the table, answer 'Pointing Plates'. If the person is pointing at a sponge on the right side of the table, answer 'Pointing Sponge'. If the person is crossing his arms to form an X shape, answer 'Cancel'. If no significant gestures are made, answer 'Idle'. Respond directly with the corresponding options [Cheers, Thumbup, ShakingHand, Pointing Can, Pointing TissueBox, Pointing Plate,

Palmup, Handing Plate, Washing, Pointing Plates, Pointing Sponge, Cancel, Idle] based on the current image and observed gestures. Directly reply with the chosen answer only, without any additional characters.

Prompt for the **Office** scenario

You are a humanoid robot sitting in front of a human and equipped with a camera slightly tilted downward on your head, providing a first-person perspective. I am assigning you a new task to recognize human gestures in front of you. Remember, the person is sitting facing you, so be mindful of their gestures. If the person is giving you a thumbs-up, answer 'Thumbup'. If the person extends their right hand to shake hands with you, answer 'ShakingHand'. If the person is waving to you with the right hand, answer 'Waving'. If the person is taking a photo of you with a cellphone, answer 'Taking Photo'. If the person is spreading out both hands in a gesture of resignation, answer 'Spreading Hands'. If the person is handing you a cap, answer 'Handing Cap'. If the person is pointing at a cap placed on the right of the table, answer 'Pointing Cap'. If the person is handing a document to you with both hands and you are NOT holding a stamp, answer 'Handing File'. If a document is placed in the center of the table in front of you, and the person is pointing to it with the right hand, answer 'Pointing File'. If the person retrieves the document from your side of the table to the other side, directly across from you, and you are still holding the stamp, answer 'Retrieve File'. If the person is lying down on the table and the lamp is ON, answer 'Lie Down'. If the person is sitting up from the table and the lamp is OFF, answer 'Sit up'. If the person is pointing at the books standing in the top left corner of the table, answer 'Pointing Book'. If the person is crossing the arms to form an X shape, answer 'Cancel'. If no significant gestures are made, answer 'Idle'. You are NOT holding a stamp right now, and the lamp is now ON. Observe the image and gestures carefully. Respond directly with the corresponding options [Pointing Book, Handing Cap, Pointing Cap, Handing File, Pointing File, Retrieve File, Lie Down, Sit up, Shaking Hand, Thumbup, Cancel, Idle]. Directly reply with the chosen answer ONLY, without any additional characters.

The sentence '*You are NOT holding a stamp right now and the lamp is now ON*' is modified at each query according to the current situation (whether the robot is holding a stamp and whether the lamp is on).

F. Implementation of RHINO Modules

1) Reactive Planner

The planner is a Transformer model, which takes the human's motion and the environment information as input, and predicts the leader's intent at a 30Hz frequency. To enhance the generalization of the model, we do not input observed images directly, but extract the human's body and hand postures, the human's hand and head position, and the nearest object to the hands from the RGB-D images, as well as the robot's hand occupancy p_t. We retarget human hand postures to a robot hand with 6 degrees of freedom (DoF) and represent the hand posture as the position of each joint.

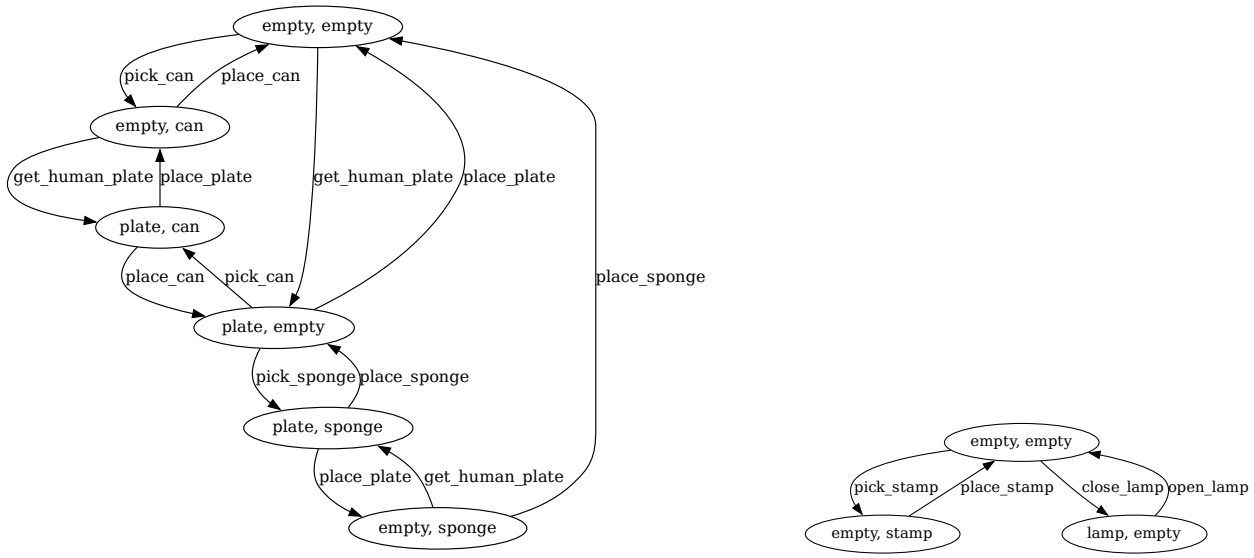
Algorithm 1 Pseudo-code for Skill Transitions of Reactive Planner.

```
Skill ← Idle
while true do
    human_intent ← Recognize_Human_intent()
    if human intent is stable for k frames and human intent != current intent then
        if human intent = Idle and Skill = Manipulation then
            Continue
        if Skill = Manipulation and interruptionAllowed then
            Skill ← Reverse_Skill(Skill)
        if Start_Condition(human_intent) is not satisfied by hand occupancy then
            path ← FindPath(occupancy, StartCondition(human_intent))
            Skill ← Execute_Path(path)
        else
            Skill ← Corresponding Skill(human_intent)
    else if SkillSucceeded(Skill) or SkillTimeout(Skill) then
        Skill ← Idle
        if SkillSucceeded(Skill) then
            Hand Occupancy ← End_Transition(Skill)
```

The humanoid robot starts from an idle state. If the reactive planner predicts a human intent I consistently for n_r time steps, the humanoid robot switches to the corresponding skill $T = f(I)$. The human intents are meant to "start" the execution of a skill, rather than "keep" the current execution. For example, the leader only needs to point at the can for a while at the beginning to get the robot to pick the can on the table, instead of pointing all the time.

The switching logic of the skill planner is listed in Algorithm 1. The directed graphs of occupancy are shown in Figure 12. Here we further explain the `Recognize_Human_intent()` function in detail, which is implemented as a transformer-based classifier. The model input includes:

- **Upper Body Human Posture**: a 36-dim human upper body skeleton, namely the 6D rotation of the wrist, elbow, and shoulder joints for each arm.
- **Human Hand Pose**: a 12-dim human hand pose vector. For each hand, we retarget the detected human hand pose



(a) Dining scenario.

(b) Office scenario.

Figure 12. Occupancy graph of two different scenarios.

to our robot hand with IK, and take the 6 joint poses as the human hand pose vector.

- **Robot Hand Occupancy**: a 10-dim robot hand occupancy label. Since we have at most 5 objects in total (Can, Cup, Plate, Sponge, Tissue), we use a 5-dim one-hot label for each hand to represent the object held in the robot's hand. If the robot is not holding anything, the label will be all zeros.
- **Human Details**: a 19-dim vector, including the x and y-axis of each human hand position, the z-axis (height) of the human head position, and a 7-dim label for the nearest object to each hand. The nearest object label is concatenated by a 5-dim one-hot label of the object type, the distance from the object to the human hand, and the average of IOU and IOFs of the object bounding box and the human hand bounding box.

We use an MLP encoder to encode the concatenated vector of **Upper Body Human Posture**, **Human Hand Pose**, and **Robot Hand Occupancy**, and another MLP to encode **Human Details** to a latent dimension. The concatenated latent vector is processed by a Transformer backbone, followed by a final MLP layer to predict the human intent class. The hyper-parameters of the Transformer backbone are listed in Table 7.

2) Interactive Motion Generation

For interactive motion generation, we use a transformer-based diffusion model, which denoises the past 30 frames of human and robot motions and the future 5 frames of robot motions. Both human motion and robot motion consist of upper-body motion (36-dim for humans and 10-dim for humanoid), hand motion (6-dim for each hand,) and hand occupancy label (5-dim one-hot label for each hand). Besides, the predicted human intent label is also conditioned during the diffusion process. The hyper-parameters of our model are listed in Table 8.

3) Manipulation Skills

Thanks to the stability of model training, most of the hyperparameters are basically consistent across all skills. The volume of data for training each skill is shown as a column in Table 6. The hyper-parameters in training ACT models [42] are shown as Table 9.

For the prediction of the success signal, we marked the last n_s frames of the recorded data as 1 (completed) and other frames as 0 to generate a 0/1 label. n_s is set to 25 in most of the skills and shifted to 10 in three of them, of which the ending frames changed sharply in motion. The special skills are *Pick Stamp*, *Stamp the Paper*, and *Place Stamp*.

4) Safety Supervisor

A safety supervisor serves as a global guarantee of safe robot actions, which forces the robot to pause immediately when potential harm is detected. In this module, the collision box of the robot is calculated based on several selected key points on the arms and is updated by forward kinematics as the movement of the arm. The key points at specific joints and their midpoints are identified as follows:

- The origins of the shoulder pitch, shoulder yaw, elbow, and wrist joints are defined as key points.

Table 7. Hyperparameters of the Reactive Planner.

hyper-parameter	value
latent dimension	128
num head	8
num layers	3
batch size	256
feed-forward dimension	128
maximum epoch	300
learning rate	0.0001

Table 8. Hyperparameters of the Motion Generation Model.

hyper-parameter	value
latent dimension	256
num head	8
num layers	4
feed-forward dimension	256
diffusion steps	300
sampling steps	30
batch size	512
maximum epoch	4000
learning rate	0.0001

Table 9. Hyper-parameters of the ACT model for manipulation skills.

hyper-parameter	value
KL weight	10
Cross-entropy weight	1
chunk size	30
hidden dimension	512
batch size	45
feed-forward dimension	3200
maximum epoch	50000
learning rate	0.00005

Table 10. Performance of manipulation module across manipulation skills.

Scenarios	**Object**	**Skill Name**	**Success Rate**	**Average Time**	**Success Rate (Human)**	**Average Time (Human)**
Scenario 1 Dining Waiter	can	Pick Can	1.00	5.31	1.00	5.77
		Place Can	1.00	4.10	0.93	4.65
	plate	Get Plate from Human	1.00	4.86	0.98	5.12
		Place Plate to Stack	0.95	8.19	0.97	6.91
		Pick Plate from Table	0.90	10.75	0.96	8.60
		Handover Plate	1.00	5.79	1.00	5.14
	sponge	Pick Sponge	0.95	8.19	1.00	7.45
		Brush with Sponge	0.90	10.02	1.00	4.18
		Place Sponge	0.85	5.57	0.98	5.41
	tissue	Pick a Piece of Tissue	0.95	9.43	0.91	9.54
Scenario 2 Office Assistant	cap	Settle Cap	1.00	7.50	0.91	8.50
		Handover Cap	0.85	8.64	0.90	10.48
	book	Pick Book	0.95	10.81	0.93	10.21
	stamp	Pick Stamp	1.00	4.80	0.92	3.91
		Stamp the Paper	0.80	5.64	0.92	3.11
		Place Stamp	1.00	4.74	0.93	4.83
	lamp	Turn off/on the Lamp	1.00	5.06	0.96	3.53

- Additional key points include the midpoints between the shoulder yaw and elbow joints, and between the elbow and wrist joints.
- A further key point is defined at one-third the distance beyond the elbow towards the wrist, extending from the segment between these two joints.

This structured delineation allows for precise calculations pertinent to robotic arm movements within a predefined spatial configuration.

Meanwhile, the global coordinates of human hands are obtained by the depth camera. The human hands are shaped by the detected key-points from the body detection model from ZED API, from which each hand is reconstructed as 5 points.

The safety module judges whether the robot's collision box is going to collide with human hands. We use the Euclidean distance from human hand key points to robot arm key points as a simple but effective approach to calculating the collision box. Once one of the points is close to any robot key point in 0.1 meters, an unsafe signal is broadcast to pause the robot control.

The safe supervisor takes strong measures to ensure the robot would not hurt humans by avoiding collision, especially in skills where they should contact at a rather close distance, such as a handshake and handing over the plate.

We also provide the visualization of the safety supervisor in Figure 13. When the human hand key points collide with any collision box, the supervisor will send an unsafe signal

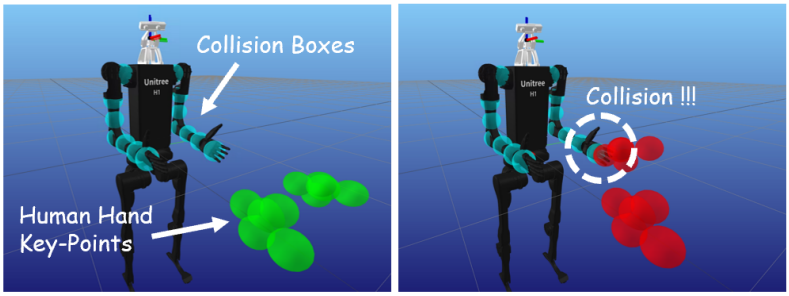


Figure 13. Visualization of the safety supervisor, with sphere markers representing the collision boxes of the human hands and robot arms. These markers move in sync with the interaction. When an unsafe collision is detected, the human hand markers change color from green to red.

to halt the robot. Our safety supervisor runs at 30Hz.

G. Detailed Experiment Results

G.1. Framework Structure

The detailed result to derive Figure 4 is shown in Table 11. For each single skill, we collected 100 slices of motion in the dataset, which is close to the average volume of all the manipulation skills. For the end-to-end model with interruption(E2E-I), an additional 20 slices of interruption data are collected for each skill. An interruption data starts with performing an arbitrary skill, then the human leader shows an intent to switch to the corresponding skill in the middle. As the number of skills increases, the end-to-end models are likely to fail to learn some skills. For example, E2E with 7 skills can only keep picking and placing, ignoring human intentions regarding the plate. As Figure 7 shows, E2E and E2E-I have similar score, which illustrate that the interruption data does not improve the interaction ability of the end-to-end model.

G.2. Human intent Prediction

We use confusion matrices to show the classification performance of our planner on the test dataset. The confusion matrices for our model, our model without human details, and GPT-4o-mini on the test datasets of the dining and office scenarios are shown in Figure 14.

As is shown in the confusion matrices, although the model mainly relies on human body motion and human hand motion input for classification, human details can help the model better deal with certain situations, such as avoiding misclassification into Idle.

G.3. Motion Generation

We compare our motion generation module with 4 baselines on all 6 expressive motions involved in our experiment (*handshake*, *wave*, *cheers*, *thumbup*, *spread hand*, *take photo*). The baselines are:

- **Zero Velocity**: the repetition of the last pose observed, as a simplest baseline.
- **InterGen**: the adapted version of human interaction motion synthesis model [21]. To align the input modalities, we use forward kinematics to calculate the joint 3D positions in the training data. While original InterGen is built for text-to-motion generation, our task is to generate 10 frames of motions based on the human intent and 30 frames of history motions. Therefore, we re-train the model with the intent as text input and input 40 continuous frames of motions. The losses remain the same as the original InterGen.
  During inference, the past 30 frames of history motions are fixed to serve as conditional information. For evaluation, we take the synthesized joint angles of one side (humanoid) from the 10 frames of future motions and calculate the metrics in the same way as other baselines.
- **Ours (w/o diffusion)**: Generate the motions directly with a Transformer model with the same structure as the denoiser used in our Diffusion-based method, without the full diffusion framework.
- **Ours (w/o human motion)**: Generate humanoid motions only conditioning on the human intent label and history of humanoid motions, without the guidance of detailed human motion.

The metrics to evaluate the performance of the motion generation modules are:

- **FID**: The FID score [16] is leveraged to assess the similarity between synthesized and real motions quantitatively. Specifically, our encoder is trained with cross-entropy loss and encodes the humanoid motions into latent embeddings.
- **JPE**: We calculate the Joint Position Error (JPE) based on the forward kinematic results of the generated robot joint positions. JPE is averaged over all wrist joints and finger joints.
- **Diversity**: We calculate the average Euclidean distances of 300 randomly sampled pairs of motions in latent space

Table 11. Detailed success rate of RHINO framework and end-to-end model on different skills.

Scene	Method	Task Num.	cheers	pick	place	handshake	wave	Get Plate from Human	Handover Plate
I.D.	E2E	1	1.00	-	-	-	-	-	-
	E2E	3	1.00	0.75	0.35	-	-	-	-
	E2E	5	0.95	0.70	0.80	0.85	0.90	-	-
	E2E-I	5	0.70	0.85	0.55	0.95	0.60	-	-
	E2E	7	0.35	0.60	0.45	0.30	0.30	0.20	0.10
	E2E-I	7	0.75	0.55	0.15	0.45	0.05	0.70	0.35
O.O.D.	E2E	1	0.95	-	-	-	-	-	-
	E2E	3	0.95	0.30	0.45	-	-	-	-
	E2E	5	1.00	0.55	0.35	0.60	0.85	-	-
	E2E-I	5	0.30	0.55	0.10	0.35	0.30	-	-
	E2E	7	0.00	0.50	0.15	0.20	0.35	0.20	0.10
	E2E-I	7	0.30	0.25	0.30	0.15	0.75	0.60	0.45
O.O.D.	Ours	5	1.00	1.00	0.85	1.00	1.00	-	-
		7	0.95	0.85	0.95	0.95	1.00	1.00	1.00

to measure motion diversity in the generated motion dataset. The Diversity of motions generated by the model is expected to be closer to the Diversity of Real Data.

- **MModality**: MModality measures the ability to generate diverse motions for the same human intent label. We sample 20 motions within one fixed human intent and motion to form 10 pairs, and measure the average latent Euclidean distances of the pairs. Note that the baseline **Ours (w/o diffusion)** will always yield the same output for fixed inputs and thus do not possess MModality.

G.4. Objects Manipulation

The detailed success rates and average execution times across skills are presented in Table 10, from which the statistics in Table 4 are derived. We compute the success rate and the average time based on 20 independent tests, on the main object in two different scenes.

In most skills, the manipulation module of RHINO autonomously executes motions following the patterns of teleoperation data within a comparable time frame. Trained exclusively on successful human teleoperation cases, the module demonstrates both effectiveness and robustness to slight scene variations during deployment. As a result, it achieves higher success rates in skills involving simple motions with abundant training data, such as *Pick Can*, *Handover Plate*, and *Place Stamp*.

However, certain skills pose challenges for the manipulation module. In *Place Plate to Stack* and *Stamp the Paper*, the robot hesitates to drop the plate or press the stamp due to prediction noise. In *Pick Plate from Table*, it must overcome increased friction against the table when joint positions deviate from those in the collected data. Another challenge arises in *Brush with Sponge*, where the success signal predictor struggles to assess the progress of the periodic motion accurately. As a result, termination is constrained by a 10-second timeout. These various factors contribute to a longer average execution time for these four skills compared to human performance.

Referring to the experiment on in-skill interruption data presented in Table 5, we select *Pick Can*, *Stamp the Paper*, and *Place Plate to Stack* as representative skills for interruptions occurring during the stages of fetching an object, operating with the object, and returning the object, respectively.

To ensure a controlled data volume across different interruption ratios, we assign a fixed data amount of M to each skill. In a full data collection for any given skill, the total data amount is N, with N_{in} representing the portion containing in-skill interruptions. The ratio of interrupted data in a selected subset is denoted as α, meaning that $\lceil \alpha M \rceil$ slices contain interruptions. To maintain this proportion, we set $M = N - N_{in} + 1$. Specifically, M is set to 69, 66, and 76 for the three skills, respectively.

Each ACT model in this experiment uses the same hyperparameters as those employed for the corresponding skill in both training and deployment with the full data collection.

G.5. Human Study

G.5.1. Participants and Experimental Setup

A total of 21 participants (17 males, 4 females) aged 20 to 28 years were recruited through an internal university platform. The cohort exhibited physical characteristics of 1.58

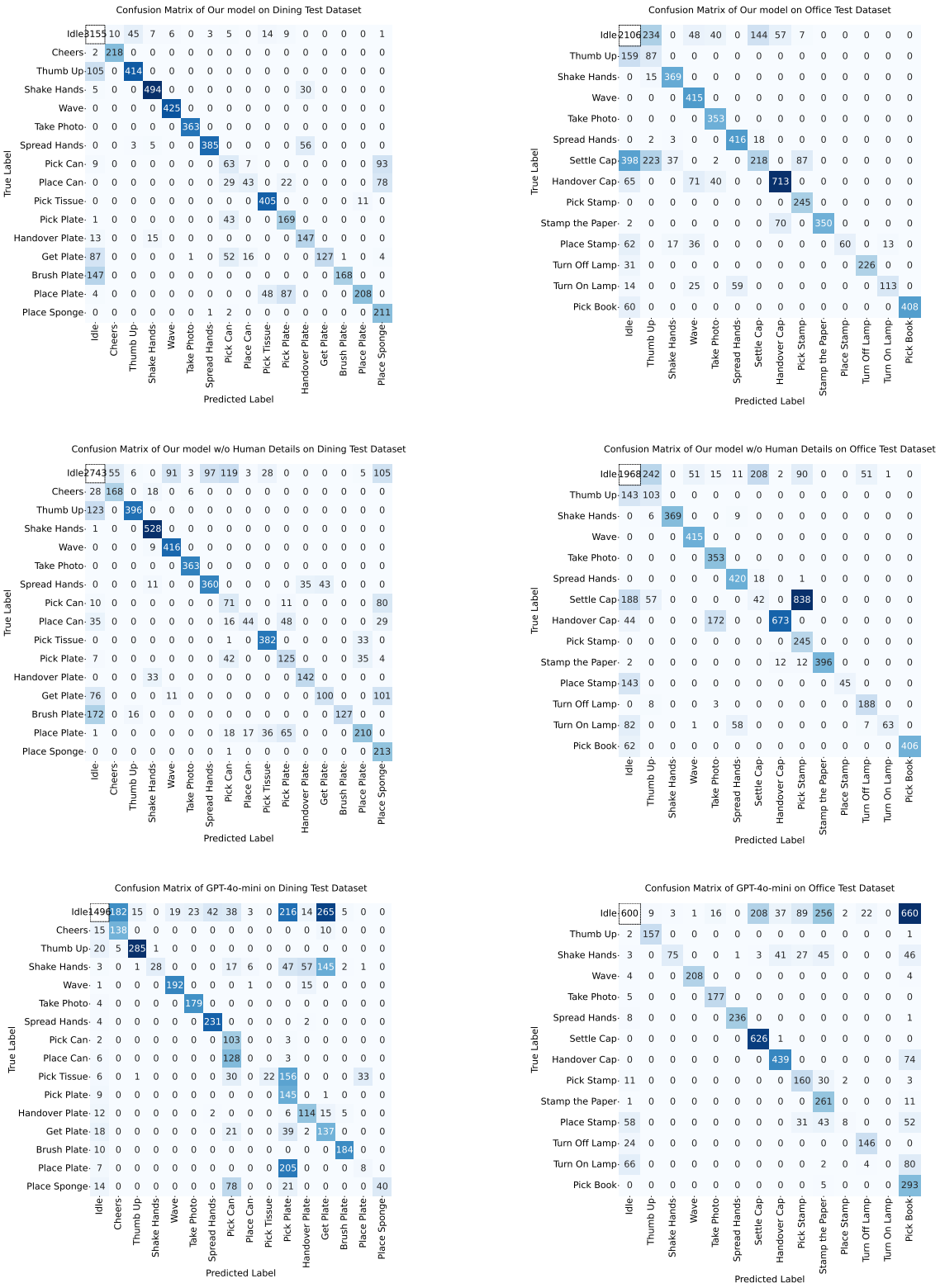


Figure 14. **Confusion Matrices of Our Model (with and without Human Details) and GPT-4o-mini.** To show the results more clearly, we did not color the cell in the top left corner since "idle" accounts for a significant proportion in the data.

to 1.81 meters in height and 50 to 80 kg in weight, with 5 participants having prior research experience in robotics-related fields. Each participant earns RMB 50 Yuan for the experiment.

Each participant independently completed a single experimental session. During the trial, participants first reviewed the experimental statement and received instructions from the experimenter on motions to express human intentions. They subsequently interacted with three different systems that implement different methodologies (RHINO, E2E, and E2E-I) presented in a randomized order. To maintain blindness, the systems were distinguished by colored badges attached to the robot rather than methodological labels. Each 5-minute interaction period was systematically recorded, with validation confirming that participants actively engaged in expressing valid human intentions across all three experimental conditions.

G.5.2. Experimental Statement

Before starting the experiment, participants were required to review and sign the following experimental statement as an informed consent form.

Experimental Statement of the Human Study

Experimental Statement

Purpose

This study investigates human-humanoid robot interaction mechanisms. You are cordially invited to participate as a research subject. The experiment utilizes a Unitree H1 humanoid robot. You will deliver intention commands through body movements, to which the robot will respond with interactive actions, including handshakes and celebratory gestures.

Procedure

The experiment comprises the following steps (20-30 minutes in total):

1. Review and sign the informed consent form after confirming understanding of the protocol, then complete a demographic questionnaire.
2. Receive operational instructions from researchers and perform device familiarization.
3. **Formal experiment phase:**
   - Transmit intention signals via predefined body movements
   - Evaluate the robot's response latency and action precision
   - Interact with 3 control strategies (randomized order), each involving seven intention command types
   - Participants are encouraged to actively assess the robot's generalization capabilities through the following optional actions:
     (a) Modifying seated postures
     (b) Altering visual identifiers (clothing, accessories)
     (c) Convey distinct intention commands before the robot completes its current response
4. Complete multidimensional assessments for both control strategies post-interaction.

Risks and Discomforts

Primary safety considerations involve potential physical contact. Safeguards include:

- Pre-tested safety protocols triggering speed reduction/cessation near participants
- Continuous monitoring by technicians with emergency termination authority
- Minimal electromagnetic emissions from monitoring systems will be present. Current research confirms that standard computer equipment poses no health risks.

Anthropomorphic features may elicit uncanny valley effects. Participants may request immediate termination at any time.

Costs

Participants completing all experimental components with valid submissions will receive compensation between ¥50 according to their performance.

Confidentiality

Research data may be published in academic venues or used for educational purposes. Recordings include:

1. Participant attire and motion patterns
2. Anonymized biometric data (face blurring, voice modulation, etc.)
3. Explicit consent is required for identifiable media usage

Participants are contractually bound to maintain confidentiality regarding experimental particulars, outcomes, and co-participant information prior to formal publication.

Sign below to confirm understanding and acceptance:

_____

Participant Signature Date: _____

G.5.3. Questionnaire

Following experimental completion, participants are asked to complete an anonymous questionnaire requesting a comparative ranking of the three methodologies across multiple performance dimensions. To mitigate ordering bias, response options were initially presented in randomized sequences. The questionnaire comprised the following components (excluding characteristic feature collection):

Questionnaire of the Human Study

Please provide a comparative ranking of the three systems:

- **Overall Performance.**
  - ☐ Blue (blossom) badge
  - ☐ Yellow (star) badge
  - ☐ Red (cherry) badge

- **Understanding:** Accurately interpreted your movement intentions.
  ☐ Yellow (star) badge
  ☐ Red (cherry) badge
  ☐ Blue (blossom) badge
- **Interruption:** Swiftly switched to new intentions when receiving updates mid-skill execution.
  ☐ Blue (blossom) badge
  ☐ Red (cherry) badge
  ☐ Yellow (star) badge

Options are omitted for readability in the following questions.

- **Task Completion:** Successfully performed actions (e.g., handshakes, cheers) to complete a task after intent recognition.
- **Naturalness:** Exhibited smoother or more human-like movements.
- **Generalization:** Maintained task completion capability when altering your posture, wearing, or other characteristics.
- **Time:** Appropriate response time between intent command delivery and the start of robot movement.
- **Safety:** Implemented strategies ensured no harm to humans or objects during operation.