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.*

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

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

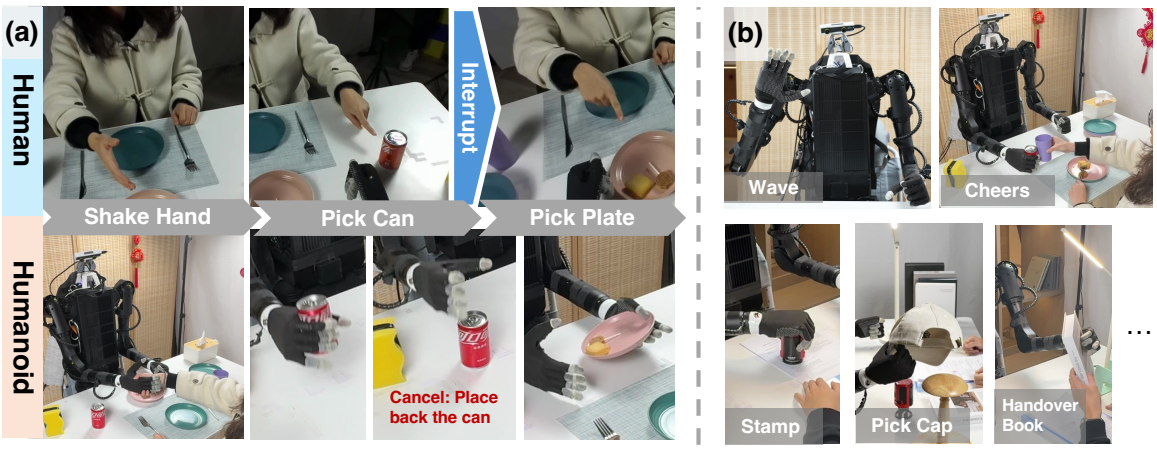

Figure 1. **RHINO has the capabilities of real-time interaction on diverse tasks.** (a) RHINO enables real-time humanoid-human-object interaction, allowing seamless task interruption and dynamic switching during operation. (b) The system demonstrates diverse capabilities, including waving, cheering,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

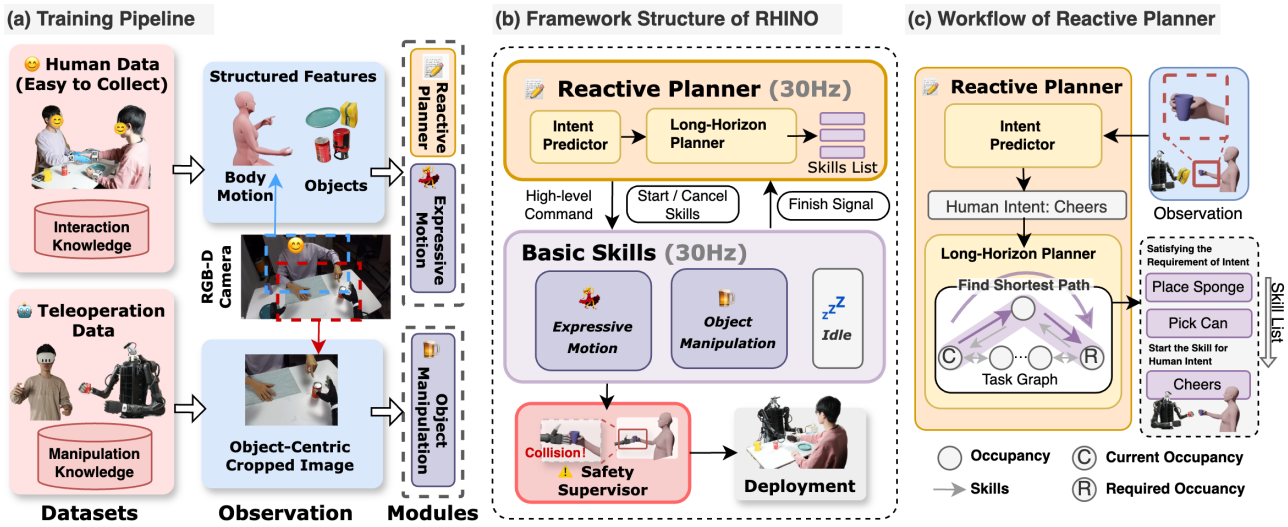

Figure 2. **Overview of RHINO framework.** (a) Training pipeline. Reactive planner and expressive motion modules are trained on human-human data, while the manipulation module is trained on teleoperation data. RHINO extracts human skeleton motion from the image to enhance the generalization of human 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-

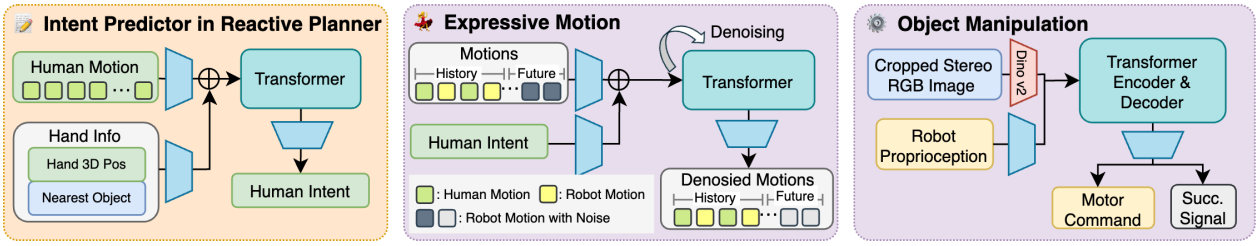

Figure 3. Network architecture of RHINO modules

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,

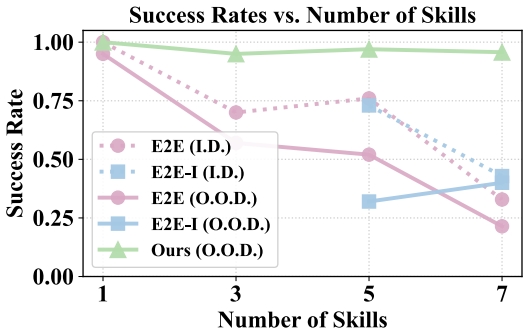

Figure 4. Model success rates with different numbers of skills.

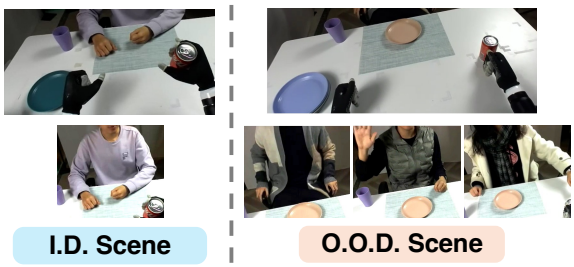

Figure 5. I.D./O.O.D. scene examples (camera view).

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

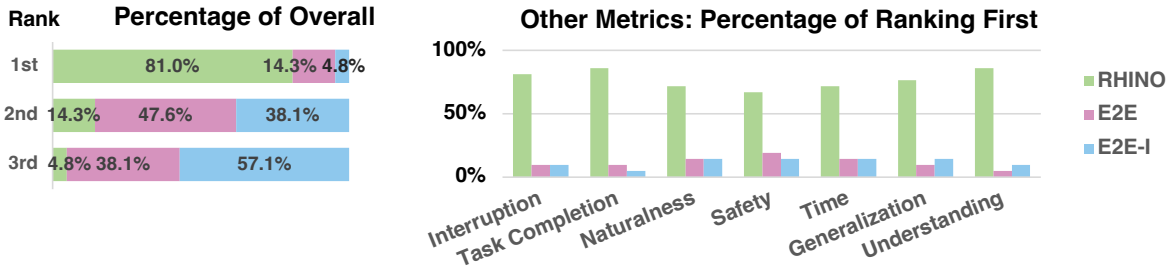

Figure 7. Results of the human study. Left: Overall ranking distribution across the three systems. Right: Percentage of participants who ranked each system first on individual metrics.

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