

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 Appendix 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.

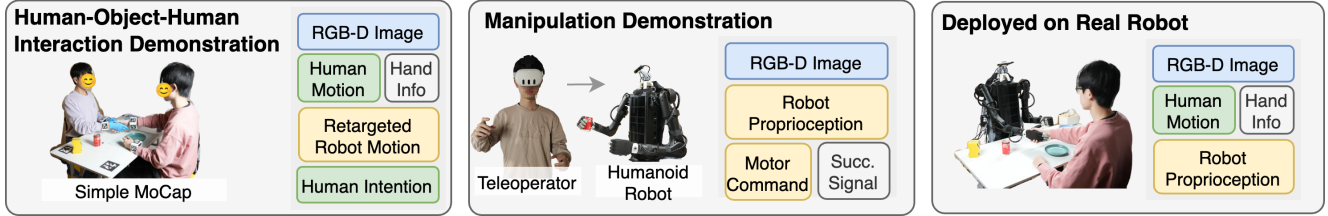


Figure 8. **Information collected in the dataset and the requirement of real robot deployment.** The left box shows our Human-Object-Human Interaction Demonstration Data, which contains human motion and hand position information collected from RGB-D images and a simple motion capture system. The box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

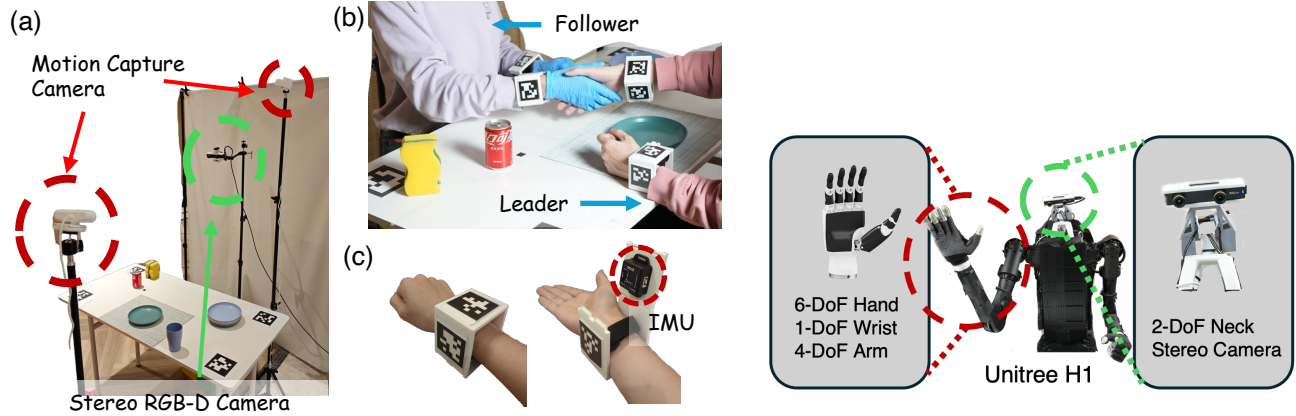


Figure 9. **Setup of motion capture system.** a) The two Motion Capture Cameras are used to detect the ArUco markers. b) Follower and Leader both wear wristbands with ArUco markers for hand position detection. c) We 3D-print our wristbands with 4 ArUco markers on 4 surfaces and embed an IMU beneath the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

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

```

1: Skill  $\leftarrow$  Idle
2: while true do
3:   human_intent  $\leftarrow$  Recognize_Human_intent()
4:   if human intent is stable for k frames and human intent  $\neq$  current intent then
5:     if human intent = Idle and Skill = Manipulation then
6:       Continue
7:     if Skill = Manipulation and interruptionAllowed then
8:       Skill  $\leftarrow$  Reverse_Skill(Skill)
9:     if Start_Condition(human_intent) is not satisfied by hand occupancy then
10:      path  $\leftarrow$  FindPath(occupancy, StartCondition(human_intent))
11:      Skill  $\leftarrow$  Execute_Path(path)
12:     else
13:       Skill  $\leftarrow$  Corresponding Skill(human_intent)
14:     else if SkillSucceeded(Skill) or SkillTimeout(Skill) then
15:       Skill  $\leftarrow$  Idle
16:       if SkillSucceeded(Skill) then
17:         Hand Occupancy  $\leftarrow$  End_Transition(Skill)

```

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.

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

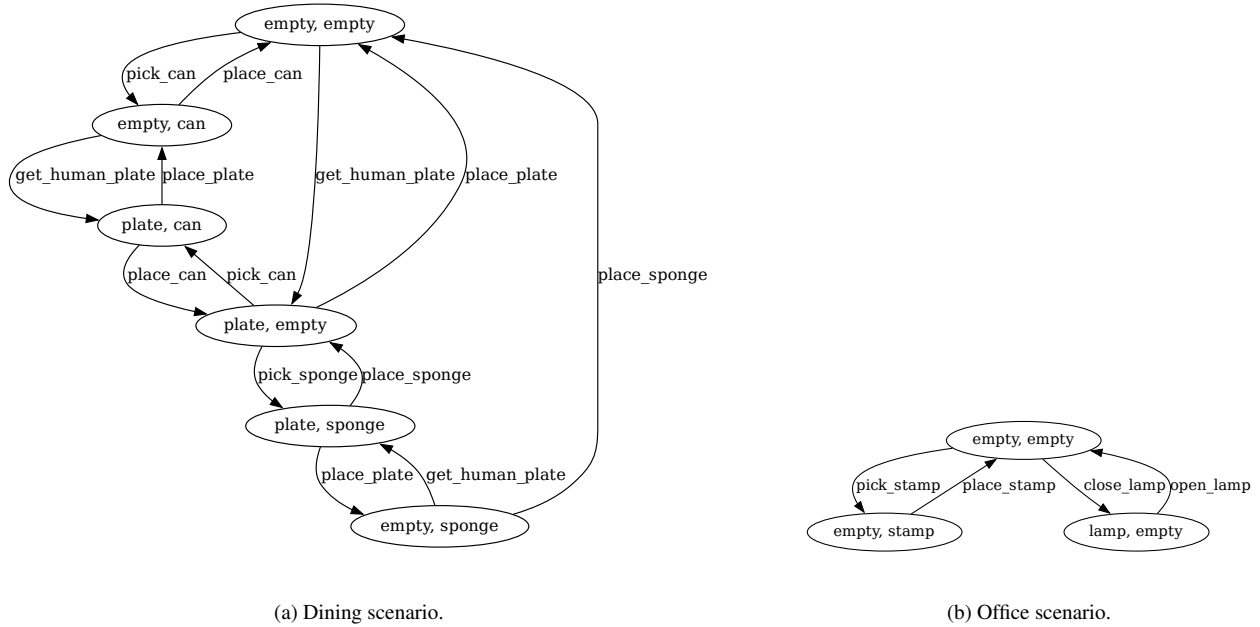


Figure 12. Occupancy graph of two different scenarios.

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

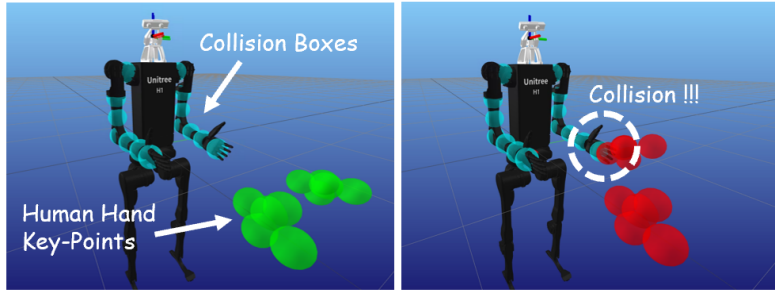


Figure 13. Visualization of the safety supervisor, with sphere markers representing the collision boxes of the human hands and robot arms. These markers move in sync with the interaction. When an unsafe collision is detected, the human hand markers change color from green to red.

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.
- **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 de-

viate 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

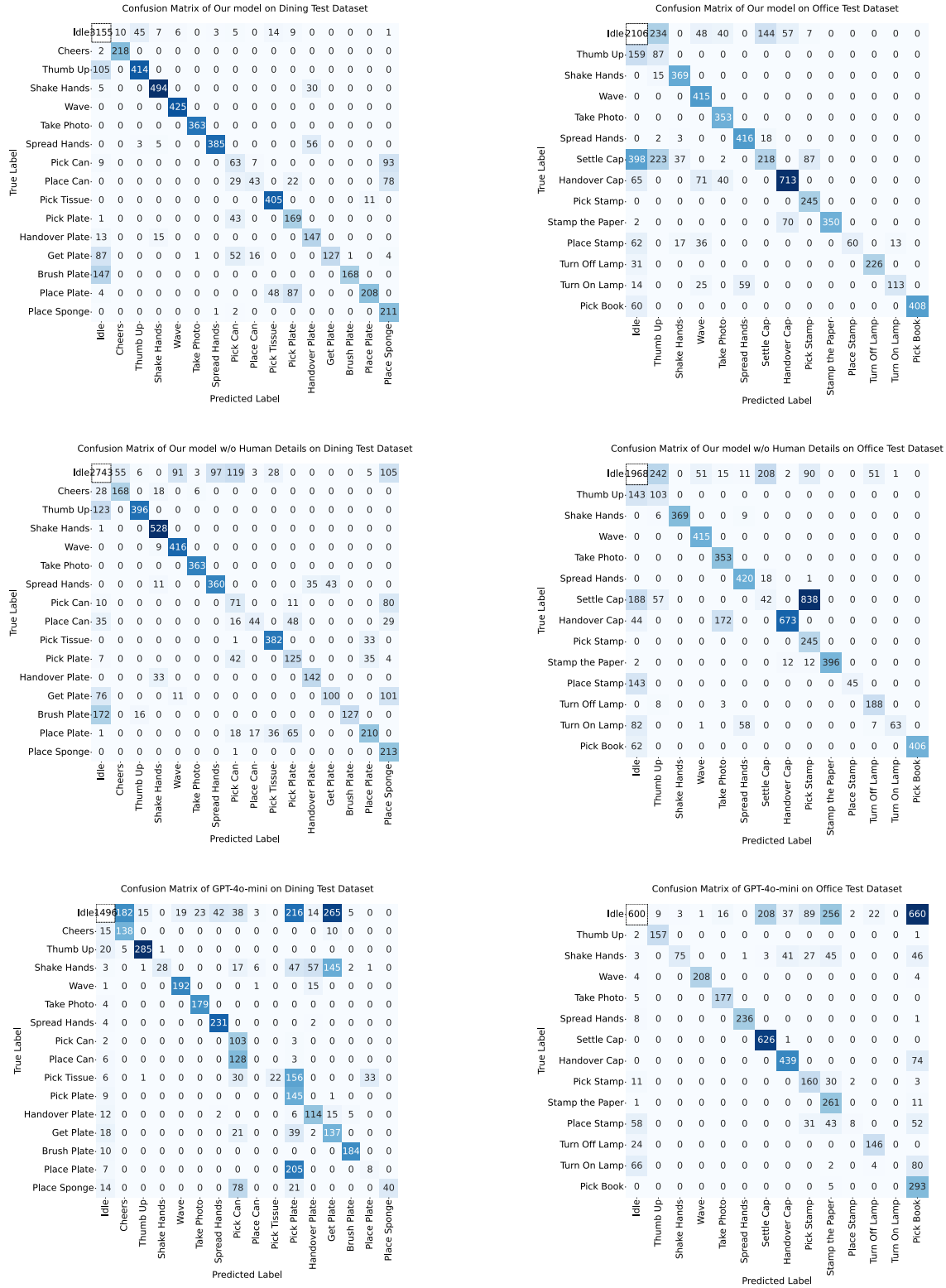


Figure 14. **Confusion Matrices of Our Model (with and without Human Details) and GPT-4o-mini.** To show the results more clearly, we did not color the cell in the top left corner since "idle" accounts for a significant proportion in the data.

1192 to 1.81 meters in height and 50 to 80 kg in weight, with
1193 5 participants having prior research experience in robotics-
1194 related fields. Each participant earns RMB 50 Yuan for the
1195 experiment.

1196 Each participant independently completed a single ex-
1197 perimental session. During the trial, participants first re-
1198 viewed the experimental statement and received instruc-
1199 tions from the experimenter on motions to express hu-
1200 man intentions. They subsequently interacted with three
1201 different systems that implement different methodologies
1202 (RHINO, E2E, and E2E-I) presented in a randomized or-
1203 der. To maintain blindness, the systems were distinguished
1204 by colored badges attached to the robot rather than method-
1205 ological labels. Each 5-minute interaction period was sys-
1206 tematically recorded, with validation confirming that par-
1207 ticipants actively engaged in expressing valid human inten-
1208 tions across all three experimental conditions.

1209 **G.5.2. Experimental Statement**

1210 Before starting the experiment, participants were required
1211 to review and sign the following experimental statement as
1212 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.