

388 Appendix

389 A Task Description

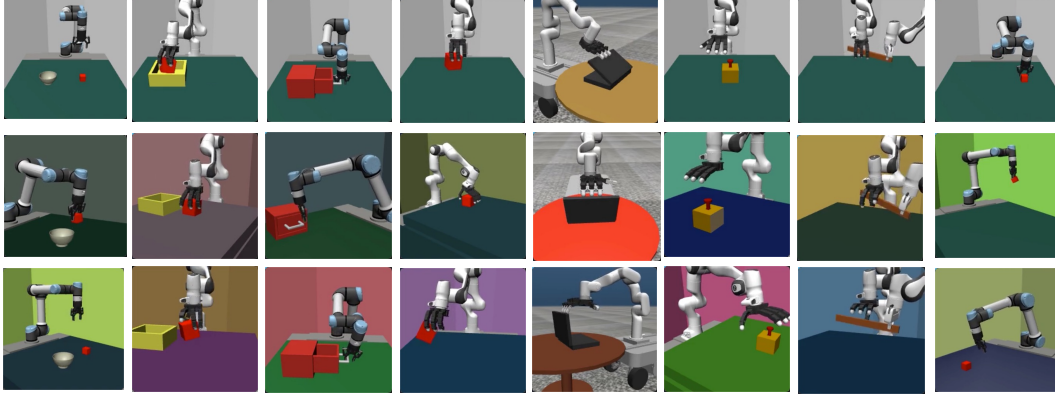


Figure 8: Snapshot of all tasks and test visual scenarios.

390 **Lift Cube:** This task involves a UR5 arm equipped with a Robotiq gripper. A red cube is placed on
 391 the table. The agents are required to grasp the cube and lift it off the table. A reward greater than
 392 250 is considered a success. We lock 3 out of the 6 DoFs of the UR5 arm to restrict unnecessary
 393 movements and reduce the action space, facilitating more efficient RL learning.

394 **Pull Drawer:** This task contains a UR5 arm equipped with a Robotiq gripper. A drawer is placed on
 395 the table. The agents need to approach the handle and pull the drawer open. A reward greater than
 396 230 is considered a success. We lock 3 out of the 6 DoFs of the UR5 arm.

397 **Pick Cube To Bowl:** Except for the red cube, we additionally place a bowl on the table. The agent
 398 needs to lift the cube and place it into the bowl. A reward greater than 230 is considered a success.
 399 We lock 3 out of the 6 DoFs of the UR5 arm.

400 **Button with Dex:** This task involves a Franka arm equipped with an Allegro Hand. The agent is
 401 required to press the button to receive the reward. A reward greater than 250 is considered a success.
 402 We lock 3 out of the 7 DoFs of the Franka arm and the DoFs of Allegro Hand.

403 **Close-Laptop Dex:** This task is equipped with a Leap Hand, an XArm, and a Ranger Mini 2 base
 404 from AgileX. The agent requires to close the laptop on the table. We lock the DoFs of Leap hand and
 405 4 DoFs of Franka Arm. When the joint of the laptop is smaller than 1.7 rad, we consider it a success.

406 **LiftCube Dex:** This task involves a Franka arm equipped with an Allegro Hand. The agent is
 407 required to grasp the cube and lift it off the table. A reward greater than 50 is considered a success.
 408 We lock 3 out of the 7 DoFs of the Franka arm and use 4 DoFs of Allegro Hand (The rest of the DoFs
 409 will be set to a default value to keep a proper gesture).

410 **PickPlace Dex:** This task involves a Franka arm equipped with an Allegro Hand. The agent is
 411 required to grasp the cube and lift it off the table and place it to the box. A reward greater than 50
 412 is considered a success. We lock 3 out of the 7 DoFs of the Franka arm and use 4 DoFs of Allegro
 413 Hand (The rest of DoFs will be set to a default value to keep a proper gesture). Additionally, we use
 414 the moving average technique to smooth the motion.

415 **Handover Dex:** We utilize two Franka arms, one equipped with a gripper and the other with an
 416 Allegro hand. This task requires cooperation between the two arms; the gripper must grasp a spatula
 417 and pass it to the hand. Success is determined if the distance between the hand and the object is less
 418 than 0.03 meters.

B Implementation Details

B.1 Environment Randomization Parameters

Table 5: Domain randomization parameters in Maniwhere.

Attribute	Value
UR5 joint armature	$0.1 \cdot (1 \pm 0.1) \text{ kg m}^2$
UR5 shoulder pan joint damping	$360 \cdot (1 \pm 0.1) \text{ N s/m}$
UR5 shoulder lift joint damping	$280 \cdot (1 \pm 0.1) \text{ N s/m}$
UR5 elbow joint damping	$250 \cdot (1 \pm 0.1) \text{ N s/m}$
UR5 wrist joint damping	$280 \cdot (1 \pm 0.1) \text{ N s/m}$
Franka joint armature	$0.1 \cdot (1 \pm 0.1) \text{ kg m}^2$
Franka joint damping	$1 \cdot (1 \pm 0.1) \text{ N s/m}$
XArm joint damping	$15 \cdot (1 \pm 0.1) \text{ N s/m}$
XArm joint frictionloss	$4 \cdot (1 \pm 0.1)$
Object Cube Size	$0.05 \cdot (1 \pm 0.1) \text{ m}$
Table height	$[-0.01, 0.01] \text{ m}$
Camera Pitch	$[10.5, 30.5]^\circ$
Camera Yaw	$[-60, 60]^\circ$
Camera Fov	$[38, 46]^\circ$
Camera Distance	$[1.12, 1.54] \text{ m}$
Action-delay	$[0, 2] \text{ timesteps}$
Control timestep	$[0.016, 0.024] \text{ s}$

B.2 Curriculum Randomization

For each task, a threshold of $2e5$ steps is established as the initial frame for domain randomization. The randomization parameters will vary exponentially within the ranges specified in Table 5 starting from the $2e5$ -step mark (the Close Laptop task beginning at $7e4$ step). Concurrently, the stabilizing objective described in Eq 4 will process augmented images from the fixed view prior to this threshold, and will incorporate augmented images from the moving view thereafter.

B.3 Hyper-Parameters

We list the training hyper-parameters used in Maniwhere in Table 6.

C Additional Results

C.1 Real-world Experiments

Real-world setup. Due to the limitation that a single PC cannot control two Franka arms simultaneously, we developed a control logic framework using `zmq` to coordinate three PCs. In this setup, one PC is regarded as the client, while the other two serve as servers. The client PC receives visual input and performs network inference, subsequently transmitting the inferred actions via socket connections to the two server PCs. The server PCs are responsible for controlling the Franka arms and executing the received actions. This process is iterative, with the servers sending new visual input back to the client for continuous processing. Given that MV-MWM has a large model size and requires substantial memory for loading, we deployed it on a desktop equipped with an RTX 3090 GPU. In contrast, the deployment of Maniwhere demands significantly less hardware, allowing it to perform inference even on CPU desktops. Regarding the camera setup, we establish the evaluation

Table 6: Hyper-parameters in Maniwhere.

Hyper-parameters	Value
Input size	128×128
Discount factor γ	0.99
Replay Buffer size	int(1e7)
Feature dim	256
Action repeat	1
N-step return	3
Optimizer	Adam
Frame stack	3
Temperature of InfoNCE	0.1
Learning Rate of STN	1e-4
λ	200

Simulation

Sim2Real Transfer

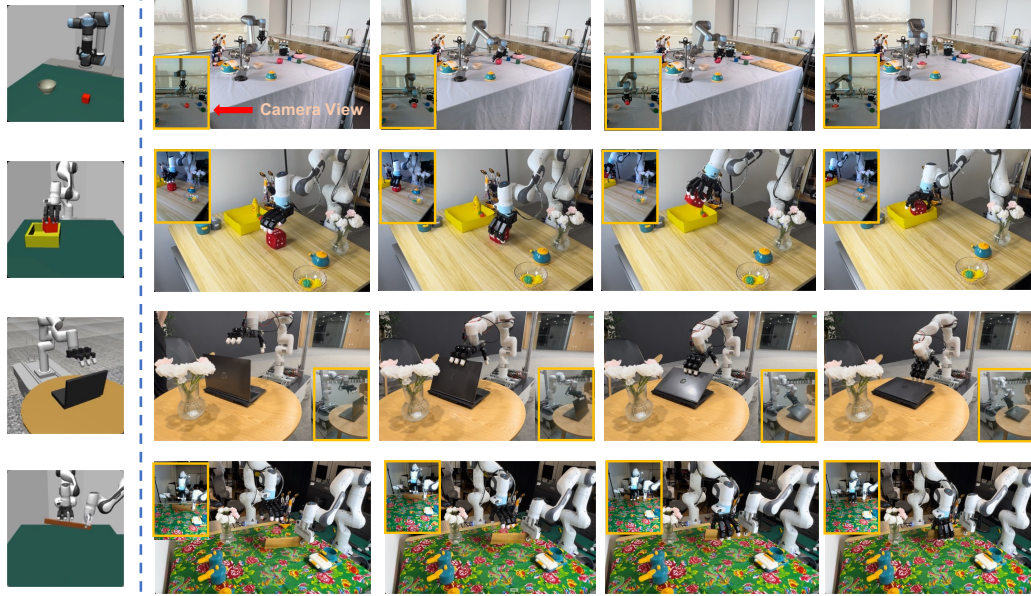


Figure 9: **More real-world snapshots..** We exhibit more real-world snapshots in challenging real-world visual scenarios.

viewpoints at three yaw angular ranges: $[0, 5^\circ]$, $[10, 25^\circ]$, and $[40, 55^\circ]$, on both the left and right sides. Additionally, across the five trials conducted at each viewpoint, the camera height will be varied within a range of -3 to 3 cm.

Instance generalization. Thanks to the general grasping capabilities of the dexterous hand, Figure 10 shows that Maniwhere is not limited to a single object when executing the *lifting* behaviours and can generalize across different instances with various shapes and sizes.

C.2 Cross Embodiment

Figure 11 illustrates that when we first select a pixel point on the UR5 original image (marked with a red pentagram) and extract its feature (enclosed in the orange square) after passing through the convolutional layer, we compute its normalized cosine similarity with the image feature of Franka arm to obtain a similarity map. The point with the highest value in this map is identified as the most similar point between two images (marked with a red pentagram). As shown in Figure 11, Maniwhere

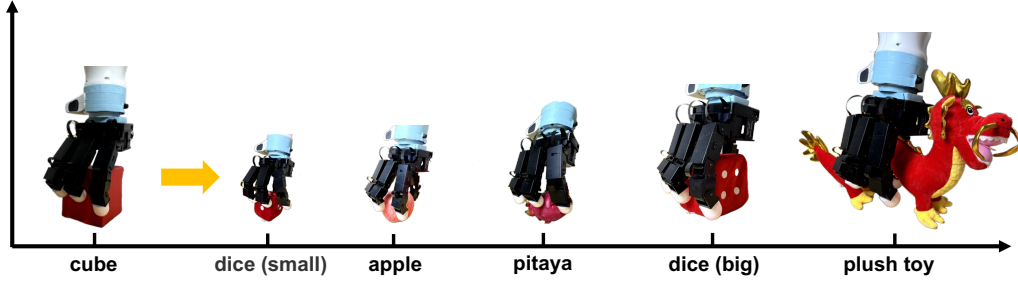


Figure 10: **Instance Generalization.** We find that Maniwhere won’t overfit to the specific object size and shape.

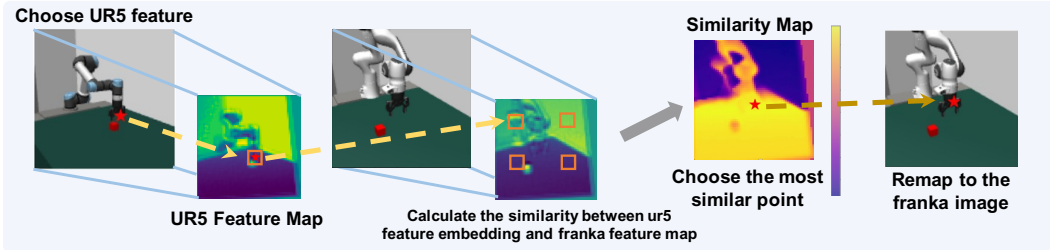


Figure 11: **Feature Correspondence.** Maniwhere can find the feature correspondence between different embodiments.

453 can effectively recognize semantically consistent positions between the two different embodiments.
 454 With respect to randomization, to enable the agent to capture the correspondence information through
 455 the multi-view representation objective, we do not augment the moving view image in Eq 4.

456 C.3 View Generalization

457 We further investigate how Maniwhere’s performance varies across different camera view ranges. We
 458 divide the randomized camera view range into three parts, within each of which the camera’s pitch
 459 and field of view are randomly altered as well. The value for each range is calculated as the average
 460 of both the left and right sides. Due to the excessive angular range in handover task potentially
 461 obscuring the other arm, we confined the range for this task to 0-30 degrees. Table 7 illustrates that,
 462 although Maniwhere’s performance exhibits a slight decline as the angle increases, it still retains the
 463 capability to handle these scenarios effectively.

Table 7: **Generalization across different camera view ranges.** Maniwhere retains the generalization capability to handle these scenarios effectively. We evaluate 20 episodes in each range.

Method / Task	LiftCube Dex	PickPlace	Pickplace dex	Button dex	Handover
range [0, 15]°	91.3%	91.0%	82.5%	97.5%	94.0%
range [20, 35]°	88.3%	88.0%	81.5%	97.5%	94.0%
range [45, 60]°	86.9%	84.0%	65.0%	94.4%	92.0%

464 C.4 Depth information helps sim2real transfer

465 To ensure the depth images closely resemble real-world conditions, we first pre-process the depth
 466 image. We introduce Gaussian noise $\mathcal{N}(0, 0.01)$ and depth-dependent noise $\mathcal{N}(0, \text{depth_scale})$,
 467 where the depth_scale equals $\text{np.abs}(\text{depth_image}) * 0.05$. Then, we apply GaussianBlur
 468 to smooth the noise. Additionally, the depth values are clipped to within 2 meters and normalized
 469 to the range [0, 255]. During sim2real, we find that depth image can largely help to alleviate the



Figure 12: **Spatial illusion.** These two figures are captured at the same timestep. Without depth information, we lose the front-to-back positional relationship between the object and the gripper in the three-dimensional world.

ambiguity situation. Figure 12 shows that when encountering large camera viewpoints, the agent cannot accurately determine the grasping position since RGB information alone does not provide the necessary front-to-back positional relationship between the object and the gripper in the 3D world. However, by incorporating depth images, we observe a significant improvement in real-world scenarios.

C.5 MV-MWM with data augmentation

We also apply the data augmentation method on MV-MWM. As shown in Table 8, MV-MWM suffers a significant performance drop while facing data augmentation. These results are consistent with the recent works [9, 12].

Task	Success Rate(w/o DA)	Success Rate (w/DA)
Button Dex	77.6 \pm 14.2 %	1.3 \pm 2.3 %
PickPlace Dex	34.0 \pm 28.9 %	8.7 \pm 13.3 %

Table 8: **MV-MWM with data augmentation.**

Naively applying data augmentation can cause instability and large variance during training. In turn, the results also prove that simultaneously handling multiple types of generalization is non-trivial and highlights the superiority of Maniwhere.

C.6 Regarding target object color

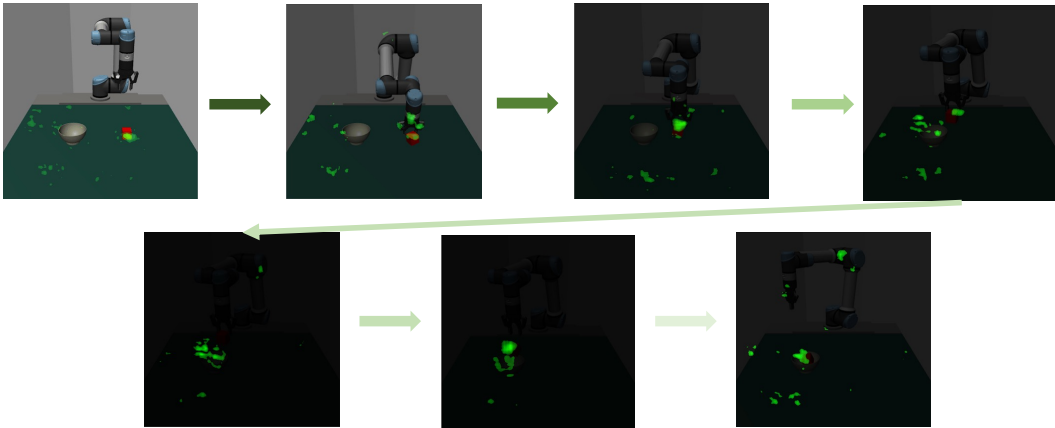


Figure 13: **Visualization of the agent's attention by Grad-CAM.**

Although we found that the agent demonstrates strong generalization capabilities when the visual scene is altered, including changes to the table, background, and the introduction of colorful dis-

tractors, it fails the task when the color of the target object is changed. Figure 13 exhibits that during executing a trajectory, the agent focuses more attention on the target object while ignoring task-irrelevant information, making it more sensitive to changes in the color of the target object. We use the Grad-CAM [53] to visualize the agent’s attention.