Supplementary Materials: One-Shot Imitation Learning with Invariance Matching for Robotic Manipulation

1. Failure Case Analysis

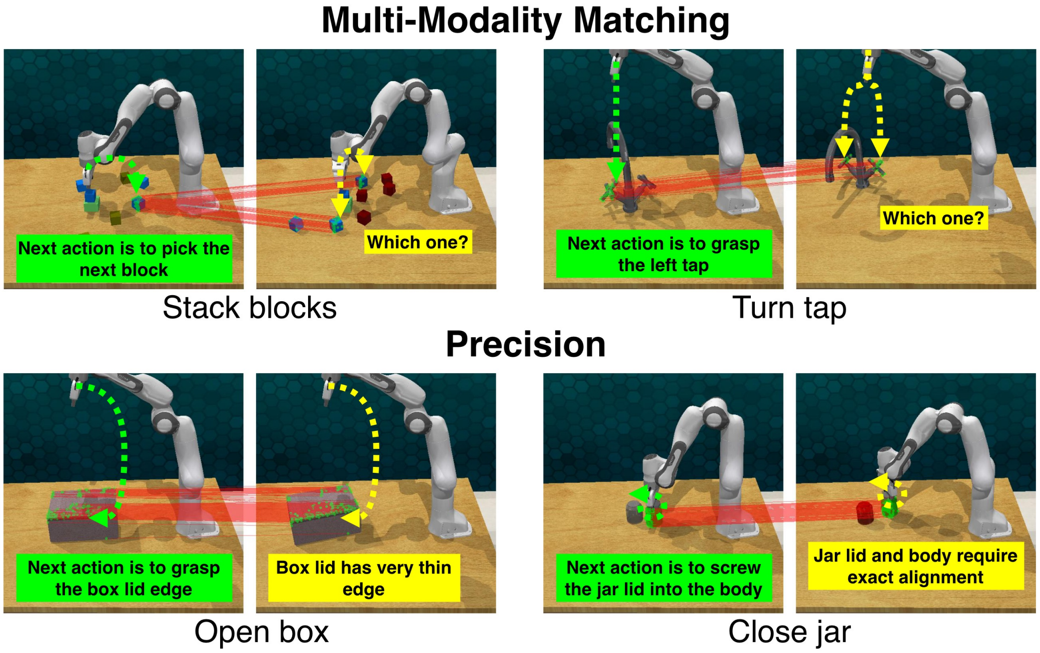
Multi-Modality Matching
Next action is to pick the next block
Which one?
Stack blocks
Next action is to grasp the left tap
Which one?
Turn tap
Precision
Next action is to grasp the box lid edge
Box lid has very thin edge
Open box
Next action is to screw the jar lid into the body
Jar lid and body require exact alignment
Close jar

Figure 1. Failure cases at tasks that require multi-modality matching and high precision. The demonstrations and test scenes are annotated in green and yellow. The arrows indicate the trajectory of the next action.

Figure 1 and 2 show failure cases of our method IMOP, which are categorized and detailed in the following. We hope this analysis provides insights for our work and potential future directions.

Multi-Modality Matching The first row in Figure 1 shows two tasks: *stack blocks* and *turn tap*. The invariant region is correctly located to the next block but is matched

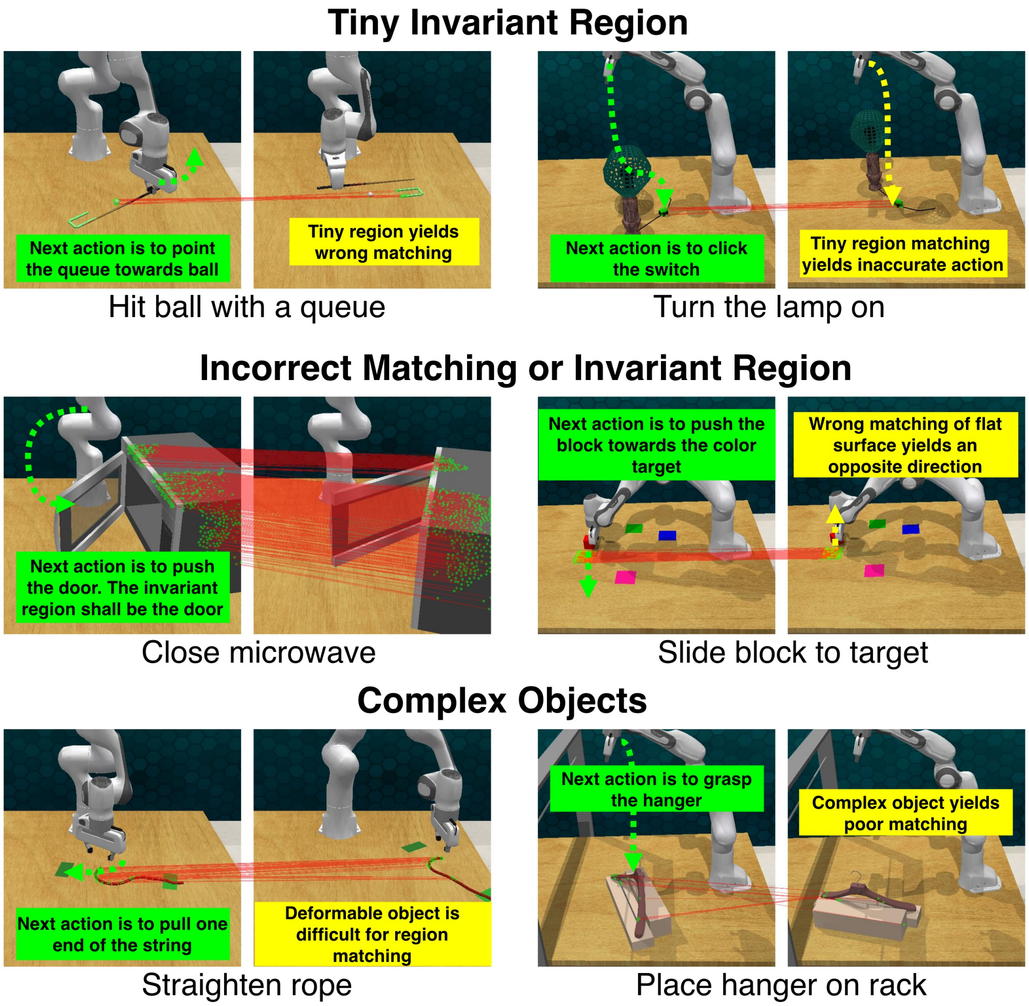
Tiny Invariant Region
Next action is to point the queue towards ball
Tiny region yields wrong matching
Hit ball with a queue
Next action is to click the switch
Tiny region matching yields inaccurate action
Turn the lamp on
Incorrect Matching or Invariant Region
Next action is to push the door. The invariant region shall be the door
Close microwave
Next action is to push the block towards the color target
Wrong matching of flat surface yields an opposite direction
Slide block to target
Complex Objects
Next action is to pull one end of the string
Deformable object is difficult for region matching
Straighten rope
Next action is to grasp the hanger
Complex object yields poor matching
Place hanger on rack

Figure 2. Failure cases at tasks with tiny invariant regions, wrong invariance matching results, and complex objects. The demonstrations and test scenes are annotated in green and yellow. The arrows indicate the trajectory of the next action.

to several blocks of the same visual appearance in the test scene. This intrinsic multi-modality affects the performance of our method IMOP. This limitation can be potentially addressed by diffusion models as a future improvement.

Precision The second row in Figure 1 shows two tasks: *open box* and *close jar*. The invariant region matching is visually reasonable as in the box lid and jar body. However, the open box task requires the grasping of a thin lid edge, and the close jar task requires an exact alignment between the jar lid and body. The action pose derived from the matched correspondence sometimes lacks the precision to complete the task.

Tiny Invariant Region The first row in Figure 2 shows two tasks: *hit ball with a queue* and *turn the lamp on*. The first task requires pointing the queue toward the ball and hitting the ball into the container, the second task requires clicking the small switch area. The invariant regions are correctly estimated in both cases, as in the ball and switch area. However, the ball is incorrectly matched to the test scene, possibly because the ball is too small which increases the matching difficulty. The switch is visually matched correctly. However, the tiny matched switch area easily leads to a less accurate pose solution.

Incorrect Matching or Invariant Region The second row in Figure 2 shows two tasks: *close microwave* and *sldie block to target*. The desirable action of the first task is to push the door towards the microwave body. Therefore, the door area shall be the invariant region. However, IMOP mistakes the microwave body as the invariant region. For the second task, the invariant region is correctly located as the target color area. However, the flat surface of the target area leads to an incorrect matching that corresponds to the opposite pushing direction.

Complex Objects The third row in Figure 3 shows two tasks: *straighten rope* and *place hanger on rack*. The first task requires the handling of a non-rigid rope object, which is intrinsically difficult for region matching because the optimization problem in Equation 1 assumes a rigid body transformation. The second task requires the grasping and placing of a cloth hanger, an object unseen during training and has a curvy surface and bar-like inner structures. This affects the IMOP's ability to estimate and match invariant regions.

2. Picking Flipped Objects

Figure 3 shows the policy behavior of picking flipped objects while providing demonstrations of non-flipped. We find that our method IMOP always predicts a picking action from the top (non-flipped) rather than an invalid action below the table surface, even in cases of vertically asymmetric objects such as mustard bottles and cups. We evaluate the action pose of picking by sampling 50 object instances and flipping them in tasks *put grocery into cupboard* and *stack cups*. IMOP predicts a successful picking pose for 46 out of 50 trials. IMOP still predicts a valid pose above the table for the failed cases. We believe the reason is that IMOP uses matching as an intermediate step to determine the future action. Since the training data only contains valid actions, IMOP learns to predict the match that corresponds to valid actions instead of naively finding the closest match.

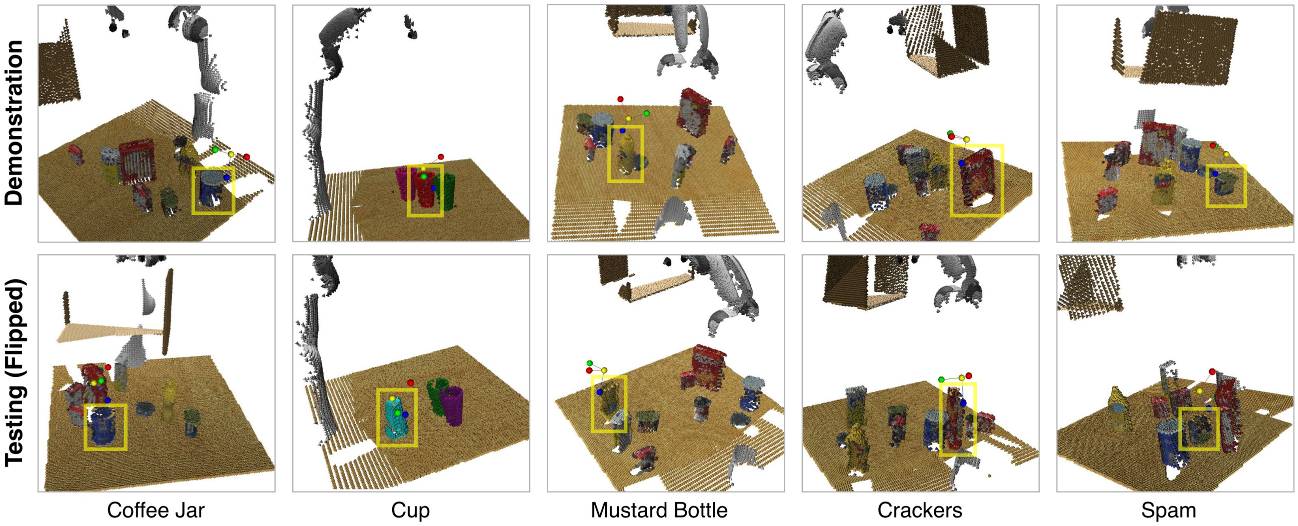

Figure 3. Point cloud visualization of picking a flipped object with a non-flipped demonstration. The object to be picked is highlighted with a yellow bounding box. The demonstrated and predicted action poses are visualized as a colored frame with red, green, blue, and yellow as the x-axis, y-axis, z-axis, and origin.